Business relation extraction from texts

Hadjer Khaldi

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Hadjer KHALDI

Le 12 décembre 2022

Extraction de relations économiques à partir de textes

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Farah BENAMARA ZITOUNE et Nathalie AUSSENAC-GILLES

Jury

M. Thierry CHARNOIS, Rapporteur
Mme Claire NEDELLEC, Rapporteure
M. Raphaël TRONCY , Examinateur
Mme Pascale ZARATE, Examinatrice
Mme Farah BENAMARA ZITOUNE, Directrice de thèse
Mme Nathalie AUSSENAC-GILLES, Co-directrice de thèse

Abstract

Relation extraction is a subtask of information extraction that attempts to identify semantic links between entities in unstructured texts. When it comes to business content, this task has received less attention than other generic or specific domains (e.g. biomedical).

This thesis is structured around three objectives: (1) Detecting various types of business relations in multilingual content, (2) Investigating the problem of data imbalance in a relation extraction task between relations of interest and the other relations, and (3) Investigating whether incorporating different sources of knowledge about entities into the relation extraction model, can improve its performances.

Most of the proposed works to extract business relations are targeting monolingual contents, where the dataset used in the case of supervised approaches are annotated using different annotation schemes. In this thesis, we propose a unified characterization to describe business relations, and we use it to annotate the first multilingual dataset for business relations in four languages (French, English, Spanish, and Chinese). A set of monolingual and cross-lingual models are proposed and tested on this dataset for automatically extracting business relations from multilingual texts on the web.

When approaching relation extraction as a classification problem, the types of relations to be extracted are predefined. This restricted set of relations of interest between entities is under-represented on the open web when compared to all other possible relations between entities (known as negative relation). This raises the issue of data imbalance in the training data used to train the relation classifiers. We investigate this issue in the context of business relationship extraction and propose/adapt various approaches based on data and model level solutions.

Finally, because entities in the relation extraction task play an important role in defining the nature of the relation type between them, incorporating external knowledge about it into the relation extraction model has been shown to be effective in the literature. We propose a systematic evaluation of incorporating different and complementary sources of knowledge about entities into the relation extraction model, using a simpler and lighter neural architecture than previous works.

Our results represent a first step towards the extraction of multilingual economic relations from the web.

Résumé

L'extraction de relations est une sous-tâche de l'extraction d'information qui vise à identifier les liens sémantiques entre les entités dans des textes non structurés. Lorsqu'il s'agit de contenu de nature économique, cette tâche a reçu moins d'attention que d'autres domaines génériques ou spécifiques (par exemple, le domaine biomédical). Cette thèse s'articule autour de trois objectifs : (1) Détecter différents types de relations économiques dans des contenus multilingues, (2) Étudier le problème du déséquilibre des données dans une tâche d'extraction de relations entre les relations d'intérêt et les autres relations, et (3) Étudier si l'incorporation de différentes sources de connaissances sur les entités dans le modèle d'extraction de relations, peut améliorer ses performances.

La plupart des travaux proposés pour l'extraction de relations économiques visent des contenus monolingues, où les jeux de données utilisés dans le cas des approches supervisées sont annotés en utilisant différents schémas d'annotation. Dans cette thèse, nous proposons une caractérisation unifiée pour décrire les relations économiques, et nous l'utilisons pour annoter le premier jeu de données multilingue pour ces relations dans quatre langues (français, anglais, espagnol et chinois). Un ensemble de modèles monolingues et interlingues sont proposés et testés sur ce jeu de données pour extraire automatiquement des relations économiques à partir de textes multilingues à partir du web.

Lorsqu'on aborde l'extraction de relations comme un problème de classification, les types de relations à extraire sont prédéfinis. Cet ensemble restreint de relations d'intérêt entre entités est sous-représenté sur le web ouvert par rapport à toutes les autres relations possibles entre entités (connues sous le nom de relations négatives). Cela soulève la question du déséquilibre des données utilisées pour former les classificateurs de relations. Nous étudions ce problème dans le contexte de l'extraction de relations économiques et nous proposons diverses approches basées sur des adaptations au niveau des données et des modèles.

Enfin, comme les entités dans la tâche d'extraction de relations jouent un rôle important dans la prédicition du type de relation, l'incorporation de connaissances externes à leur sujet dans le modèle d'extraction de relations s'est avérée efficace dans la littérature. Nous proposons une évaluation systématique de l'incorporation de sources différentes et com-

plémentaires de connaissances sur les entités dans le modèle d'extraction de relations, en utilisant une architecture neuronale plus simple et plus légère que les travaux précédents.

Nos résultats représentent un premier pas vers l'automatisation de l'extraction de relations économiques multilingues à partir de contenus textuels sur le web.

Acknowledgements

First and foremost, I want to express my gratitude to Allah, the Almighty, for providing me with the strength, wisdom, and faith to complete this dissertation and cherish every moment of it. Then, I express my gratitude to a few people who have supported and been patient with me throughout this journey.

I'm extremely grateful to Prof. Claire Nédellec and Prof. Thierry Charnois for the time spent reviewing this dissertation and careful attention to detail. Your valuable suggestions and comments helped me improve this work. I'd like to extend sincere thanks to Prof. Pascale Zarate, Dr. Raphael Troncy, Dr. Camille Pradel, and Mr. Grégoire Sigel for serving on my dissertation committee.

I'd like to express my heartfelt appreciation to my supervisor, Dr. Farah Benamara, for her insightful comments and constructive feedback on my work. She has been a great mentor for me, with her great empathy, humility, and availability. Thank you for allowing me to grow as a research scientist. My deepest gratitude goes to co-supervisor Nathalie Aussenac-Gilles for her guidance and support, and for the insightful discussions we had during our meetings.

I also could not have undertaken this journey without the generous support of Geotrend company who financed my research and hosted me for three years of thesis, and I am more than happy to continue the adventure with them as a data scientist. A special thanks goes to Mr. Grégoire Sigel, and Thomas Binant, the co-founders, who trusted me, and to Dr. Amin Abdaoui and Dr. Camille Pradel, for supervising and following closely my research work. I would not forget my Geotrend colleagues, EunBee, Sonia, Amina, Gwen, Daisy, Clara, Auriane and all the others, who made my thesis journey less stressful and more enjoyable.

To my husband, Taqiy Eddine, the person who brought joy and laughter into my life, the one who stood by me during thesis writing, helping me organize my ideas, providing unconditional support, and an endless amount of advice. Thank you for always being there for me

and celebrating every single small accomplishment I've made along the way.

I can't thank my parents enough for their faith in me, their love and support, their encouragement, and for giving me the opportunity to explore new horizons and achieve my dreams. I am nothing without you and your help, I love you. I also would like to thank my sister Sarra, my sister in law Hizia, and my brother Amine, for the joy they bring to my life, for answering my calls at all hours of the day, and for sharing special moments that occur at home while I am away.

Finally, I'd like to thank my pandemic lockdown buddies, Yasmine and Rayhane, you were my family and friends when I was dealing with both home sickness and thesis stress. Thank you for all the planned and unplanned trips, for the endless talks and food we shared, and for your help and support when I needed it the most. May our friendship last forever.

Now, I can say, It was not an easy journey, but I did it elhamdouli'Allah.

Contents

List of Figures

List of Tables

Introduction

Context and Motivations

On the Importance of Structuring Business Relations

The economy of the twenty-first century has changed how market participants interact with one another in a global market where national borders have dissolved and trade has become more open and free (Hameed et al., 2021). Rivalry has moved from the local market level to the multinational level (Gorodnichenko et al., 2008). This incites companies and industries to strengthen their capacity for innovation to deliver competitive products and services, and increase their economic growth and performances (Hameed et al., 2021; Passaris, 2006).

In a complex, rapidly evolving business environment, competitive intelligence (CI) refers to the process of gathering, analyzing and delivering information about the business environment such as the capabilities and intentions of the competitors, and then transforming them into knowledge that can be used by managers for decision-making (Gilad and Gilad, 1986; Kahaner, 1997; Montgomery and Weinberg, 1979; Oberlechner and Hocking, 2004). Due to the huge amount of public information shared and disseminated everyday on the internet, unstructured web contents have become a crucial source for CI, making their manual exploitation impractical (Boncella, 2003). The automatic extraction of business information is therefore a valuable tool for identifying links between specific market stakeholders and building business networks.

One possible way to structure business relations and make the generation of business networks easier, is to organize textual content into financial knowledge graphs, where nodes are financial and business entities and edges linking those entities represent the business interactions between them. Figure A.1 illustrates such a knowledge graph as given by the Geotrend[1] platform. Geotrend is a French SME, who developed a "Market Intelligence" platform that aims to support the discovery, analysis, and monitoring of any market in real time. This platform is based on information extraction components that

[1] https://www.Geotrend.fr/fr/

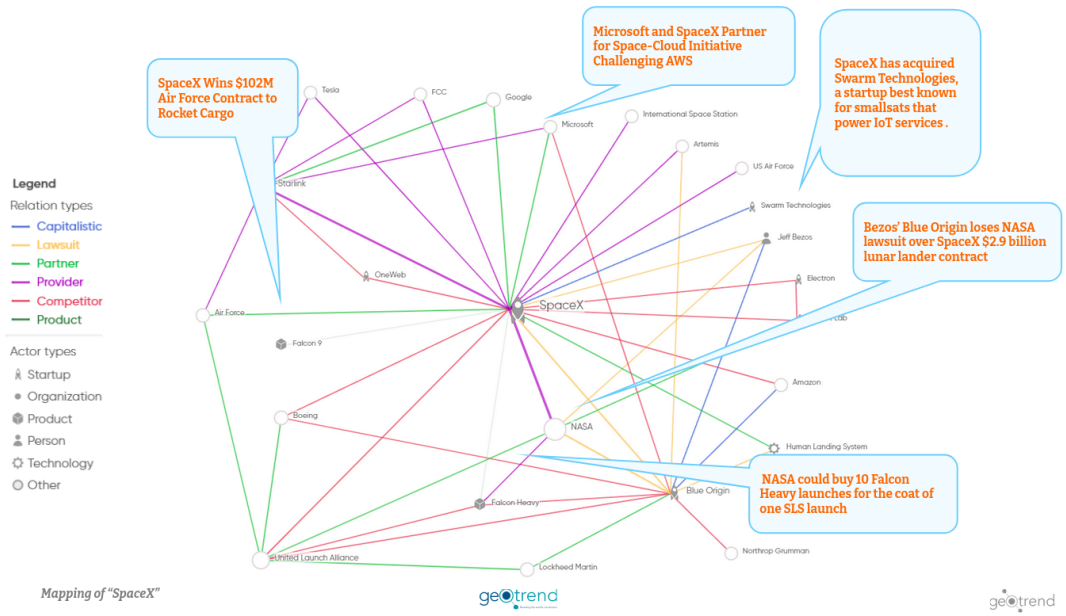

Figure 1 An example of knowledge graph about SpaceX, as given by the Geotrend platform.

extract from the web the main market actors, the relationships that exist between them (e.g., partnership, competition, subsidiarity, etc.), the monetary values that characterize the market, and the important dates and places related to it. In this figure, the graph was created by analyzing hundreds of web-retrieved documents about the *SpaceX company*. Then, business relations are extracted at the sentence level using a Relation Extraction (RE) component. For example, from the sentence in (1), the platform can identify the entities *SpaceX*, and *Swarm Technologies*, and the business relation between them (`acquired_by`).

(1) SpaceX has acquired Swarm Technologies, a startup best known for small satellites that power IoT services.

Competitors of SpaceX can use the generated graph to detect potential threats or opportunities based on the company's activities, allowing them to adjust their strategies to prosper and remain competitive in the market (Sewlal, 2004). This business network can also be used by banks and investors to analyze the business relationships of their clients and investees, in order to assess the risks of making a loan or an investment, therefore maximize any gains while minimizing potential losses (Yan et al., 2019; Zuo et al., 2017).

In 2019, the Geotrend RE component was initially based on manually designed regular expressions created by domain experts. Writing these rules and adapting them to new languages, on the other hand, is costly. Furthermore, these relationships can be expressed

indirectly or metaphorically in text, making rule-based extraction even more difficult. This is illustrated in the sentence in (2) which expresses a `compete_with` relation between the entities *Delphi Automotive* and *Volkswagen* using the expression *has issued Autonomous Vehicle Testing Permits to*, implying that they are all autonomous vehicle manufacturers.

(2) Wheego and Valeo now join the likes of Google, Tesla, GM Cruise and Ford on the list of companies the Californian DMV has issued Autonomous Vehicle Testing Permits to, as well as Volkswagen, Mercedes Benz, Delphi Automotive and Bosch.

Despite their strategic importance, business relations extraction has received less attention in the literature compared to other specific-domains, such as the biomedical (Bunescu et al., 2005; Krallinger et al., 2017; Lee et al., 2013; Segura-Bedmar et al., 2013; Van Mulligen et al., 2012; Wu et al., 2019) and scientific domains (Bruches et al., 2020; Buscaldi et al., 2018; Luan et al., 2018a; Ma et al., 2022). Zhao et al. (2010) presented the first work in this direction aiming at identifying the taxonomy of business relations to be extract between companies, persons, dates, locations, etc. and its application in the context of CI. Few papers were published in the years that followed, with a typical industrial focus, where the pool of targeted relations is mostly limited to competition and cooperation interactions between organizations (Lau and Zhang, 2011; Yamamoto et al., 2017). Recently, various research workshops have been dedicated to analyzing financial information on the web, thereby encouraging advances in this field (e.g., Financial Technology and Natural Language Processing Workshop,[2] Knowledge Discovery from Unstructured Data in Financial Services Workshop, and Financial Narrative Processing Workshop[3]).

The objective of this thesis, carried out under the terms of a CIFRE contract between the IRIT laboratory in Toulouse and Geotrend, is to propose new supervised learning approaches based on modern deep learning architectures to detect business relations in a multilingual context, improving therefore the initial Geotrend rule-based RE component.

Business Relation Extraction as an NLP task

So far, various approaches have been proposed in the literature to extract relations between entities. The first ones, *pattern-based*, relied on the manual definition of patterns that identify the type of semantic relations between entities in text based on various lexico-syntaxic linguistic patterns used to express a given type of relation (Akbik and Broß, 2009; Aussenac-

[2]https://aclanthology.org/venues/finnlp/
[3]https://aclanthology.org/venues/fnp/

Gilles and Jacques, 2008; Batista et al., 2015; Hearst, 1992; Snow et al., 2004; Suchanek et al., 2006). However, this approach has a lower recall and requires human expertise to create these patterns as well as to adapt them to new domains.

To overcome these limitations, *supervised approaches* based on machine learning algorithms were proposed, mainly due to the increasing volume of textual corpora (especially on the web) that can be used as training data after being annotated by domain experts. *Feature-based* (Kambhatla, 2004; Nguyen et al., 2007b; Zhou et al., 2005) and *kernel-based* (Collins and Duffy, 2001; Culotta and Sorensen, 2004; Mooney and Bunescu, 2006) are among the first approaches where dedicated lexical, semantic, and syntactic features representing a training sentence are manually designed then fed into a classification algorithm that learns to predict the type of the relation linking two previously identified entities. Although these approaches are noticeably more efficient, choosing the sub-optimal set of representative features is not an easy task. In addition, annotating training data comes at a high cost whenever a new domain is tackled.

To reduce the tedious phase of identifying the most relevant features, *neural models* (in particular Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)) have been proposed to automate feature extraction. In these models, sentences are represented by static semantic vectors known as *word embeddings*, which are computed on large corpora to learn word representations from their various contexts (Mikolov et al., 2013b; Pennington et al., 2014). Neural architectures are trained on these vectors to automatically extract features and predict the type of relation expressed in the input sentence. Recently, *transformer* architectures (Vaswani et al., 2017) based on multi-head self-attention mechanisms have resulted in contextualized word representations generated by pre-trained language models on large-scale text corpora (Devlin et al., 2019), achieving maximum scores on very well-known RE datasets (Tao et al., 2019; Wu and He, 2019). These architectures (and their associated performances) have further been improved by injecting knowledge about the target entities as given by external linguistic resources and additional pre-training tasks (Wang et al., 2019; Yamada et al., 2020b; Zhang et al., 2019).

In this study, we experimented with various new transformer-based architectures while evaluating their perfomances on a new multilingual dataset for business relation extraction.

Why Business Relation Extraction is a Difficult Task?

In the case of business relations, most existing methods rely on manually or automatically generated patterns which are hard to maintain (Braun et al., 2018; Burdick et al., 2015; Lau and Zhang, 2011). Supervised approaches were recently proposed (Collovini et al., 2020;

De Los Reyes et al., 2021; Yamamoto et al., 2017; Yan et al., 2019) but the datasets used in these studies present the following shortcomings:

- They focus all on a single language. However, due to a lack of multilingual models, a large amount of business and financial textual information generated online in various languages is difficult to exploit automatically by professionals;
- They are either small or not always available to the research community;
- They are annotated using various annotation schemes, making it difficult to compare various works carried out by different researchers in this context.

Furthermore, because supervised approaches have a limited set of targeted relations, models that extract relations from the open web suffer from a scarcity of positive relations of interest. In the context of business relations, for example, every two companies mentioned in a single sentence are not necessarily linked with a semantic business relation, as shown in (3). This sentence expresses a negative relation (i.e., `None`) between the two companies, *Intel* and *Tesla*. At the same time, the relation `acquired_by` exists between *Intel* and *Mobileye*, and `partner_with` between *Tesla* and *Mobileye*. This causes a data imbalance problem, which hinders the learning of trained models.

(3) Mobileye was acquired by Intel in 2017 for 15.3 billion U.S. dollars. This Israeli vision company was also a partner of Tesla, to have the first generation of Autopilot.

In this dissertation, we aim to bridge the gap by proposing solutions to each of the aforementioned shortcomings.

Research Questions and Contributions

To explore business relations extraction, our research can be formed into the following research questions (RQ).

($RQ1$) How business relations are characterized and annotated in multilingual textual content?

($RQ2$) Can training a single RE model on multilingual data outperform training multiple single models on monolingual data?

($RQ3$) How can a RE model handle data imbalance between business vs. non business relations?

(*RQ*4) Can injecting factual knowledge about entities into a RE model at different levels of granularity improve its performances?

(*RQ*5) How can the results of our research be used in a market intelligence real-time application?

Based on the above research questions, the main contributions (C) of this dissertation are summarized as follows:

(C1) A unified characterization for business relations between *Organizations*, based on a taxonomy composed of five relations, namely: INVESTMENT, COOPERATION, SALE-PURCHASE, COMPETITION, LEGAL PROCEEDINGS, and a negative relation OTHERS which groups the remaining non-targeted types of relations;

(C2) A multilingual manually annotated business RE dataset annotated using this characterization in four languages: *French*, *Spanish*, *English*, and *Chinese*. Part of the dataset is available to the research community.[4] As far as we know, this is the first multilingual dataset in this field, as all previously proposed datasets focused on a single language at a time (Khaldi et al., 2022c).

(C3) A set of Bert-like models for multilingual business RE relying on both monolingual and multilingual pre-trained language models (Khaldi et al., 2022c).

(C4) An empirical evaluation of various data-level and model-level approaches to tackle the problem of data imbalance between business and non-business (i.e., negative) relations (Khaldi et al., 2022a,b). We investigate in particular three new solutions: data augmentation using sentence similarity, multitask relation extraction, and binary soft labels generated by knowledge distillation to supervise RE.

(C5) An empirical evaluation of the impact of integrating different sources of knowledge about entities into the RE model, at different levels of granularity (Khaldi et al., 2020, 2021), going beyond recent studies which focused on a single source of knowledge (Papaluca et al., 2022; Poerner et al., 2020; Zhang et al., 2019).

(C6) The integration of our models into the Geotrend Market Intelligence platform showing that a multilingual entity-informed business relation extraction that handles data imbalance is crucial in an industrial context.

[4]https://github.com/Geotrend-research/business-relation-dataset

Dissertation Outline

The dissertation is organized in six chapters. The first two present an overview of the state of the art in binary relation extraction, while the four others focus on one of the aforementioned contributions.

In Chapter 1, we introduce the task of *generic relation extraction* at the sentence level. We begin by defining the key related concepts and the overall RE pipeline. We then present the main existing manually annotated datasets for RE, along with a description of how they were constructed and a quantitative characterization of their annotated relation instances. We finally focus on supervised approaches for RE, with a particular emphasis on neural and transformer-based approaches, which serve as a basis for our work.

Chapter 2 continues state of the art, focusing this time on three *domain-specific relationships*, namely the biomedical, scientific, and business domains, which have sparked significant interest in the research community. For each domain, we go over the main proposed datasets and approaches trained on these datasets, as well as some applications that exploit the extracted relations in different deployed systems. We end this chapter by highlighting the main contributions of this work.

Chapter 3 details the data collection process that we followed and characterizes the typology of business relations that has been used to annotate our multilingual dataset (i.e., *(RQ1)* and *(RQ2)*). We then present the experiments performed to detect business relations from multilingual content with various cross-lingual transfer settings, ranging from zero-shot to joint transfer (i.e., *(RQ3)*).

Chapter 4 addresses the issue of data imbalance in RE models (i.e., *(RQ3)*), specifically the imbalance between the negative relation and positive relations of interest. We begin by providing a broad overview of existing data and model-level approaches for data imbalance in NLP. We then present the three approaches we newly propose namely: data augmentation based on sentence similarity, multitasking the identification and classification of relations to improve their extraction, and finally using binary soft labels generated through knowledge distillation to supervise relation extraction.

We attempt to answer our fourth research question *(RQ4)* in Chapter 5. We begin by providing an overview of knowledge enhanced pre-trained language models, highlighting their key features. We then present our proposed architecture that injects multiple sources of knowledge about target entities into a RE model at multiple levels. The experiments conducted to investigate the importance of each level of knowledge are then presented.

In Chapter 6, we describe the Geotrend platform, a market intelligence industrial pipeline for extracting business relations. We present the overall architecture of this pipeline and

detail how the models proposed in this thesis have been integrated into it (i.e., *(RQ5)*).

Finally, we conclude by providing an overview of this work, emphasizing its contributions and limitations. We also highlight our perspectives for future work.

Chapter 1

Generic Relation Extraction

The first attempt in extracting structured information from texts dates back to the 1990s, at the Message Understanding Conferences (MUC), where several evaluations of Information Extraction (IE) tasks were organized for conference participants to extract specific information about business and defense-related activities from news articles (MUC, 1991, 1992, 1993).

MUC evaluations were designed as a template filling task in which participants used pattern matching techniques based on lexical and syntactic analysis of input text to fill template slots with event information such as event type, event agent, event time and place, effect, and so on. Entity extraction was not introduced as a domain independent task until MUC-6 (Grishman and Sundheim, 1996), which aimed to identify named entities such as persons, organizations, and locations, or numeric entities such as time, date, currencies, and percentages. MUC-7 (Chinchor, 1998) later added a relation extraction (RE) task to identify relationships between these entities.

Roughly, IE can be divided into three main subtasks: Entity Extraction, Relationship Extraction, and Event Extraction, defined as follows:

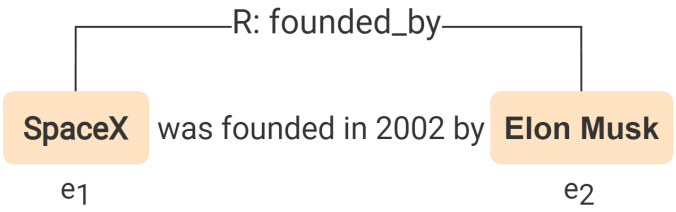

Figure 1.1 Example of *founded_by* relation type.

- **Entity Extraction.** This task aims to locate and categorize (extract) a sequence of tokens referring either to named entities such as person (e.g., *Abu Bakr*), organization

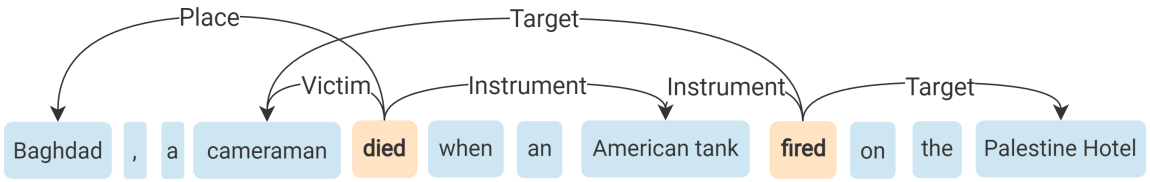

Figure 1.2 Event types, triggers, arguments, and their roles.

(e.g., *Google*), etc. or to concepts which are nominal referring to a group of individuals such as *Knowledge Management*, *Science*, etc. (Sekine, 2004; Yadav and Bethard, 2018).

- **Relation Extraction.** This task concerns the extraction of semantic links expressed between entities in text (Aydar et al., 2020; Bach and Badaskar, 2007; Bassignana and Plank, 2022b; Han et al., 2020; Pawar et al., 2017; Smirnova and Cudré-Mauroux, 2018; Wang et al., 2022). Figure 1.1 depicts an example of *founded_by* relation at the sentence level, between two entities, the company *SpaceX* and the person *Elon Musk*, triggered by the verb *was founded by*.

- **Event Extraction**. An event is a change of state happening at a certain time, in a given place, involving one or more participants (Doddington et al., 2004). Event extraction aims at identifying event information from unstructured plain texts by classifying the event type usually represented by the type of the event trigger, identifying its arguments, and judging the arguments' roles (Li et al., 2021). For example, in Figure 1.2, two types of event are expressed: *Die* triggered by the verb *died* with three argument roles of *Place*, *Victim*, and *Instrument* and the event *Attack* triggered by the verb *fired* with three argument roles of *Place*, *Target*, and *Instrument*.

We begin this chapter by defining the task of relation extraction as well as the related task of entity extraction. We then provide an overview of available datasets and their construction methods. We also go over prior works for extracting relations from text, and review proposed supervised models, including traditional and neural ones. We finally report the metrics used to evaluate them.

1.1 What is Relation Extraction?

Relation extraction (RE) is a subtask of information extraction (IE) that aims at discovering *semantic relationships* between at least two entity mentions in unstructured natural language texts (Culotta et al., 2006). Semantic relations are important because they connect entities

in a text, and together with entities they make up a good chunk of the meaning of that text. The extracted relations play a crucial role in many NLP applications such as information extraction (Wu and Weld, 2010), search engines (Schlichtkrull et al., 2018; Xiong et al., 2017), recommendation systems (Betancourt and Ilarri, 2020), question answering (Bordes et al., 2014; Dai et al., 2016; Dong et al., 2015; Mohamed et al., 2017; Yao and Van Durme, 2014; Yih et al., 2015), and textual entailment (Androutsopoulos and Malakasiotis, 2010). Knowledge base (KB) population (also known as ontology population) is probably the main application of RE, where a relation – and the entities involved in it – are used to discover facts about entities and to augment a KB with these facts (Augenstein et al., 2016a; Cimiano, 2006; Ji and Grishman, 2011).

Formally, a relation is an n-ary tuple $R(e_1, \ldots, e_n)$ $(n \geq 2)$ of entities e_i, with a predicate term R denoting the type of relation. A relation can be oriented and expressed in only one way, from entity e_a to entity e_b, where $R(e_a, e_b) <> R(e_b, e_a)$, or non-oriented where $R(e_a, e_b) = R(e_b, e_a)$. For example, from the sentence in (1), the relation `aerial-bombardment` is extracted along with four entities.

(1) On Thursday, there was a massive U.S. aerial bombardment in which more than 300 Tomahawk cruise missiles rained down on Baghdad.
`aerial-bombardment (U.S., Baghdad, Tomahawk cruise missiles, Thursday)`

RE focuses on the extraction and/or linking of two elements: entities and relations. Extraction involves identifying textual mentions of entities/relations, while linking involves associating each of such mentions with an appropriate disambiguated identifier referring to the same element in a Semantic Web KB (or ontology) (Martinez-Rodriguez et al., 2020). We provide in the next two sections the basis behind extracting and linking each of these two elements.

1.1.1 Entity Extraction and Linking (EEL)

Relation extraction builds on top of entity extraction, also known as *entity recognition*. An entity can either be a *named entity* that can refer to a proper name or a *concept* which is a conceptual grouping of elements, set, or collection of entities.

Martinez-Rodriguez et al. (2020) defined a named entity as a person (e.g., *Bill Gates*), a location, or an organization (e.g., *Microsoft*). Jurafsky and Martin (2018) extended this definition by including temporal expressions (e.g., dates, times) and even numerical value (e.g., percent, currency). With the advancement of the IE field, entity types have been expanded to include domain-related entities such as proteins names (e.g., *Collagen*), chemicals names (e.g., *Sodium hydrogen carbonate*), and diseases names (e.g., *Brucellosis*) in the biomedical

domain (Eltyeb and Salim, 2014) but alos more fine-grained types such as city names (e.g., *Toulouse*), road names (e.g., *Road of Saint Simon*), or facilities names and monuments (e.g., *Empire State Building*).

Martinez-Rodriguez et al. (2020) further considered two types of concepts:

- Classes: that represent a named set of objects that share the same characteristics. For example, the class *Google CEOs* groups all CEOs of Google since the foundation of the company, such as: *Larry Page, Sundar Pichai*.
- Topics: that are categories to which objects relate. For example, the topic *Cancer* groups all objects that relate to it, such as *cancer, breast, doctor, chemotherapy*.

Relations between concepts aim to capture knowledge about the world, while relation between named entities describe particular events/situations expressed in texts. More formally, entities can be considered as atomic elements from the domain, while concepts as unary predicates. Entity extraction or entity recognition is then the task of locating and classifying entities in text into predefined categories of entities listed above (i.e., named entities, classes, and topics).

Before starting to extract relations, it is often good to proceed with entity resolution and linking (EEL), which aims to group words that refer to the same entity in text, then link them to a unique identifier in a KB. For example, in (2), the two terms *He* and *Steve Jobs* are related to the same real-world entity, which is *Steven Paul Jobs*, while the entity *Apple* to the real-world named entity *Apple Inc.* which corresponds to the unique Wikidata identifier *Q312*. In this example, EEL will provide additional information about the same person, which enable extracting the following triples: `CEO (Steven Paul Jobs, Apple)`, `Co-founder (Steven Paul Jobs, Apple)` and `Is-a (Steven Paul Jobs, American business magnate)`.

(2) Steven Paul Jobs was an American business magnate and investor. He was the chief executive officer (CEO) of Apple. Steve Jobs was also its co-founder.

EEL either relies on off-the-shelf named entity recognition tools that extract entities for limited numbers of types (e.g., persons, organizations, places) or on specific methods using entity labels in dedicated KBs (like Wikipedia, DBPedia, etc.) as a dictionary to guide the extraction. This is illustrated in Figure 1.3 that shows the output given by the online DBpedia Spotlight demo[1] when processing the sentence "*Bryan Cranston is an American actor*". The output, in JSON format, shows a selected identifier obtained from DBpedia

[1] https://www.dbpedia-spotlight.org/demo/

("@URI" attribute), the "@types" list matches classes from the KB, the "@surfaceForm" represents the text of the entity mention, and the "@offset" indicates the character position of the mention in the text.[2]


```
{
" @text ": " Bryan Cranston is an American actor
" Resources ": [
    { " @URI ": "http://dbpedia .org/resource /Bryan_Cranston ",
    " @types ": " DBpedia :Agent , Schema :Person , Http ://xmlns .com/foaf /0.1/ Person,
                DBpedia : Person",
    " @surfaceForm ": " Bryan Cranston ",
    " @offset ": "0" ,
    },
    { " @URI ": "http://dbpedia .org/resource/United_States ",
      " @types ": " Schema:Place, DBpedia :Place, DBpedia:PopulatedPlace,
                    Schema:Country , DBpedia:Country ",
      " @surfaceForm ": " American ",
      " @offset ": "21",
    }
    { " @URI ": "http://dbpedia.org/resource/Actor ",
      " @types ": "",
      " @surfaceForm ": " actor ",
      " @offset ": "30" ,
     }]
 }
```


Figure 1.3 An example of EEL output as given by DBpedia Spotlight (Martinez-Rodriguez et al., 2020)

1.1.2 Relation Extraction and Classification

Main Concepts

We define some concepts related to the relation extraction task that we will use in the remainder of this dissertation.

- **Entity type.** It is the category to which an entity belongs, such as *Organization*, *Person*, *Date*, *Location*, etc.
- **Relation type.** It refers to the type of the semantic link R between entities, for example: $R = employee\_of$, is a relation type to express that an entity of type *person* is an employee of an entity of type *organization*.
- **Relation candidate.** It is a sentence S, with a set of two tagged entities (e_1, e_2) (for binary relation) or more $(e_1, e_2, ..., e_n)$, that may be connected with a certain relation type (see next Section). We note it $(S, e_1, e_2, .., e_n)$.
- **Relation instance.** It is composed of a relation candidate $(S, e_1, e_2, .., e_n)$ and a relation type R linking these entities in a sentence. We write it down as $R(S, e_1, e_2, .., e_n)$.

[2]The output provides several other attributes that are not shown in the figure.

Relation Types

Bach and Badaskar (2007) identified two main types of relation based on the number of entities it links:

- **Binary relations** that link two entities (e.g., `husband-of (Barack Obama, Michelle Obama)`). Most of the research on relation extraction and classification focuses on binary relations only.
- **N-ary relations** that link three or more entities. They are good for verbs which can take multiple arguments (e.g., *sell*) or for event representation. Such relations can be expressed as frames. For example, a *selling* relation can invoke a frame covering relations between *a buyer, a seller, an object_bought* and *price_paid*. Here is another example of such n-ary relations, extracted from the sentence below (cf. (3)) between three entities: *conference, institution* and *location*.

  (3) CMU conference was organized by ACL at Pittsburgh.
  `organize-conference-at(CMU, ACL, Pittsburgh)`

Relations can be further classified according to the argument involved which gives rise to another useful distinction:

- **First-order relations** that connect two or more entities;
- **Higher-order relations** that link an entity with one or many relations, as in `believes (Mark, is-a (banana, fruit))` where `is-a (banana, fruit)` is a first order relation. Mapping sentences to hierarchical representations of their underlying meaning is a fundamental step towards natural language understanding (Kim et al., 2008; Liang et al., 2011; Lu et al., 2008; Raphael, 1964). Usually, such higher order relations are better expressed as conceptual graphs (Sowa, 1984).

For example, in (4), `Is-a(SnowBall, method for RE)` is a first order relation whereas `Proposed (author, Is-a(SnowBall, method for RE))` is a higher order relation.

(4) The author proposed SnowBall, a new method for RE.

The extracted relations can either be part of a set of predefined relations or not. [3] In the

[3] There is no consensus on a comprehensive list of relations that can fit all purposes and all domains. See Ó Séaghdha (2007) for a discussion.

first case, RE (also called *targeted* RE) is similar to a classification task, where already identified entity mentions have to be linked by a set of known relations. [4] In the second case, RE (called *emergent* RE) does not assume a selected set of relations and tries to target all relations that can be extracted in an unsupervised or semi-supervised fashion. This idea is behind *Open Information Extraction* (OpenRE) "an extraction paradigm that tackles an unbounded number of relations" (Etzioni et al., 2008).

Finally, Pawar et al. (2017) further clarified the usage of the term RE which can refer to either relation expressed between two entities in a single sentence, across different sentences, or in a document:

- *Mention-level*, also known as *sentence level*, that takes as input the entity mentions, and the sentence which contains it, and tells if a relation exists between them and if yes, what's its type. Mention-level RE has been the focus of the Relation Detection and Characterization task at the Automatic Content Extraction evaluation campaign (ACE[5]) (Doddington et al., 2004).

- *Global level*, also known as *document level*, where the system takes a large text as input, and produces as output a list of entity linked by a certain semantic relation. This is more difficult and open-ended than the task of Relation Classification. For example, applications in the field of semantic web and ontology building require extraction of all possible relations without knowledge of the entities of interest (Xu et al., 2021, 2022).

We focus in the reminder of this chapter (and dissertation) on targeted binary RE that may hold in single sentences, assuming a set of predefined relations and the knowledge about boundaries and types of entity mentions known before hand. Here, entity mentions refer to named entities (excluding concept mentions).

1.1.3 Relation Extraction Pipeline

Hachey (2009) described relation extraction as a pipeline that includes two main sub-tasks: *relation identification* that concerns the identification of entity pairs linked by a semantic relation, and *relation characterization* that determines the type of the identified link.

Let's illustrate this pipeline, assuming the sentence in (5) as an input (cf. Figure 1.4).

[4] A prior step, before predicting the more suited relation that may hold between entity mentions, is to predict whether they are linked by a relation or not. In practice, these two tasks are often combined by making a multi-class classification problem with an extra `NoRelation` class. More details about these steps are given in the next sections of this Chapter.

[5] http://www.ldc.upenn.edu/Projects/ACE

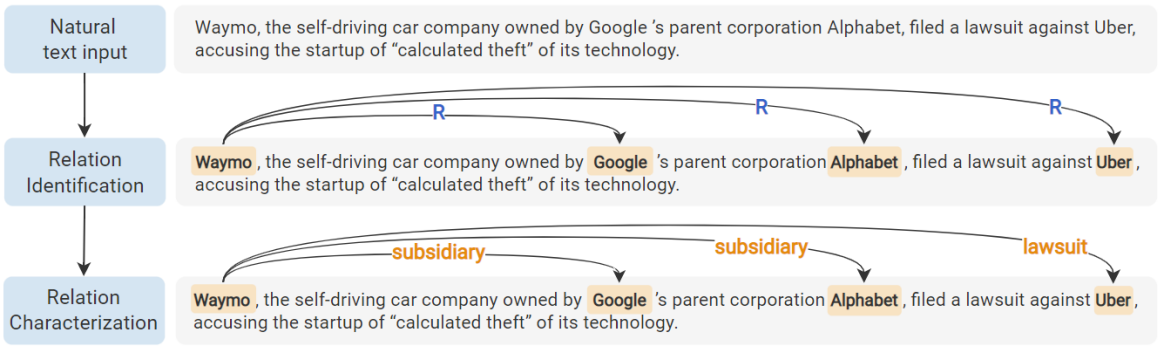

Figure 1.4 Illustration of relation extraction pipeline as proposed in (Hachey, 2009).

(5) Waymo, the self-driving car company owned by Google's parent corporation Alphabet, filed a lawsuit against Uber, accusing the startup of "calculated theft" of its technology.

The pipeline starts by feeding up unstructured textual data into the relation identification module which identifies pairs of entity mentions in sentences that can be arguments for relation instances: `(Waymo, Google)`, `(Waymo, Alphabet)`, and `(Waymo, Uber)`. The remainder of the entity pairs such as `(Waymo, Google)`, `(Google, Uber)` are not identified since no semantic relation is expressed between them in the input sentence. Next, the relation characterization annotates the relation candidate with a label describing the relation type. In this example, the (*Subsidiary*) label describes the relation between *Waymo* and *Alphabet*, and between *Google* and *Alphabet* while the (*Lawsuit*) label describes the relation between *Uber* and *Waymo*). Finally, the extracted information is represented in a structured format as follows: `Subsidiary (Waymo, Alphabet)`, `Subsidiary (Google, Alphabet)`, and `Lawsuit (Waymo, Ubert)`.

Most RE systems treat entity extraction and relation extraction tasks separately: first they identify the entity mentions, then they classify the possible relation that may connect them. Entities may be extracted and linked before or after relations are extracted. Martinez-Rodriguez et al. (2020) observed that pre-entity relation processing can help to filter out sentences that do not involve relationships between known entities, while post-entity processing can help extract more complex relations (e.g., n-ary) using traditional RE or Open IE tools that can identify entities that are not supported by EEL. Recent studies show that joining the two sub-tasks is important for high performance, since relations interact closely with entity information (Finkel et al., 2006; McCallum and Jensen, 2003; Miwa and Bansal, 2016).

When a semantic link of interest is expressed between two given entities in a sentence, the relation type assigned to this candidate is called a *positive relation type* (R^+). However, identifying a positive relation for a relation candidate is not always possible. A *negative relation* (R^-) is assigned in this scenario. R^- can refer to the absence of a semantic link between two entities – in other words, the expressed relation is none of the ones to extract (usually named *Others*, referring to any other relation types) – or to the complete absence of any semantic link between the relation candidates (often referred to as *None* or *no_relation*). As an example, the entity pair `(Google, Uber)` in (5) is linked with a negative relation because no semantic link is expressed between the two entities.

In this dissertation, the relation extraction task is performed independently of the entity extraction task, where the RE task takes as input a relation candidate consisting of a sentence with a pre-identified pair of entities. We therefore redefine the RE pipeline initially proposed by (Hachey, 2009) by the pipeline in Figure 1.5 where a RE task is divided into two sub-tasks (Ye et al., 2019): relation identification and relation classification:

- **Relation identification.** This task entails distinguishing between relation candidates linked by a positive relation R^+ and those linked by a negative relation R^-.
- **Relation classification.** Positive relation candidates are classified into one of the positive relation types R_i^+ of interest that connect target entities.

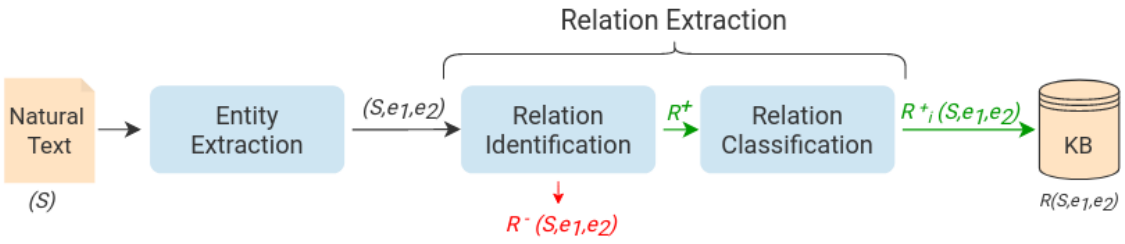

Figure 1.5 Overview of relation extraction sub-tasks

In the literature, these two tasks are typically combined into a single task known as *relation extraction* or *relation classification*. It is intended to be a single classification problem with the goal of identifying the relation type of relation candidate, where the relation type can be either a negative relation type or one of the positive relations of interest. In this dissertation, the expressions "relation extraction" and "relation classification" are used interchangeably to refer to the combination of relation identification and relation classification as a single task.

1.1.4 Automatic Relation Extraction: Main Approaches

The first proposed approaches for RE relied on manually or automatically generated linguistic patterns that are matched against documents to identify and extract relevant information in a structured format (Hearst, 1992; Snow et al., 2004). With the advance of machine learning (ML) algorithms, and the abundance of textual data online, it was important to automate this extraction process. ML models which can learn from historical data to make predictions on new data were then used. Depending on the type of data used and the way they were generated, RE models can be classified into five types of approaches:

- **Pattern-based approaches** employ a set of lexical and syntactic rules derived from a small set of annotated data (Agichtein and Gravano, 2000; Brin, 1998; Hearst, 1992).
- **Supervised approaches** rely on models trained on manually annotated datasets (Dos Santos et al., 2015; Kambhatla, 2004; Wu and He, 2019; Zhang et al., 2015a).
- **Distantly supervised approaches** leverage on knowledge bases to annotate raw data; annotated data is later used to train ML models (Augenstein et al., 2016b; Kamel et al., 2017b; Lin et al., 2021; Mintz et al., 2009; Quirk and Poon, 2017; Smirnova and Cudré-Mauroux, 2018; Sui et al., 2021; Zhang et al., 2021).
- **Unsupervised approaches** rely on non-annotated data to train a RE model (Fader et al., 2011; Gashteovski et al., 2021; Kolluru et al., 2022; Léchelle et al., 2019; Mausam, 2016; Stanovsky and Dagan, 2016; Wang et al., 2021a).

Supervised approaches are the main focus of this dissertation, given the quality of the generated data through the annotation process. Each entry of the training dataset is a relation instance $R(S, e_1, e_2)$, where e_1 and e_2 are the target entities, and S is the sentence expressing the relation R to be predicted between those entities. RE as a supervised task is generally cast into a multi-class classification problem, where a classifier learns to distinguish between representations of relation candidates while optimizing the distance between generated predictions and ground-truth labels (Liu et al., 2013; Zhang et al., 2015a, 2017b).

To obtain these candidate representations, different techniques were explored in the literature. Let f be the representation function that takes as input a relation candidate (S, e_1, e_2) and generates the relation representation vector v_r that is fed to a relation classifier to predict the relation type R.

$$f(S, e_1, e_2) = v_r \tag{1.1}$$

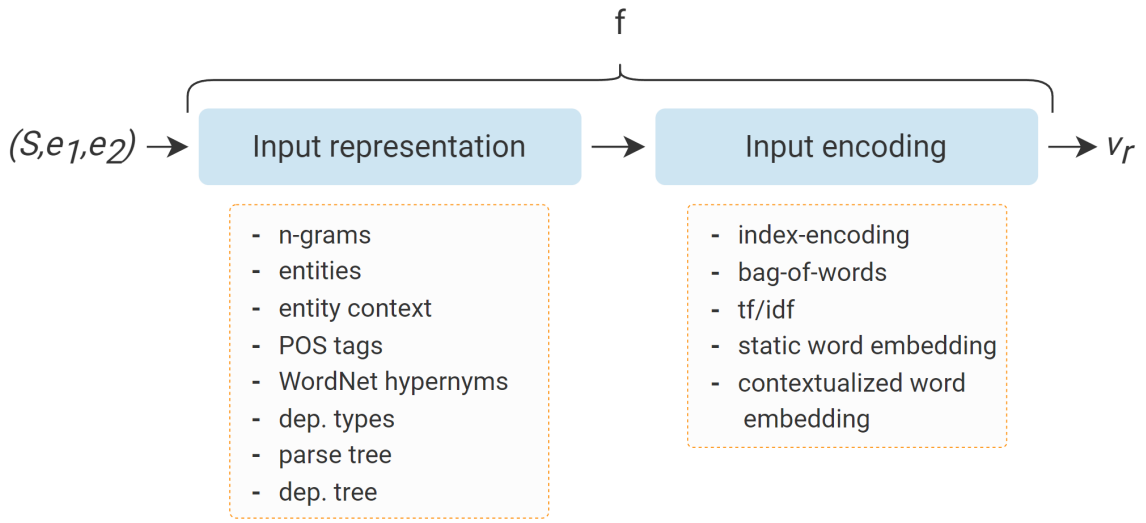

Figure 1.6 Input representation function.

Supervised approaches are classified into three types of methods based on the nature of the representation function f and the relation representation vector v_r :

- **Feature-based methods** in which f is designed manually followed by an encoding phase in which features are encoded into v_r using conventional NLP techniques such as TF-IDF, bag-of-words, and so on (cf. Section 1.3.2).
- **Kernel-based methods** where f relies on a set of parsers, taggers, and gazetteers to generate rich structural representations and encode them into v_r using conventional NLP techniques such as TF-IDF, bag-of-words, and so on (cf. Section 1.3.3).
- **Neural methods** where f is a superposition of a mapping function that maps each word in the input to a semantic dense space, and an extraction function that uses neural architectures to automatically extract the features on top of these representations to generate v_r (cf. Section 1.4).

The general framework for relation representation is described in Figure 1.6, highlighting possible input representations such as n-grams, parse trees, and WordNet hypernyms of words or entities; as well as different methods to encode these representations including: TF/IDF, static word embeddings (cf. Section 1.4), and contextualized word embeddings (cf. Section 1.4.5).

Most approaches are usually evaluated using existing annotated datasets, making them therefore dependent on the relation schemata used to annotate these datasets. Before delving

into these methods, we present in the next section the most important existing datasets for binary targeted RE task. We conclude this chapter by discussing various model architectures and features that can be used to represent a relation candidate.

1.2 Annotated Datasets For Binary Relation Extraction

1.2.1 Dataset Construction

The availability of textual data has contributed to the advancement of various NLP tasks, including RE, and has encouraged the development of machine learning algorithms that have significantly advanced the field. To train a RE model, annotated data is required. As a result, several methods for creating relation extraction datasets, ranging from entirely manual, semi-automated, to fully automatic, have been developed.

Automatic Construction. Methods like machine translation or parallel data exploitation were used to generate multilingual training data starting from an annotated dataset in one language (Faruqui and Kumar, 2015; Seganti et al., 2021; Yanan, 2013; Zou et al., 2018). The quality of the generated data depends on the performances and the availability of external resources and machine translation systems, which is not straightforward for many languages and domains.

Semi-automatic Construction. Distant supervision for relation extraction was designed to reduce the need of labeled data and domain dependency of existing RE datasets. These methods aim to predict semantic relations between pairs of entities of any domain while being supervised by Knowledge Bases (KBs). They heuristically align entities in texts to entity labels from a given KB and use this alignment to learn a relation classifier. The training data are labelled automatically as follows: for a triplet $R(e_1, e_2)$ in the KB, all sentences that mention both entities e_1 and e_2 are regarded as the training instances of the relation R (Ji et al., 2017).

Distant supervision was first introduced by Mintz et al. (2009) who used a large semantic database along with a large unlabeled corpus instead of a manually annotated corpus for supervision. They used Freebase (Rouces et al., 2017), a KB that contains more than 7,300 relations between 9 million named entities. Their approach achieved a precision of 67.6% but the idea gained popularity quickly. Later on, other approaches with various improvements over Mintz et al's basic distant supervision were proposed (Chen et al., 2014; Koch et al., 2014; Nagesh, 2015; Zhang et al., 2013). After deep learning-based methods

gained popularity for supervised learning of relation classification, researchers started using it in distant supervision as well (Lin et al., 2016; Qin et al., 2018a,b).

Distant supervision was first exclusively used to generate monolingual training data (Mandya et al., 2019; Nam et al., 2018; Norman et al., 2019; Riedel et al., 2010), then recently it has been used to generate data in a multilingual setting (Abdou et al., 2019; Bhartiya et al., 2022; Köksal and Özgür, 2020).

Distant supervision is an elegant solution to overcome the lack of training examples, but rises many complications. First, the availability of knowledge resources related to the domain of the studied relations and corpora is required, and significant errors in labels may occur leading to noisy training data which may hurt models precision (Riedel et al., 2010; Xie et al., 2021). Second, when multiple KBs are used, relation mentions may overlap. And even with a single KB, several relations may exist in this KB between a pair of entities, which makes the selection of the relation type hazardous. Moreover, a KB can be incomplete and should be interpreted under an Open World Assumption: if a relation is not present in the KB, it should not be considered as a negative example for training. Finally, distance supervision depends heavily on EEL outputs for entity recognition, where entity mentions can be linked to an incorrect KB identifier. Several approaches have been proposed to address this issue. See Martinez-Rodriguez et al. (2020) and Zhao et al. (2019) for an overview.

Manual Construction. Finally, manual data annotation has been used to generate monolingual or multilingual data (Hendrickx et al., 2010; Mitchell et al., 2005; Zhang et al., 2017b), relying on clear and well-defined annotation guidelines about relation and entity types (Han, 2010; LDC, 2004). To reduce human errors and biases that may occur during annotation and to produce a high-quality annotated dataset, the annotation is performed in an iteratively assessed process (Grosman et al., 2020), where inter-annotator agreement is evaluated using standard metrics, such as Cohen's coefficient (Cohen, 1960) or Fleiss's kappa (Fleiss, 1971).

The following section primarily focuses on datasets generated semi-automatically or manually.

1.2.2 Available Datasets

Training datasets have many important properties that may impact the extraction process, among which:

- Relation taxonomy: A useful distinction is between coarse-grained and fine-grained relation extraction. The number of possible relations can be infinite in the extreme, as

in the case of emergent relation extraction. Supervised learning is used to handle the targeted relation extraction which are usually coarse-grained, i.e., have a few number of possible relations.

- General vs. domain-specific: General datasets have a mixed bag of sources which are likely to be useful in processing all kinds of text or in representing knowledge in any domain (e.g., relations like `is-a`, `member-collection`, `possession`, `cause-effect`, `location`, `part-of`, etc.). On the other hand, domain-specific relations have a very homogeneous source and are only relevant to a specific text genre or to a narrow domain (e.g., `inhibits`, `activates`, `phosphorylates` for gene/protein events). News articles (Doddington et al., 2004; Hendrickx et al., 2010; Riedel et al., 2010; Zhang et al., 2017b) and Wikipedia pages (Köksal and Özgür, 2020; Lyu and Chen, 2021) are the main source of data for generic relations. Others rely on domain-specific documents such as scientific publications (Bunescu et al., 2005; Lee et al., 2017; Luo et al., 2022; Xing et al., 2020).

We review in this section publicly available datasets annotated for generic RE. Domain-specific RE (both used data and automatic approaches) will be detailed in the next Chapter.

Manually Annotated Datasets at the Sentence Level. After the MUC conferences, the program proposed a dataset collected from news articles and manually annotated to evaluate entity, relation, and event extraction tasks at the sentence level in three languages: English, Chinese, and Arabic (Doddington et al., 2004). This dataset served for almost a decade as a reference for evaluating relation extraction models. Figure 1.7 shows some examples of domain-independent relation types from the ACE2004 dataset. Then, in 2007 and 2010, new datasets for relation extraction between nominal were created in the context of SemEval shared tasks (Girju et al., 2007; Hendrickx et al., 2010), another challenge to compare scientific contributions to IE. The 2010 dataset became the new benchmark for relation extraction at the sentence level (cf. Figure 1.8 for relation types and their frequencies in this dataset).

the TACRED dataset (Zhang et al., 2017b) expresses relations between named entities covering 41 relation types and accounting for 106,264 examples. When compared to SemEval 2010 and ACE 2004, it is the largest manually annotated dateset (9 and 24 relations types for SemEval 2010 Task 8 and ACE respectively; and 10,000 instances for SemEval 2010 Task 8). It was recently built using texts from TAC Knowledge Base Population (TAC KBP) challenges corpora [6] extracted from newswire and web forums. Other versions of this dataset

[6] https://catalog.ldc.upenn.edu/LDC2018T03

ACE 2003		
Type	**Subtype**	**Count**
AT	based-in	496
	located	2879
	residence	395
NEAR	relative-location	288
PART	other	6
	part-of	1178
	subsidiary	366
ROLE	affiliate-partner	219
	citizen-of	450
	client	159
	founder	37
	general-staff	1507
	management	1559
	member	1404
	other	174
	owner	274
SOCIAL	associate	119
	grandparent	10
	other-personal	108
	other-professional	415
	other-relative	86
	parent	149
	sibling	23
	spouse	89

ACE 2004		
Type	**Subtype**	**Count**
PHYS	LOCATED	745
	NEAR	87
	PART-WHOLE	384
PER-SOC	BUSINESS	179
	FAMILY	130
	OTHER	56
EMP-ORG	EMPLOY-EXEC	503
	EMPLOY-STAFF	554
	EMPLOY-undetermined	79
	MEMBER-OF-GROUP	192
	SUBSIDIARY	209
	PARTNER	12
	OTHER	82
ART	USER/OWNER	200
	INVENTOR/MANUFACTURER	9
	OTHER	3
OTHER-AFF	ETHNIC	39
	IDEOLOGY	49
	OTHER	54
GPE-AFF	CITIZEN/RESIDENT	273
	BASED-IN	216
	OTHER	40
DISC	DISC	279

Figure 1.7 Relation taxonomy in ACE 2003 and ACE 2004 datasets, taken from (Pawar et al., 2017)

Relation	Freq
Cause-Effect	1331 (12.4%)
Component-Whole	1253 (11.7%)
Entity-Destination	1137 (10.6%)
Entity-Origin	974 (9.1%)
Product-Producer	948 (8.8%)
Member-Collection	923 (8.6%)
Message-Topic	895 (8.4%)
Content-Container	732 (6.8%)
Instrument-Agency	660 (6.2%)
Other	1864 (17.4%)
Total	10717 (100%)

Figure 1.8 Semantic relation typology with their frequencies in SemEval 2010 task 8, taken from (Hendrickx et al., 2010).

were later published by researchers in an effort to correct its flaws, such as incorrect instance labels or ambiguous relations in the annotation scheme (Alt et al., 2020; Stoica et al., 2021).

Datasets annotated by Distant Supervision Approach. NYT dataset (Riedel et al., 2010) was created based on Freebase as a knowledge source and New York Times articles as a source of data. Following the same method, other datasets were generated using the corpus

of TAC KBP challenges for supervision (Angeli et al., 2014; Zhang and Wang, 2015). T-REx (Elsahar et al., 2018), is the largest dataset, consisting of an alignment between free text documents from DBPedia abstracts and 11M KB triples from Wikidata. Recently, RELX-Distant (Köksal and Özgür, 2020), a multilingual dataset was created. It includes hundreds of thousands of sentences with relations from Wikipedia and Wikidata collected by distant supervision for five languages including: English, French, German, Spanish, and Turkish. According to Bhartiya et al. (2022), this dataset has some flaws, such as a low frequency of negative instances and a limit of one relation type per entity pair. As a result, they proposed DiS-ReX, which has over 1.5 million instances in four languages: English, German, Spanish, and French.

Manually Annotated Datasets at the Document Level. Some relations are expressed across many sentences for single entity pairs in a document. (Yao et al., 2019) proposed DocRED, a large-scale human-annotated document-level RE dataset constructed from Wikipedia and Wikidata, containing 132,375 entities and 56,354 relational facts annotated on 5,053 Wikipedia documents. This dataset was revisited by (Tan et al., 2022b) to investigate the causes and consequences of the massive false negative problem. By adding the missing relation triples back to the original DocRED, 4,053 documents were re-annotated. Re-DocRED is the name given to the newly generated dataset. DWIE by Zaporojets et al. (2021) is another document-level dataset, specifically designed for multitask IE (Named Entity Recognition, Co-reference Resolution, Relation Extraction, and Entity Linking). Recently, Yao et al. (2021) proposed CodRED the first human-annotated cross-document RE dataset allowing reasoning across multiple documents.

It should be noted that corpora built for practical applications inherently contain relation instances that are more difficult to extract than those expressed in RE benchmarck datasets, leading to a drop in the performance of RE models when evaluated on real world data. Cheng et al. (2021) proposed a case-oriented construction framework to create a Hard Case Relation Extraction Dataset in order to improve the robustness of RE models against such hard cases (HacRED). The proposed HacRED is composed of 65,225 relational facts annotated from 9,231 documents with sufficient and diverse hard cases. Notably, with a data quality score of 96% F1, HacRED is one of the largest Chinese document-level RE datasets.

Other Interesting Manually Annotated Datasets. As entity and relation extraction can be performed jointly, CONLL04 (Roth and Yih, 2004) was the first dataset created for this purpose. Lately, Seganti et al. (2021) proposed the SMiLER dataset consisting of 1.1 M annotated sentences, representing 36 relations, and 14 languages

To perform relation extraction with insufficient training instances, FewRel (Han et al., 2018b) dataset was created for few-shot relation classification. This dataset was improved to handle domain adaptation and new relations not existing in the predefined relation set (Gao et al., 2019).

1.2.3 Summary

To summarize, the availability of annotated datasets not only facilitates the development of powerful automated models for relation extraction, but also serves as benchmarks to unify their evaluation process and metrics, allowing new models to be compared to existing ones. The table 1.1 summarizes the main characteristics of the previously cited datasets. We consider both quantitative and qualitative comparison criteria in this table, as follows:

- Qualitative characteristics:
  - Source of data;
  - Granularity level of the relation (*gran.*);
  - Annotation method (*gold* for manual annotation, *DS* for distantly supervised);
  - Type of the relation extraction task (*NER+Rel* for joint entity-relation extraction, *Rel* for relation extraction, and *few-shot Rel* for few shot relation extraction);
  - Language concerned (*lang.*);
- Quantitative characteristics
  - Number of instances (*#inst.*);
  - Number of relations (*#rel.*);
  - Number of entities (*#ent.*);
  - Number of words (*#words.*);
  - Negative relation rate in the dataset (*%neg.*).

Table 1.1 Datasets for supervised, distantly supervised, few shot, or joint relation extraction from a sentence or a document.

dataset	source	gran.	annotation	type	lang.	#inst.	#rel.	#ent.	#words	% neg.
ACE03-04	news	sent.	gold	Rel	3 lang.	16,771	24	46,108	297k	-
SE10T8	web	sent.	gold	Rel	EN	10,717	19	21,434	205k	17.4
TACRED	news	sent.	gold	Rel	EN	106,264	42	29,943	1,823k	78.7
NYT10	news	sent.	DS	Rel	EN	742,748	58	69,063	21,457k	
T-REx	Wiki	sent.	DS	Rel	EN	11M	642	-	-	-
RELX-Distant	Wiki	sent.	DS	Rel	5 lang.	2M	37	-	-	-
DiS-ReX	Wiki	sent.	DS	Rel	4 lang.	1.5M	37	-	-	-
CONLL04	news	sent.	gold	NER+Rel	EN	1,700	5	-	-	-
SMiLER	wiki	sent.	DS	NER+Rel	14 lang.	1.1M	36	-	-	-
FewRel	Wiki	sent.	DS+gold	few-shot Rel	EN	70,000	100	72,124	1,397k	-
DocRED	Wiki	doc.	DS+gold	Rel	EN	63,427	96	132,375	1,002k	-
DWIE	news	doc.	gold	NER+Rel	EN	317,204	65	43,373	501,095	-
CoRED	news	doc.	DS+gold	Rel	EN	30,504	65	-	-	84.4
HacRED	news	doc.	DS	Rel	ZH	65,225	26	-	-	-

Overall, the majority of the datasets are in English, with only few initiatives addressing multiple languages annotated by distant supervision approach, like RELX-Distant (Köksal and Özgür, 2020) (English, French, German, Spanish, and Turkish), DiS-ReX (Bhartiya et al., 2022) (English, German, Spanish, and French), and SMiLER (Seganti et al., 2021) (14 languages). Given the richness of multilingual content in Wikipedia, the majority of them used it as a data source.

Furthermore, the manually annotated datasets are smaller in size than those generated by the distantly supervised method, due to the cost of manual annotation. The latter method, on the other hand, produces much higher quality data. Finally, the majority of the proposed datasets have been designed for a relation extraction task that is performed independently of the named entity recognition task. This is mainly intended to reduce the error propagation between the two tasks when performed together.

In the next sections, we focus on the methods that have been proposed for binary RE at the sentence level when evaluated on manually annotated datasets.

1.3 Traditional Methods

1.3.1 Pattern-based RE

The primary idea behind pattern-based techniques is to convert the linguistic feature space into lexical and syntactic patterns that can then be applied to natural language texts to extract relations. These patterns are created manually by analyzing a group of relation instances to discover the surface form of a certain relation type and design rules that can distinguish it from other types (Aussenac-Gilles and Jacques, 2008; Aussenac-Gilles and Séguéla, 2000; Fauconnier and Kamel, 2015; Jacques and Aussenac-Gilles, 2006; Séguéla, 1999).

Hearst (Hearst, 1992) used lexico-syntactic domain-independent patterns to extract hyponym relations between entities and could enrich the lexical database WordNet[7] with 152 new relations. The following example (cf. (6)) explains this approach.

(6) Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

Hearst (Hearst, 1992) suggested the following lexico-syntactic pattern :

$$NP_0 \textit{ such as } NP_1\{,NP_2\ldots,(and|or)NP_i\}, i \geq 1$$

that implies the following semantics :

[7] https://wordnet.princeton.edu/

$$\forall NP_i, i \geq 1, hyponym(NP_i, NP_0)$$

allowing us to infer :

$$hyponym(Gelidium, redalgae)$$

Other works used the dependency path between entities, defined by the parse tree or the dependency grammar, as a pattern to identify novel entity pairs to generalize these generated rules (Akbik and Broß, 2009; Snow et al., 2004). Suchanek et al. (2006) proposed that pattern matching can be extended by using deep linguistic structures rather than shallow text patterns. They trained a model that can distinguish between positive and negative patterns and used it to identify positive patterns to determine the type of relationship between two entities.

According to Konstantinova (2014), rule-based approaches may produce acceptable results if the primary goal is to quickly extract relations in well-defined linguistic domains. These patterns typically have a high precision but a low recall. Furthermore, creating these handcrafted rules and considering all possible ones takes a significant amount of time and energy. Small variations from these patterns in relation candidates can prevent the discovery of appropriate relationships. Their adaptation to extract new relationships or target new languages is extremely weak.

Automatically generated patterns can be used in a bootstrapping step to facilitate the construction of the training dataset. Methods using pattern-based bootstrapping are often referred to as *weakly supervised information extraction*. They use an initial small set of seeds or a set of hand-constructed extraction patterns to begin the training process. They operate over large corpora of unlabeled data to learn patterns that extract new relation instances. For example, for the relation `SubsidiaryOf`, the starting seed could be `(Waymo, Alphabet)`, `(GitHub, Microsoft)`. First, the algorithm collects sentences containing these two entity mentions from the large corpora. Then, it generates extraction patterns from them. At the end, the bootstrapping algorithm extracts new instances of this relation like `(Avanade, Microsoft)`, using the generated patterns. Three main systems have been developed following this approach: DIPRE (Brin, 1998), SnowBall (Agichtein and Gravano, 2000), and BREDS (Batista et al., 2015).

In DIPRE (Dual Iterative Pattern Relation Expansion) (Brin, 1998), the algorithm is based on two main principles:

- Given a good set of patterns, a good set of tuples (entity pairs following a certain relation type) can be found;
- Given a good set of tuples, a good set of patterns can be learned;

Here, a relation instance representation takes into consideration three string contexts: words before the first entity mention, words between the two mentions, words after the second entity mention. Extraction patterns are generated by grouping contexts based on exact string matching.

Error propagation and semantic drift are common issues in bootstrapping methods. Any error made in the initial stages leads to many mistakes further on. These problems significantly impact the accuracy of these methods. To overcome this problem, Agichtein and Gravano (2000) built SnowBall RE, where string contexts extended with entity types, are represented by word vectors using TF-IDF. The patterns are clustered based on the similarity degree between their three context vectors. To control the semantic drift, Snowball scores the patterns and ranks the extracted instances using a confidence function and discards the ones where the confidence degree is under a certain threshold.

A recent work by (Batista et al., 2015) has improved bootstrapping algorithms. He used word embeddings to represent relation extraction patterns. His system, called BREDS, obtained better performances compared to Snowball.

1.3.2 Feature-based RE

Given the labeled data, lexical, syntactic and semantic features can be manually extracted and used to train a classifier, generally a Support Vector Machine (SVM) classifier, to classify newly unseen relation instances (Fauconnier et al., 2015; Ghamnia et al., 2017; Kamel et al., 2017a). Because the performance of created RE systems is strongly reliant on the quality of the extracted features, numerous efforts have been made to investigate the efficiency of various feature sources. (see Table 1.2 for an overview of main features).

Kambhatla (2004) combined a range of lexical, syntactic, and semantic features extracted from the text and used a Maximum Entropy (maxEnt) model for classification. The features used can be related to the words in the phrase, the entity type, the context between the two entities, but they can also be retrieved from the dependency and parse trees of the input sentence. Their findings suggest that using a variety of information sources can improve recall and the overall F-measure. Zhou et al. (2005) also explored the incorporation of various lexical, syntactic, and semantic knowledge using SVM. They noticed that the most important information in complete parse trees for relation extraction is shallow and can be obtained by chunking. They have shown that more information from full parsing provides only limited additional enhancement. They also demonstrated how semantic information, such as WordNet and Name List, may be used to increase performance even further.

According to Nguyen et al. (2007b), the dependency path between the two entities in the dependency tree may be missing some important word indicators that hint at the relation

type. Thus, they turned this dependency path between two named entities in a sentence into a tree by including more paths between the secondary entity and the relation-related keywords. This core tree can accurately reflect a relationship between a given entity pair because it contains as many clues for this relationship as possible. On top of this tree, syntactic features were manually selected and used to train an SVM classifier.

Furthermore, Chan and Roth (2010) demonstrated that utilizing multiple external knowledge such as co-reference relations between entities, information about entities from Wikipedia, and the global hierarchy of relations provided coherent models and predictions.

Given the structured shape of the input sentence, it can be challenging to arrive at an adequate subset of features to describe a relation instance. To solve this issue, new methods for relation extraction based on kernel functions have been developed that rely on rich representations of the input data, such as dependency parse trees.

1.3.3 Kernel-based RE

Kernel approaches have been offered as an alternative to feature-based methods for indirectly exploring features in a high dimensional space by directly calculating the similarity between two objects using a kernel function. Kernel approaches, in particular, could be helpful at reducing the load of feature engineering for structured objects in NLP research. Instead of manually enumerating the properties of two discrete structured items, these approaches can estimate their similarity directly, utilizing the original representation of the objects (Culotta and Sorensen, 2004).

Sub-sequence kernel. Inspired by the string sub-sequence kernel (Lodhi et al., 2002), Mooney and Bunescu (2006) described a new sequence kernel method for relation extraction systems by considering the sequence of words around the named entities as features, as follows:

- the set of all words, the set of all POS tags (e.g., NNP, NN, etc.):
- the set of all generalized POS tags (e.g., NOUN, VERB, ADJ, etc.);
- and the set of entity types (e.g., PER, ORG, LOC, GPE, etc.).

This work was generalized in (Mooney and Bunescu, 2006) that proposed a kernel combining three sub-kernels based on the similarity of different contexts of entity pairs (the before, middle and after entity pairs). The system gave better results when compared to the

Table 1.2 Main lexical, syntactic and semantic features used in supervised RE.

Type of features	References	Features
Lexical	(Kambhatla, 2004) (Zhou et al., 2005)	— words before and after mentioned entities; — words between mentioned entities; — existence of selected headwords related to a relation/entities; — nb. words separating the mentioned entities;
Syntactic	(Kambhatla, 2004) (Zhou et al., 2005) (Nguyen et al., 2007b) (Nguyen et al., 2007a)	— entities type — flags indicating whether the two mentioned entities are in the same NP, VP or PP; — paths and nb. paths in dependency tree between entities; — Words, POS and chunk labels of words on which the mentioned entities are dependent in the dependency tree; — shortest path between entities in dependency tree; — shortest path between entities and headwords in dependency tree; — paths in the parse tree between entities, that passes by headwords;
Semantic	(Zhou et al., 2005) (Chan and Roth, 2010)	- list of words from WordNet to differentiate between relation types; — co-reference relations between entities.

existing rule-based systems.

Syntactic tree kernel. Syntactic trees have been used in kernel-based methods because they encapsulate the structural properties of a sentence in terms of constituents such as noun phrases (NP), verb phrases (VP), prepositional phrases (PP), and POS tags (NN, VB, IN). Collins and Duffy (2001) presented a kernel function: convolution parse tree kernel, which computes similarity between any syntactic tree by counting the number of shared subtrees between them. Zhang et al. (2006) proposed five different possible syntactic tree representations for a given relation instance to find the most efficient subtree in syntactic parse trees. The best performance was achieved by the subtree surrounded by the shortest path connecting two entities. Later, (Sun and Han, 2014) proposed that a set of discriminant features be used to enrich the nodes in a syntactic tree (like WordNet senses, context information, properties of entity mentions, etc.). To compute similarity between these types of syntactic trees, a feature-enriched tree kernel was designed.

Dependency tree kernel. In Culotta and Sorensen (2004), dependency trees were also investigated, and a kernel function to compute similarity between these rich representations was devised on the assumption that instances containing comparable relations will have similar substructures in their dependency trees. To improve the relation representation and include more informative context, Zhou et al. (2007) proposed a tree kernel with context-sensitive structured parse tree information.

Dependency Graph Path Kernel. Bunescu and Mooney (2005); Mooney and Bunescu (2006) claimed that the shortest path between two entity mentions in the dependency tree holds much information that can discriminate relation instances. Based on this observation, a new kernel has been proposed: dependency graph path kernel that captures similarity between the shortest dependency paths representing two relation instances. To avoid data sparsity, generalized paths are used where every node in the path is extended with additional information like: POS tags, general POS tags, entity types.

The above-mentioned representations, including sequences, syntactic parse trees, and dependency parse trees, were systematically evaluated by Jiang and Zhai (2007). According to the results, each representation is effective on its own, with the syntactic parse tree form being the most effective. Furthermore, integrating the three representations yields no benefit.

External knowledge or NLP resources are required for both feature-based and kernel-based supervised algorithms for relation extraction. These resources could be the cause of errors that spread to the relation extraction system. Furthermore, manually creating features

may not capture all the necessary information to represent a relation instance. Thus, neural-based models have been proposed to address the issue of feature engineering and lessen reliance on external parsers.

1.4 Neural methods

The performance of the previously described methods is heavily dependent on manually designed features. Furthermore, these features are frequently derived from the output of previously existing NLP systems, which propagates errors in the existing RE systems and reduces their performance. Deep learning methods have received a lot of attention in recent decades because of their ability to reduce reliance on external resources and reduce the number of hand-crafted features by learning them automatically from continuous representations of words in a semantic space.

These representations are known as word embeddings. The basic idea behind learning them is the distributional hypothesis (Harris, 1954), which states that words that occur in the same contexts tend to have similar meanings. To learn them, the following unsupervised or self-supervised machine and deep learning algorithms were proposed : Word2vec (Mikolov et al., 2013b), which either predicts neighboring words (CBOW) or the focus word (Skip-Gram) in a context, and Glove (Pennington et al., 2014), which relies on local context information but also incorporates a global co-occurrence statistic to represent a word (cf. Figure 1.9). These algorithms can learn dense representations (i.e. dense vectors) of words by providing semantically similar words similar representations. Each word in the relation candidate's input sentence is mapped to its semantic space corresponding dense vector, then concatenated with other word representations and fed into the neural network alone or with other extracted features.

Different neural architectures have been proposed for RE, the most commonly used ones are described in the following sections, namely: Convolutional neural network (CNN), Recurrent neural network (RNN), Graph neural network (GNN), Transformers, and hybrid models that combine different architectures. Figure 1.10 depicts the general framework of neural based models. It is mainly composed of three modules:

- *Input representation module* that aims at representing the input sentence in a dense vector space using embedding vectors. Different types of embeddings were considered: embedding of words (WE), entities (NE), positions (PE), part-of-speech tags (POS), dependency tags (Dep), etc.

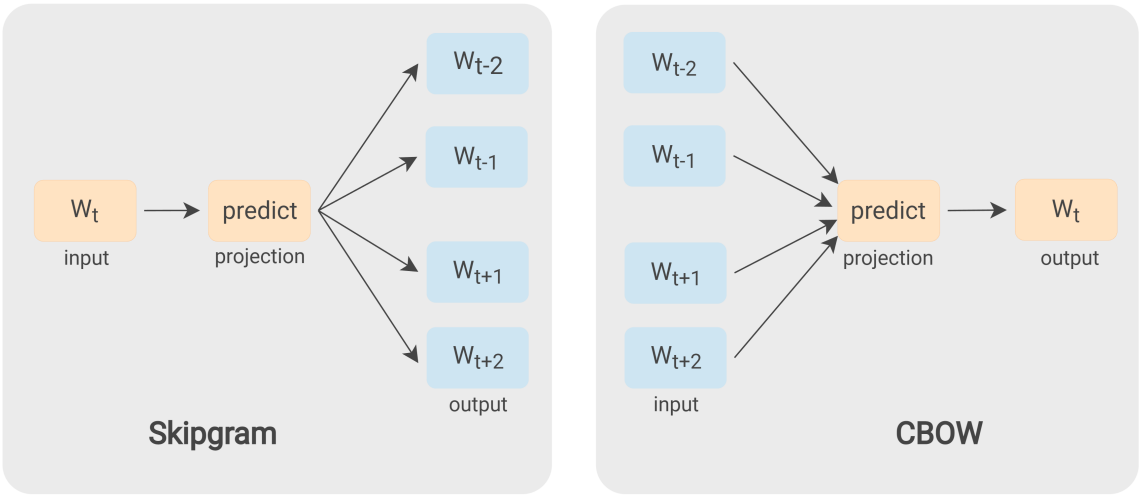

Figure 1.9 Word2vec algorithms to learn word embedding vectors (Mikolov et al., 2013a)

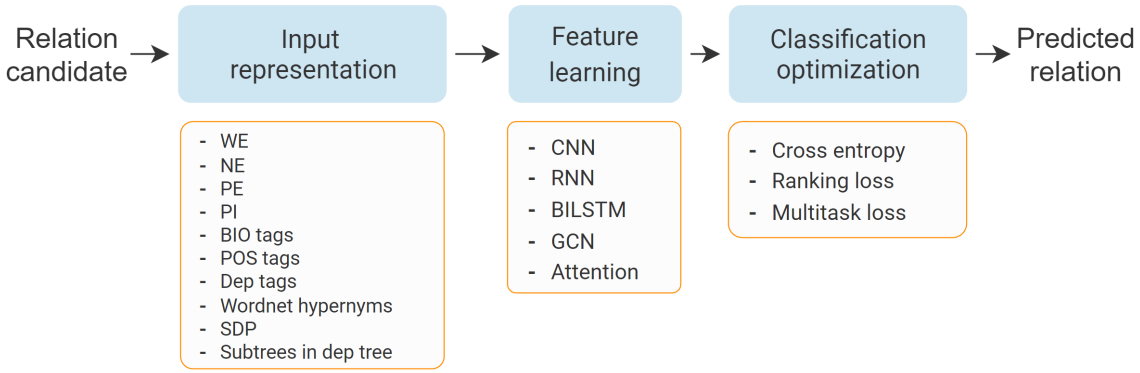

Figure 1.10 Generic neural-based models framework. WE:Word Embedding, PE: Position Embedding, PI: Position Indicators.

- *Feature learning module* that uses different neural architectures (e.g., CNN, RNN, GCN) to automatically extract relation representations on top of the considered input representations;
- *Classification optimization module* that consists in reducing the distance between relation representation and the ground truth relations using a loss function.

In the following sections, we will mainly focus on different architectures proposed to learn features.

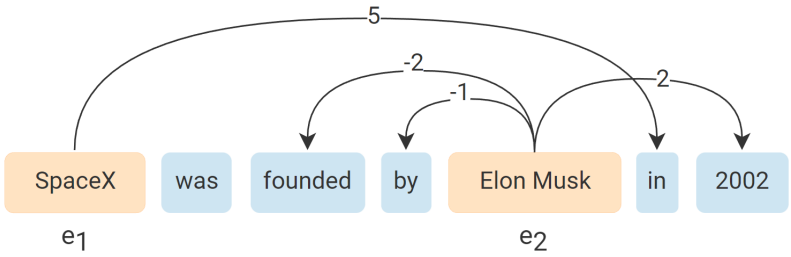

Figure 1.11 Relative position between words in a sentence and target entities.

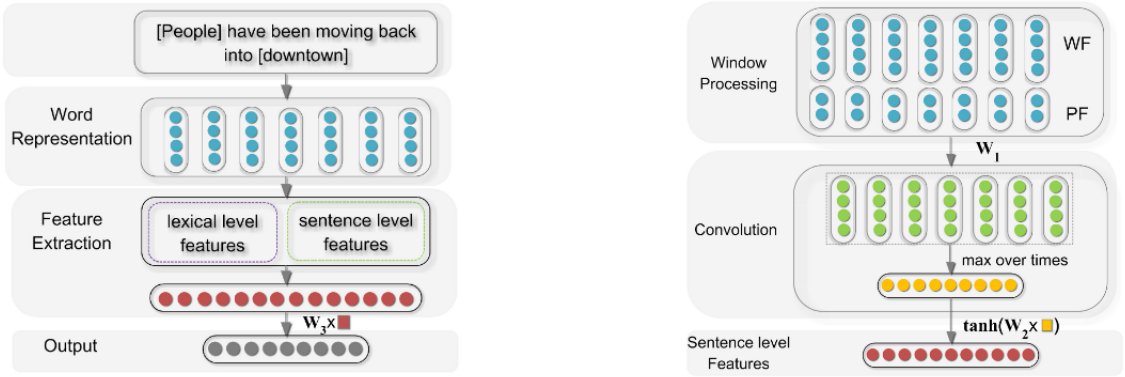

Figure 1.12 CNN architecture to extract sentence features (Zeng et al., 2014).

1.4.1 Convolutional Neural Network

CNN (LeCun et al., 1998) is a multi-layers architecture composed of convolutional layers, pooling layers and fully connected layers. The core component is a convolutional layer which generates a feature map from an input representation by multiplying it by a kernel matrix. The kernel does this by sliding step by step through every element in the input data. The generated feature map is smaller than the input vector and contains all the important local features (O'Shea and Nash, 2015).

Zeng et al. (2014) was the first work who used a CNN model for RE to learn sentence level features from word and position embeddings. Position embedding encodes the relative distance of each word to the target entities (cf. Figure 1.11). The extracted sentence features are combined with lexical level features, then fed into a softmax classifier to predict the relationship between two marked nouns (cf. Figure 1.12).

This model, however, can only learn short-distance patterns when using a kernel with small window size. To tackle this problem, Nguyen and Grishman (2015) proposed a CNN model with multiple window sizes for kernels, which allows learning patterns of different

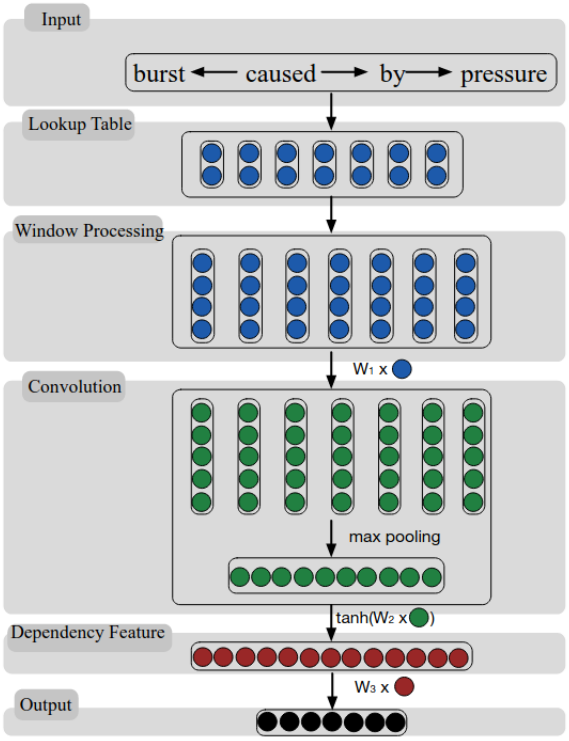

Figure 1.13 CNN architecture relying on an SDP, the *shortest dependency path between target entities* (Xu et al., 2015a).

lengths. Besides, Dos Santos et al. (2015) used a ranking loss to optimize a CNN model training to reduce the influence of the R^- class in RE (cf. Section 1.1.3.

Ye et al. (2019) also proposed to jointly perform relation identification with cross-entropy loss and relation classification with ranking loss to mitigate the negative relation impact on overall performance by learning features that distinguish negative instances from positive instances. Furthermore, as entities in the input sentence might bring semantic information about relation type, their model used BIO tags[8] embedding to highlight these entities, which are concatenated to word and position embeddings to represent a relation instance.

When the two target entities are separated by a long distance in the sentence, using all words to represent the relation instance may introduce a lot of irrelevant information. Syntactic features were thus introduced to capitalize on the dependency relationships between words in a sentence. For example, Xu et al. (2015a) used a CNN to learn more robust relation representations from the Shortest Dependency Path between target entities (SDP) (cf. Figure 1.13). Their findings support the effectiveness of dependency paths in implicitly representing

[8]short for (beginning, inside, outside), a common tagging format for tagging tokens in a chunking task such as named entity recognition.

the relative positions of target entities via path directions, as well as in determining the direction of the relation to extract via a negative sampling technique.

Because not all words contribute equally to expressing a relationship type, the **attention mechanism** was used to determine which parts of the sentence are most influential regarding the two entities of interest, and without relying on external resources. The attention mechanism for RE was first used by Wang et al. (2016). To better discern patterns in heterogeneous contexts, they proposed a novel CNN architecture based on two levels of attention. The first type of attention is entity-specific attention, which is applied at the input level to the target entities. The second attention is a relation-specific pooling attention, which is applied to features extracted from convolutional layers according to the target relations.

Finally, Shen and Huang (2016) combined a sentence convolution feature vector learned by a CNN using full word embedding, part-of-speech tag embedding, and position embedding; with an attention-based context vector obtained via a word-level attention mechanism using target entities vectors. Their model was successful in identifying the most influential parts of the sentence in relation to the two entities of interest, and could achieve a competitive performance just with minimal feature engineering.

CNN has the potential disadvantage of being unable to learn long-distance patterns in relation learning because it can only learn local patterns. Simply increasing the convolutional filter window size does not work: this reduces the power of CNNs in modeling local or short-term patterns. As a result, RNN-based RE models were introduced.

1.4.2 Recurrent Neural Network

RNN (Elman, 1990), as opposed to CNNs, are powerful sequential data models that can learn long-distance patterns between entity pairs. Because of their internal memory, these models allow the entire history of previous inputs to influence the network's outputs.

Zhang and Wang (2015) proposed a simple RNN-based framework for RE that learns relation representation from only word embeddings and does not include relative position embeddings. Given the sequential learning capability of RNN, only annotating the beginning and end of target entities can aid in implicitly learning the relative position of words.

When more context is needed to understand a relationship, RNN models fail to account for very long-distance patterns. This is known as a vanishing gradient problem, and it causes long-term contributions to be lost. Long Short-Term Memory (LSTM) networks were proposed (Hochreiter and Schmidhuber, 1997) to address this problem. Zhang et al. (2015a) was the first to use a bidirectional LSTM for RE to model the sentence with complete, sequential information about all words. The basic idea behind bidirectional LSTM is to

present each training sequence to two separate recurrent nets, both of which are connected to the same output layer. This means that the network has complete, sequential information about all positions before and after each token in a given sequence. Training such a model on word embeddings alone could achieve state-of-the-art results, and importing more features from WordNet and dependency trees could improve the results even further.

Because the SDP in a relation candidate condenses the most interesting and insightful information for entities' relationships, it has been used to represent a relation in (Xu et al., 2015b) rather than the full sentence. A multi-channels LSTM was trained on top of multiple information sources, including POS tags, WordNet hypernyms, word representation, and grammatical relations, allowing effective information integration from heterogeneous sources across dependency paths.

Rather than relying on external lexical resources or NLP systems, Zhou et al. (2016) used an Attention-Based Bidirectional Long Short-Term Memory Network (Att-BiLSTM) to capture the most important semantic information in a sentence. Word level attention could automatically focus on the words that have decisive effect on classification, then merge them into a sentence-level feature vector.

Nonetheless, this attention mechanism does not fully exploit information about the target entities, which may be the most important feature for relation classification. Therefore, entity-aware attention mechanism with a latent entity types was proposed in (Lee et al., 2019). This type of attention focuses on the most important semantic information by considering entity pairs with along their latent type representation. In addition, to capture the meaning of the sentence's words with consideration of their context, self-attention mechanism was used on the sequence of words before giving it to the BiLSTM model to extract (cf. Figure 1.14).

Finally, position-aware attention mechanism was explored by Zhang et al. (2017b). They argue that their model can not only integrate the semantic information about the input sequence, but also information related to global positions of the entities within the sequence.

1.4.3 Graph Neural Network

Dependency trees increased relation extraction models' ability to capture long-range word relationships. Reducing the tree to the SDP between the target entities is more efficient computationally, but it limits the dependency information that can be included in the relation representation.

Example (7) shows a sentence extracted from the TAC KBP challenge corpus,[9] that expresses a relationship between the two entities, **he** and **Mike Cane**. [*he* ← *relative* →

[9]https://catalog.ldc.upenn.edu/LDC2018T03

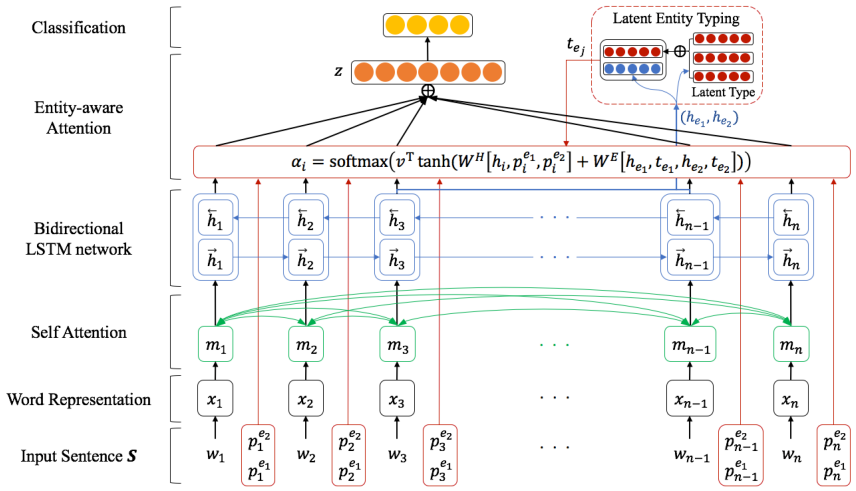

Figure 1.14 : BiLSTM with entity-aware attention using latent entity typing (Lee et al., 2019).

Cane] is the shortest dependency path between these two entities. We observe that the negation word **not** is off the SDP. Thus, when considering this path for relation prediction, the predicted relation type is *per:other_family*, however the gold label is *no_relation*.

(7) I had an e-mail exchange with Benjamin Cane of Popular Mechanics, which showed that [**he**]$_{e_1}$ was not a relative of [**Mike Cane**]$_{e_2}$.

Given the importance of dependency relations in RE, it is possible that a relation instance would be best expressed using the full tree. GNNs were designed to handle such complex structures and learn features on top of them. These architectures generate graph embeddings by learning embeddings for each node in the graph and aggregating the node embeddings. For a more in-depth understanding of these neural networks, see (Wu et al., 2021).

Different pruning strategies were proposed to use the dependency tree efficiently by incorporating relevant information while removing as much irrelevant content as possible. For example, Zhang et al. (2018c) applied GNN over a pruned tree, which only keeps words immediately around the SDP between the two target entities. This tree includes tokens that are up to distance K-away from the dependency path in the subtree, including the lowest common ancestor of the target entities.

Guo et al. (2019b) proposed to prune the dependency tree automatically by transforming it into a fully connected edge-weighted graph using an attention mechanism. This soft-pruning approach automatically learns how to selectively attend to the relevant sub-structures useful for the RE task. Recently, Tian et al. (2021) proposed attentive graph convolutional networks (A-GCN). They applied an attention mechanism upon graph convolutional networks to

different contextual words in the dependency tree to distinguish the importance of different word dependencies.

Rather than using dependency trees that are generated by off-the-shelf parsers, which are not always of high-quality, Qin et al. (2021) proposed building a graph from n-grams extracted from a lexicon constructed from pointwise mutual information in an unsupervised manner. Then attentive graph convolutional networks are applied over this graph to weight different word pairs from contexts within and across n-grams, improving the relation representation.

1.4.4 Hybrid Networks

Many works tried to utilize advantages from diverse neural architectures and propose hybrid architectures. For example, RNN and CNN models have been combined in many works to join their capacities in learning local and global long-distance features from sequential textual data or dependency tree substructures (Ren et al., 2018; Tran et al., 2019). Some works have combined these models at the evaluation stage through a voting process that consists in applying several CNN and RNN models on each sentence of the test set, and predict the class the most voted by these models (Vu et al., 2016), while others adopted a joint training strategy to extract features at different levels.

Liu et al. (2015), for example, enhanced the SDP between target entities by adding subtrees connected to it. Then, a CNN was used to learn features from the SDP flat structure, while a RNN was used to extract hierarchical features from the subtrees. Later, a BiLSTM and a CNN were merged to learn bidirectional relation representation from the SDP (Cai et al., 2016). BiLSTM could encode global patterns, whereas CNN could catch local ones on the SDP. Furthermore, the LSTM's bidirectionality aided in classifying the direction of relations in SemEval 2010 Task 8 dataset. In contrast, Le et al. (2019) suggested improving the SDP using an attention mechanism applied on SDP attached child nodes, then to use a CNN to learn syntactic features from it and a LSTM to learn lexical features from word sequences.

Recently, Chen et al. (2021) recently proposed a hybrid model that combines a CNN neural model with hand-crafted features. Given the advancement of neural models in designing efficient architectures, as well as the developed experience in manually selecting optimal features to represent a relation instance, the proposed combination could achieve an impressive +4% improvement on the ACE dataset when compared to related works.

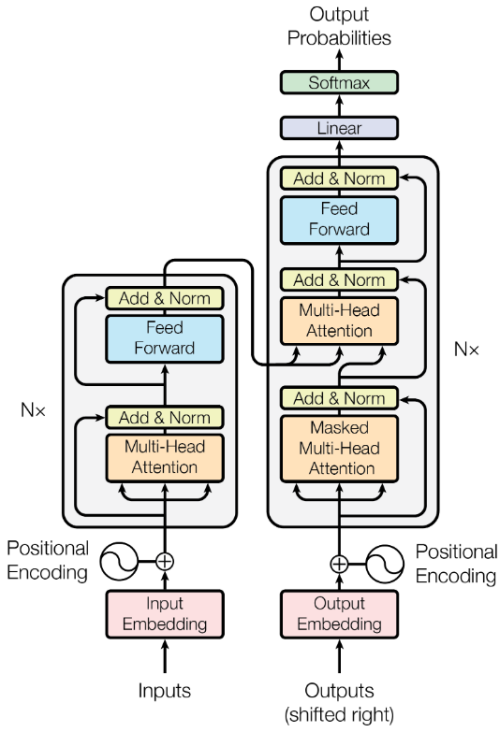

Figure 1.15 Transformer architecture (Vaswani et al., 2017).

1.4.5 Transformers

RNNs are sequential models that process one token at a time. For a current token at a position i, its hidden state is a function of the previous hidden state at the position $i-1$ and the token representation at the position i. This sequential processing prevents parallelization inside training samples, which gets more difficult as sequence length increases (Vaswani et al., 2017). To overcome this issue, the transformer model was developed, which relies on the attention mechanism that allows modeling of dependencies regardless of where they appear in the input or output sequence. This model is based on an encoder-decoder architecture (cf. Figure 1.15) that combines stacked multi-head self-attention and point-wise, fully connected layers for both the encoder and the decoder (Vaswani et al., 2017).

Devlin et al. (2019) pre-trained transformer encoder architecture on two unsupervised tasks, namely Masked Language Model (MLM) and Next Sentence Prediction (NSP) using the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words), to obtain contextual word representations that fuse the left and the right context of a word, giving rise to BERT: Bidirectional Encoder Representations from Transformers.

With just one additional output layer, the pre-trained BERT model was fine-tuned to create state-of-the-art models for numerous NLP tasks (Rajpurkar et al., 2016; Wang et al.,

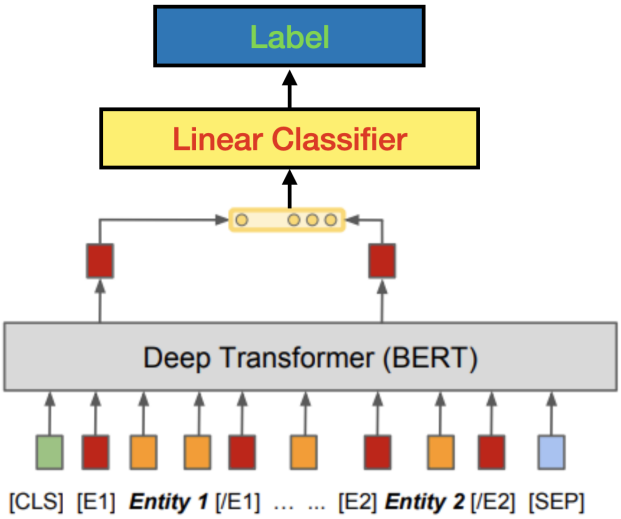

Figure 1.16 BERT architecture for RE (Baldini Soares et al., 2019).

2018), including RE task (Tao et al., 2019; Wang et al., 2019; Wu and He, 2019; Yamada et al., 2020b). In this case, the relation type is predicted by fine-tuning a classification layer, whose size equals to the number of relations, along BERT parameters. The general architecture of BERT for RE is depicted in Figure 1.16.

Other studies followed the same general idea of developing BERT to propose new language models that target new languages (Le et al., 2020) or incorporate new forms of information (Lee et al., 2020; Liu et al., 2021c; Peters et al., 2019; Wang et al., 2019; Zhang et al., 2019). In this dissertation, "transformer", "transformer encoder", and "pre-trained language models" are all expressions that are used interchangeably to refer to a transformer encoder architecture that has been pre-trained on language modeling objectives on a large set of non-annotated data. We group the contributions for transformer-based RE according to the types of models, considering:

(a) The various *input representations* of a relation instance,

(b) The use of *different strategies for learning* RE or *additional training objectives*,

(c) Reformulating RE as a non classification task,

(d) Enhancing the model by *syntactic knowledge* as given by NLP parsers, and

(e) Integrating other kinds of *background knowledge* as given by external resources.

In the following, we detail state-of-the-art models that addressed each of these categories, focusing on (a) to (d). We review *background knowledge* enhanced models in Chapter 5, Section 5.1.

Figure 1.17 Entity markers used to identify target entities in the input sentence.

Investigating Input Representations.

Because RE task requires knowledge of both the context of the sentence and the targeted entities, many works using transformers have sought the most efficient representation for relation candidates while leveraging contextualized word vectors generated by pre-trained language models (PLM).

Alt et al. (2018) was the first to use a transformer encoder for relation extraction after pre-training it on language model objective proposed in (Radford et al., 2018) on BookCorpus dataset (Zhu et al., 2015). They also looked into various entity masking strategies, such as replacing target entities with an UNK token, their types, or their grammatical roles (subject/object). These strategies produced good results and were efficient in preventing overfitting and allowing for better generalization to previously unseen entities. Shi and Lin (2019) also followed this entity masking strategy where the target entities are replaced by the combination of their roles and their types, for example: SUB-LOC is the mask for a *subject* entity of type *location*. The BERT outputs are injected into a BiLSTM layer, which generates relation representations that are fed into the relation classifier.

In (Wu and He, 2019), entity markers—special tokens positioned before and after the target entities—were used to localize the target entities in the input sentence (cf. Figure 1.17). The information about target entities in the RE model was then enhanced by fusing the sentence representation for relation prediction (CLS token) with the entity representations created by BERT (cf. Figure 1.18).

Furthermore, Baldini Soares et al. (2019) experimented with different architectures to extract fixed-length relation representation from BERT for RE. Different inputs, such as entity mentions, adding entity markers, or using token type indicator, were tested with various BERT outputs, such as CLS token, entity mentions pooling, and entity markers pooling. The best results were obtained when entity markers were included in the input sentence and the BERT's output of the entity marker at the start of each target entity was used as a relation representation.

Besides, Zhou and Chen (2021) looked into the impact of different types of entity information that can be injected at the input level. They experimented with combination of various entity indicators such as entity mention, entity types, entity roles, and punctuation

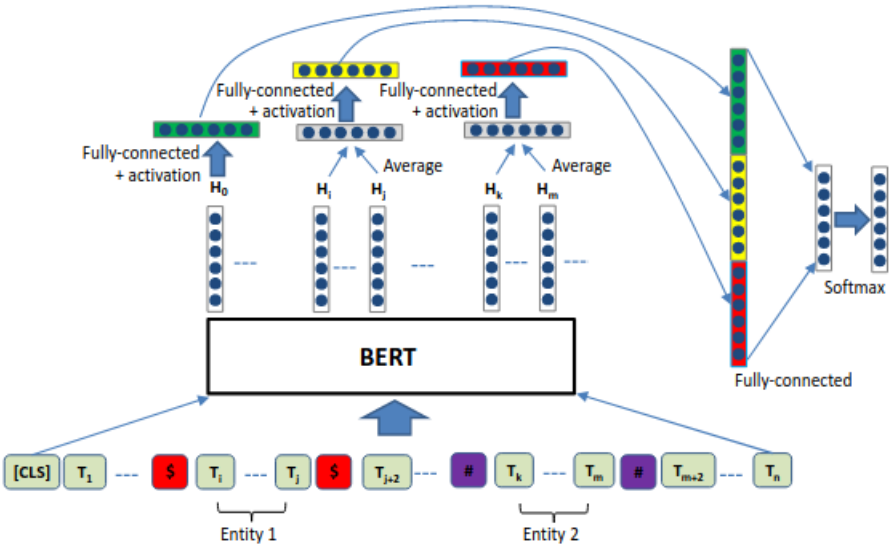

Figure 1.18 Enriching BERT with entity information for RE (Wu and He, 2019).

or special tokens as entity indicators. Their findings confirm that both the entity mention and type, as well as a punctuation entity indicator, help to improve relation representation, achieving an F1 of 74.6% on TACRED dataset. In (Peng et al., 2020a), the contribution of context information and entity information in the RE task was evaluated. They discovered that while context is the most important source of information for relation prediction, entity information, particularly entity type, is also critical for RE.

Recently, to automate the search for the optimal architecture for a RE task, Zhu (2021) used reinforcement leaning (RL) strategy following the efficient neural architecture search proposed in (Pham et al., 2018). The evaluation of their system on eight benchmark datasets for RE could result in the most optimal design in terms of entity and context representation layers, outperforming a simple BERT baseline.

The attention mechanism was also employed to build more accurate relation representations from pre-trained transformer encoders. Liang et al. (2022) proposed a novel approach for extracting multi-granularity features from the original input texts. For varied granularity feature extraction, three levels of attention were used: 1) mention attention, which is intended to extract entity mention features from given entity pairs; 2) mention-aware segment attention, which is based on entity mention features extracted from previous mention attention and aims to extract core segment level feature related to entity mentions; and 3) global semantic attention, which focuses on sentence level feature. To construct the relation representation, the acquired features are combined. BERT and SpanBERT incorporating this approach could produce better results than without it.

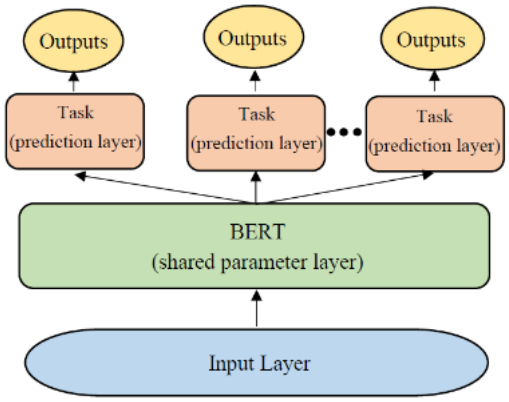

Figure 1.19 The structure of multitask RE model (Wang and Hu, 2020).

Learning Strategies.

Various strategies for improving RE learning have been proposed in the literature to improve the classifier's ability to predict relations. Some work to enhance the learning of difficult relations, while others work to improve the extraction of under-represented relations.

Lyu and Chen (2021) suggested learning relations with entity type restriction. For example, given the relation type *ORG:parent*, only sentences with *ORG* target entities can be candidates for this relation. Based on this evidence, the authors proposed training one classifier on a smaller and more precise set of candidate relations for each pair of entity types. This learning paradigm could improve SpanBERT performances on TACRED dataset by 4.4%, reaching 75.2% F1. Kim (2021), on the other hand, proposed learning relation extraction gradually through a curriculum learning process. Where the model quickly learns easy data by finding a parameter space that is more appropriate for the task, then solves the problem of local minima in that space while learning difficult data. This is accomplished by first categorizing the entire corpus into several groups based on the difficulty of the data, and then feeding this data to the model, allowing it to gradually learn the entire corpus based on the difficulty. On the TACRED dataset, this learning process applied to the RoBERTa large model yielded a 75.0% F1.

Modifying the learning objective of the RE model has also been investigated to improve RE performance. For example, joint learning, also known as auxiliary or multitask learning, was explored. It involves optimizing a model to perform more than one task at a time (cf. Figure 1.19). Wang and Hu (2020) selected tasks from the GLUE benchmark to be optimized along RE to give the model the power to extract implicit information that is difficult to learn from RE. Other works, on the other hand, rely heavily on the RE dataset to generate additional learning objectives (cf. Section 4.1.2, Chapter 4).

Peng et al. (2020b) used contrastive learning (Hadsell et al., 2006) as an additional pre-training objective for BERT, along with masked-language modeling, before fine-tuning it for the RE task, with the goal of learning representations by pulling "neighbors" together and pushing "non-neighbors" apart. For pre-training, sentences from Wikidata were used, with a training objective of producing closer representation vectors for examples expressing the same relation and distant representation vectors for examples expressing a different relation. Furthermore, entities in the input sentences were masked to avoid entity memorization and instead focus on context.

Finally, Baek and Choi (2022) proposed a model that addresses high misclassification errors for minority classes in RE by employing an attention module that detects noisy instances of minority relations and then employs the none-noisy ones to perform data-augmentation. Their model achieves the best score for the TACRED dataset to date, with an F1 of 75.4%.

Task Reformulation.

For a long time, the RE problem was cast as a multi-class classification task in which a classifier is trained to generate relation probabilities based on encoded features of the input sentence. This is a straightforward and natural formulation of the RE problem given the shape of the input data. Furthermore, when using this formulation, good results have been obtained. Recently, some attempts have been made to use an intermediate task to solve the RE problem.

Researchers believe that because transformer encoders have been trained on masked language objectives, it would be more appropriate to reformulate the fine-tuning tasks to be optimized using the same pre-training objective. This is referred to as prompting, and it has been shown to be effective in a variety of NLP tasks (see (Liu et al., 2021b) for a more detailed survey). For example, Sainz et al. (2021) proposed reformulating relation extraction as an entailment task, where the task is to infer whether the premise entails the hypothesis based on the input sentence containing the two target entities as the premise and the verbalized description of a relation as the hypothesis using the natural language inference model. On the TACRED dataset, their model achieved 63% F1 zero-shot, 69% with 16 examples per relation (17% points better than the best supervised system under the same conditions), and was only 4 points short of the state-of-the-art (which uses 20 times more training data).

The RE task has also been viewed as a question-answering problem, with the task to perform being a span prediction, in which the relation type is used as a query over a context consisting of one input sentence and one target entity. The second target entity is the query's

answer, which is represented as a span across the context. Cohen et al. (2020) used this method while considering two-way span prediction, with one entity serving as context and the other as the answer, and vice versa.

Other works made use of the entire transformer architecture (encoder and decoder) by reformulating the RE as a sequence to sequence task. Following on from this concept, Cabot and Navigli (2021) proposed REBEL, an autoregressive approach to extracting relationships from various domain datasets after linearizing triplets into text sequences using token markers. When tested on six different RE datasets, their approach turned out to be highly adaptable to new domains or longer documents with no adaptation. Besides, given that summarization tasks aim to extract concise expressions of information from larger contexts, it naturally aligns with the goal of RE, i.e., extracting a type of concise information that describes the relationship of entity mentions. Lu et al. (2022) proposed SuRe that performs RE through summarization using the generative pre-trained models BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020a).

Syntactically Enhanced Models.

Tao et al. (2019) was the first to use syntactic information from external parsers to reduce the input sentence to just what was needed to determine the semantic relationship between target entities. Modifiers like adjectives and adverbs, as well as unrelated named entities, conjunctions, and compound words, are removed from the input sentence, and the generated syntactic indicators are concatenated with the original sentence before being fed into BERT to generate the relation representation. On the SemEval 2010 Task 8 dataset, their model had an F1 of 90.36%.

Huang et al. (2021b) proposed DBERT, a model that integrates information about the dependency between each word and the target entities into the PLM through an attention mechanism. Dependencies are obtained using Stanford Core NLP. Then their embeddings are randomly initialized and used to apply attention of BERT's output vectors of each word (cf. Figure 1.20). The attention operation can automatically learn the contribution of each word in a sentence, given its dependency link to the target entities. Dynamically integrating dependency information into transformer encoder in a flexible manner was investigated in (Tian et al., 2022). Where the authors proposed using dependency masking to create a new dependency-guided pre-training objective for language models. This strategy ensures a more flexible learning process on those frequent and important dependency relations, contrasted to using dependency parser (with noises) as fixed knowledge in the language model.

Recently, Guo et al. (2022) proposed modifying the transformer encoder's self-attention mechanism to account for dependency types of words and identify their importance based

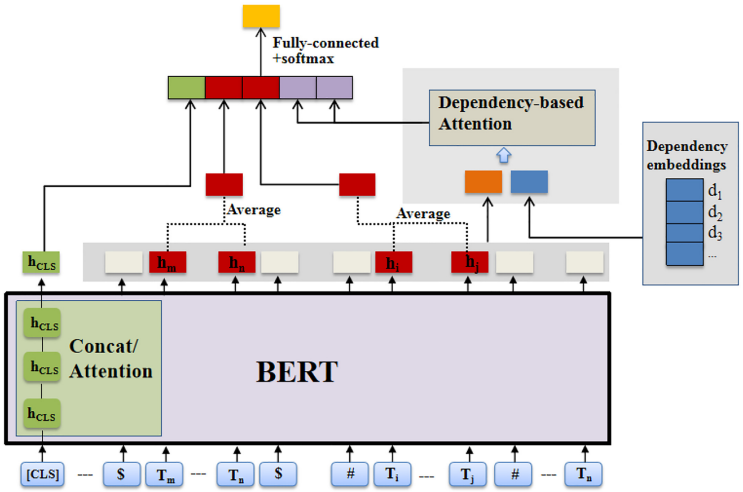

Figure 1.20 Incorporating dependency-based attention into BERT (Huang et al., 2021b).

on them. Their method ensures that context and dependency information are integrated into the encoder in a single stream. Word dependencies are extracted using a dependency parser, then expressed as an adjacency matrix. Each matrix element is mapped to an embedding vector. To incorporate dependency links and type into contextualized word representations, the embedding matrix is used in the self-attention mechanism.

1.5 Evaluation Metrics

The metrics used to evaluate RE models depend on the type of approach used to extract these relations. In a supervised setting, as RE is generally framed as a classification task *Precision (P)*, *Recall (R)*, and *F1-score (F1)* are used to evaluate the performances of the models. These metrics are defined in the following.

- Precision (P): is a measure of how many of the positive predictions made are correct. It is the number of correctly extracted relations divided by the total number of extracted relations.

$$P = \frac{TP}{TP + FP} \tag{1.2}$$

- Recall (R): is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data. It is the number of correctly extracted relations divided by the actual number of extracted relations.

$$R = \frac{TP}{TP + FN} \tag{1.3}$$

- F1-score ($F1$): is a measure that combines precision and recall. It is commonly referred to as the harmonic mean of the two.

$$F1 = 2 * \frac{P * R}{P + R} \tag{1.4}$$

  Two variants of $F1$ are used :

  - The macro-averaged F1 score (or macro F1 score) is computed using the arithmetic mean of all the per-class F1 scores. In an imbalanced dataset, this measure assigns equal importance to all classes. It served as an official metric for assessing RE from the Semeval 2010 Task 8 dataset (Hendrickx et al., 2010).
  - Micro averaging computes a global average F1 score by counting the sums of the True Positives (TP), False Negatives (FN), and False Positives (FP), without considering the relation type. This score gives an idea of overall performance regardless of class and does not account for class imbalance. This metric was used in the TACRED dataset to evaluate the performance of RE models (Zhang et al., 2017b).

When relation identification is performed along with relation classification, and the negative relation is included in RE dataset, some datasets tend to ignore this relation at the evaluation stage and do not include it in the calculation of the evaluation measure. This approach has been used in well-known RE shard tasks, such as in the Semeval 2010 Task 8 (Hendrickx et al., 2010) and TACRED (Zhang et al., 2017b).

Also, when relation extraction consists in extracting one or more tokens in texts expressing this relation, the Jaccard Similarity Coefficient is generally used (Reyes et al., 2021). This coefficient is defined as the size of the intersection divided by the size of the union of the sample sets, assesses the similarity between two sets (cf. Equation 1.5).

$$JACCARD_{sim}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1.5}$$

This can provide a more accurate perspective of the model's efficiency when recognizing the tokens contained in the relationship, and it can penalize it less in the case of relationships represented with many tokens.

1.6 Conclusion

RE is an important IE task and the literature is vast. Several excellent surveys of the field exist. The first one was proposed by (Bach and Badaskar, 2007), followed by (Sharma

et al., 2016) who focused on binary and complex relation extraction techniques in the biomedical domain. Recently, (Pawar et al., 2017) surveyed advances in supervised and semi-supervised methods while (Smirnova and Cudré-Mauroux, 2018), (Niklaus et al., 2018) and (Kumar, 2017) focused on distant supervision, Open IE and deep learning methods for RE, respectively. Finally, a good survey of RE (and IE in general) in a semantic web settings has been proposed by (Martinez-Rodriguez et al., 2020), followed by Wang et al. (2022)'survey on deep learning-based models for RE.

Based on these surveys and relevant literature in the field, this chapter attempts to review main existing techniques on first-order binary generic relation extraction at the sentence level, discussing in particular main related concepts, typology of relations, and the RE overall pipeline. We also presented main existing annotated datasets, how they were built as well as a quantitative description of the annotated relation instances.

The models trained on these datasets were also presented and classified into five groups based on how the training data was generated. Because manually annotated data is of higher quality, we devoted a significant part of this chapter to describing supervised methods that are trained on it. Supervised methods include traditional models that rely on manually generated or off-the-shelf NLP tools, as well as neural models that learn them automatically. When using dense semantic representations of words, known as word embeddings, neural-based models outperform traditional models when trained on benchmark datasets such as TACRED and SemEval 2010 Task 8. Pre-trained transformers, in particular, achieve cutting-edge results nowadays due to their ability to represent words based on their context.

The datasets and approaches for generic RE has been used as a basis for domain-specific RE that are crucial in many real-world applications. We review in the next chapter related work in extracting relations in the scientific and biomedical domains, but also in business and financial contents, the focus of this thesis.

Chapter 2

Domain Specific Relation Extraction: A Focus on Business Relations

Many specific domains have benefited from the advancement of information extraction tools to automate the extraction of structured information from texts, including Biomedical (Krallinger et al., 2017; Xing et al., 2020), Sports (Surdeanu et al., 2011), Political (Sundheim, 1992), Scientific (Augenstein et al., 2017; Lee et al., 2017), Food and Health (Nanba et al., 2014; Wiegand et al., 2012; Zhang et al., 2022), and Business domains (Jabbari et al., 2020; Sharma et al., 2022).

As opposed to generic relations, the extraction of domain-specific relations requires significantly more domain expertise provided by specialists or knowledge bases. This chapter provides a more comprehensive overview of the datasets, approaches, and applications proposed for domain-specific relation extraction, with a focus on biomedical (Section 2.2) and scientific domains (Section 2.1) because they have received the most attention in the literature, and business domain (Section 2.3) as it is the primary focus of this dissertation.

2.1 RE in the Scientific Domain

Keeping track of scientific challenges, advancements, and future directions is an important aspect of research. However, researchers are confronted with a flood of scientific publications, making their manual processing tedious. Therefore, automating the extraction of information expressed in scientific publications allows researchers to gain key insights from these documents.

Relations in scientific papers can be expressed between identified concepts which can refer to metrics, tasks, models/approaches, or datasets (Hou et al., 2019). Figure 2.1 depicts

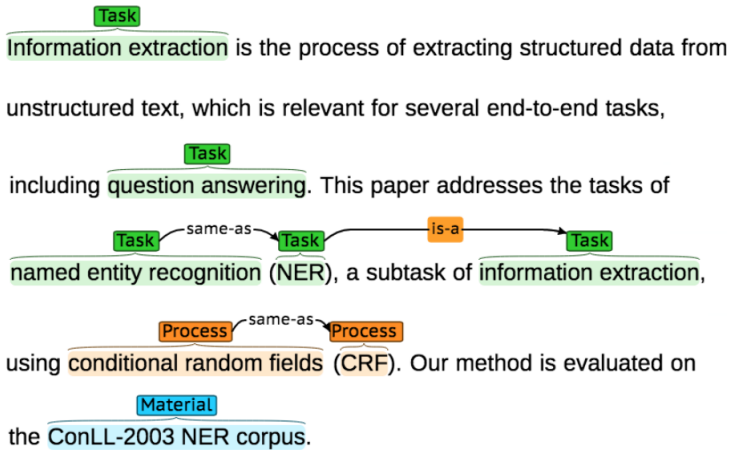

Figure 2.1 An example of annotated concepts and relations in an instance from SemEval 2017 Task 10 dataset (Lee et al., 2017).

an example from SemEval 2017 Task 10 dataset (Lee et al., 2017) of a *hypernym* relation (i.e., `is-a`) between two *tasks* and a *synonym* relation (i.e., `same-as`) between two *processes*, expressed in a scientific paper.

Generic relations have been targeted between domain-specific entities from scientific papers, such as *synonyms*, *hypernyms*, and *part-of* (Augenstein et al., 2017; Bruches et al., 2020; Luan et al., 2018a). Other works focus on more specific relations like *trade-off* that expresses a problem space in terms of mutual exclusivity constraints between competing demands (Kruiper et al., 2020), *usage* to refer to either *used by*, *used for task*, or *used on data* (Bruches et al., 2020; Buscaldi et al., 2018), or *compare* to compare between entities (e.g., datasets, models) in a paper (Bruches et al., 2020; Buscaldi et al., 2018; Luan et al., 2018a). Ma et al. (2022) extracted a metrics-driven mechanism triples (Operation, Effect, Direction), where they considered three relation types (positive effect, negative effect, and other) describing the influence direction between the operation entity and the effect entity.

2.1.1 Datasets

SemEval 2017 Task 10 was the first published benchmark dataset for relation extraction from scientific documents. It focuses on the extraction of hyponym and synonymy relations between three types of key-phrases rather than named entities (Augenstein et al., 2017) (as shown in Figure 2.1). Buscaldi et al. (2018) proposed SemEval 2018 Task 7, a second shared task, that focuses on 6 major relations between 7 entity types, where each target relation in their dataset is divided into more fine-grained sub-relations. Figure 2.2 describes the semantic relation typology used to annotate this dataset.

RELATION TYPE	Explanation	Example
USAGE	*Methods, tasks, and data are linked by usage relations.*	
used_by	ARG1: *method, system* ARG2: *other method*	approach – model
used_for_task	ARG1: *method/system* ARG2: *task*	approach – parsing
used_on_data	ARG1: *method* applied to ARG2: *data*	MT system – Japanese
task_on_data	ARG1: *task* performed on ARG2: *data*	parse – sentence
RESULT	*An entity affects or yields a result.*	
affects	ARG1: *specific property of data* ARG2: *results*	order – performance
problem	ARG1: *phenomenon* is a problem in a ARG2: *field/task*	ambiguity – sentence
yields	ARG1: *experiment/method* ARG2: *result*	parser – performance
MODEL	*An entity is a analytic characteristic or abstract model of another entity.*	
char	ARG1: *observed characteristics* of an observed ARG2: *entity*	order – constituents
model	ARG1: *abstract representation* of an ARG2: *observed entity*	interpretation – utterance
tag	ARG1: *tag/meta-information* associated to an ARG2: *entity*	categories – words
PART_WHOLE	*Entities are in a part-whole relationship.*	
composed_of	ARG2: *database/resource* ARG1: *data*	ontology – concepts
datasource	ARG1: *information* extracted from ARG2: *kind of data*	knowledge – domain
phenomenon	ARG1: *entity, a phenomenon* found in ARG2: *context*	expressions – text
TOPIC	*This category relates a scientific work with its topic.*	
propose	ARG1: *paper/author* presents ARG2: *an idea*	paper – method
study	ARG1: *analysis* of a ARG2: *phenomenon*	research – speech
COMPARISON	*An entity is compared to another entity.*	
compare	ARG1: *result, experiment* compared to ARG2: *result, experiment*	result – standard

Figure 2.2 Semantic relation typology of SemEval 2018 Task 7 dataset by (Buscaldi et al., 2018).

To jointly extract scientific entities, relations, and co-reference links, SCIERC dataset was proposed (Luan et al., 2018a). It focuses on 7 relation types, including co-reference, between 6 types of entities. Recently, Jain et al. (2020) introduced SciREX, a document level IE dataset that covers multiple IE tasks, including entity identification (Dataset, Method, Metric, and Task) and document level N-ary relation identification from scientific articles.

Instead of only extracting relations between coarse-grained entities from scientific papers such as *method*, *dataset*, etc. that describes how research is carried out at a higher-level, Magnusson and Friedman (2021) target scientific claims which focus on the subtleties of how experimental associations are presented. Scientific claims are represented using entities, attributes that applies to these entities, and relations linking them. These authors proposed SciClaim, a dataset of 901 sentences having a total 12,738 labels about entities, attributes, and relations. The dataset is composed of manually annotated claims from different sources: claims identified by experts from Social and Behavior Science (Alipourfard et al., 2021), causal language phrases identified in PubMed papers (Yu et al., 2019), and claims and causal language heuristically identified from CORD-19 abstracts (Wang et al., 2020a). Figure 2.3 presents an example of a knowledge graph that has been extracted from a scientific claim.

Table 2.1 summarizes the key features of the proposed datasets for RE from scientific journals. Mainly, the extraction of relations is performed at the mention level, where named-entity extraction is considered in the majority of datasets, along with relation extraction. Manually annotated datasets like SemEval 2017 and SciERC are small in comparison to

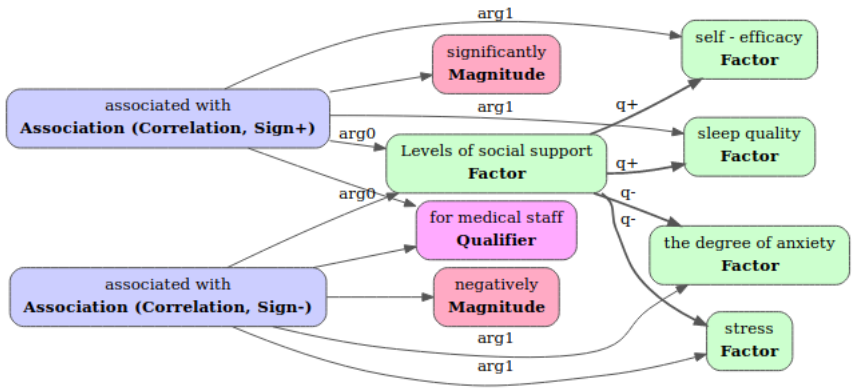

Figure 2.3 SciClaim knowledge graph with entities (nodes), relations (edges), and attributes.

Table 2.1 An Overview of scientific RE datasets. NER: Named-entity extraction, RE:Relation extraction. DS: distantly supervised.

DATASET	#Words	#Entities	#Inst.	#Rel.	Annot.	Rel. level	Tasks
SemEval 2017	84,010	9,946	672	2	gold	mention	RE
SemEval2018	58,144	7,483	1,595	6	gold	mention	RE
SciERC	65,334	8,089	4,716	7	gold	mention	NER+RE
SciClaim	20,070	5,548	5,346	7	gold	mention	NER+RE
SciRex	2,512,806	157,680	9,198	16 (binary) 5 (4-ary)	DS	doc	NER+RE

the distantly supervised dataset SciRex (672 and 4,716 vs. 9,198 instances). Furthermore, except for the dataset SciRex, which focuses on 4-ary relations, the majority of the datasets target binary relations.

2.1.2 Main Approaches

The majority of proposed systems for relation extraction from scientific papers use supervised approaches, with only a few using semi-supervised (Luan et al., 2017) or unsupervised methods (Groth et al., 2018; Kruiper et al., 2020; Lauscher et al., 2019) leveraging Open IE models. This section provides an overview of main approaches focusing on supervised models.

Hettinger et al. (2018) explored feature-based models thanks to an SVM classifier on handcrafted context and entity features combined with domain-specific word embeddings and entity embeddings. Neural based models were widely explored either by using ensemble

learning (Jin et al., 2018; Rotsztejn et al., 2018), multitask leaning (Luan et al., 2018a), multiple or specific losses' optimization (Jin et al., 2018; Pratap et al., 2018), or by training a vanilla neural model (Luan et al., 2018b; Nooralahzadeh et al., 2018). Both CNN and RNN architectures have been explored. Generally, these models are trained on pre-trained domain-specific word embeddings trained on ACL anthology (Jin et al., 2018; Nooralahzadeh et al., 2018; Rotsztejn et al., 2018) or on arXiv data (Hettinger et al., 2018). Few works have evaluated generic word embedding, such as Word2vec or Glove (Nooralahzadeh et al., 2018; Pratap et al., 2018). Other additional features were added to word embeddings such as WordNet features, entity embeddings, POS embeddings or position embeddings (Hettinger et al., 2018; Jin et al., 2018; Pratap et al., 2018).

Rather than using a sequence of words as input to the neural model, Nooralahzadeh et al. (2018) used the shortest dependency path between entity pairs as an input given the importance of information it contains. A character level attention mechanism was also used in (Luan et al., 2018b) to select the most accurate pre-trained concept candidates embeddings to include in the relation representation before feeding the classification layer with it.

More recently, transformer-based models have been used. Jain et al. (2020) combined BERT with a BiLSTM model to get token representations and perform an end-to-end entity relation extraction. SciBERT, a BERT architecture trained on scientific articles was used in (Magnusson and Friedman, 2021) to generate span representations that are used to classify entities first and only infer relations on pairs of identified entities.

Most works on relation extraction from scientific papers focus solely on the content of a single paper, without considering how the document is linked to the whole collection of scientific papers. Viswanathan et al. (2021) recently proposed a citation-aware scientific IE architecture that uses both structural information from the citation graph of referential links between citing and cited papers, as well as textual information from the content of citing and cited documents. These details are incorporated into the relation extraction task, thus greatly improving the outcome of that task.

2.1.3 Applications

The extracted relations from scientific publications can be used to automatically build knowledge graphs from a large collection of documents and analyze information in the scientific literature (Luan et al., 2018a).

We review here some interesting applications that have been proposed for RE in the scientific domain.

Rank	Model	F1	F1 (10% Few-Shot)	F1 (5% Few-Shot)	F1 (1% Few-Shot)	F1 (Zero-Shot)	Extra Training Data	Paper	Code	Result	Year	Tags
1	**DeepStruct multi-task w/ finetune**	76.8					×	DeepStruct: Pretraining of Language Models for Structure Prediction			2022	
2	**RE-MC**	75.4					✓	Enhancing Targeted Minority Class Prediction in Sentence-Level Relation Extraction			2022	
3	**RECENT+SpanBERT**	75.2					×	Relation Classification with Entity Type Restriction			2021	
4	**SuRE** (PEGASUS-large)	75.1	70.7	64.9	52	20.6	×	Summarization as Indirect Supervision for Relation Extraction			2022	
5	**EXOBRAIN**	75.0					×	Improving Sentence-Level Relation Extraction through Curriculum Learning			2021	

Figure 2.4 TACRED dataset leaderboard from PaperWithCode website.

To identify papers dealing with the same task and to track the evolution of the task's results in IA for example, Leaderboards were constructed manually such as PaperWithCode[1] or NLPIndex.[2] Information extraction from scientific documents helps the automatic generation of such boards, while ensuring a frequent and a rapid update (Kabongo et al., 2021). For example, Figure 2.4 represents TACRED dataset leaderboard on the PaperWithCode website where the name of the model, its score and the link to its code are extracted from papers about RE evaluated on this dataset.

Finally, Lahav et al. (2022) focused on the COVID-19 pandemic by analyzing a large corpus of interdisciplinary work ranging from biomedicine to AI and economy. Their system identifies challenges and directions from a collection of publications, and creates a dedicated search engine to assist scientists and medical professionals in analyzing scientific literature (cf. Figure 2.5).[3]

2.2 RE in the Biomedical Domain

2.2.1 Datasets

Biomedical RE (BioRE) refers to the identification and classification of relation mentions between different biomedical concepts within a document.

Various datasets have been built for the biomedical relation extraction task involving different relation types such as drug-drug interactions (Segura-Bedmar et al., 2013), genes-

[1]https://paperswithcode.com
[2]https://index.quantumstat.com/
[3]https://challenges.apps.allenai.org/

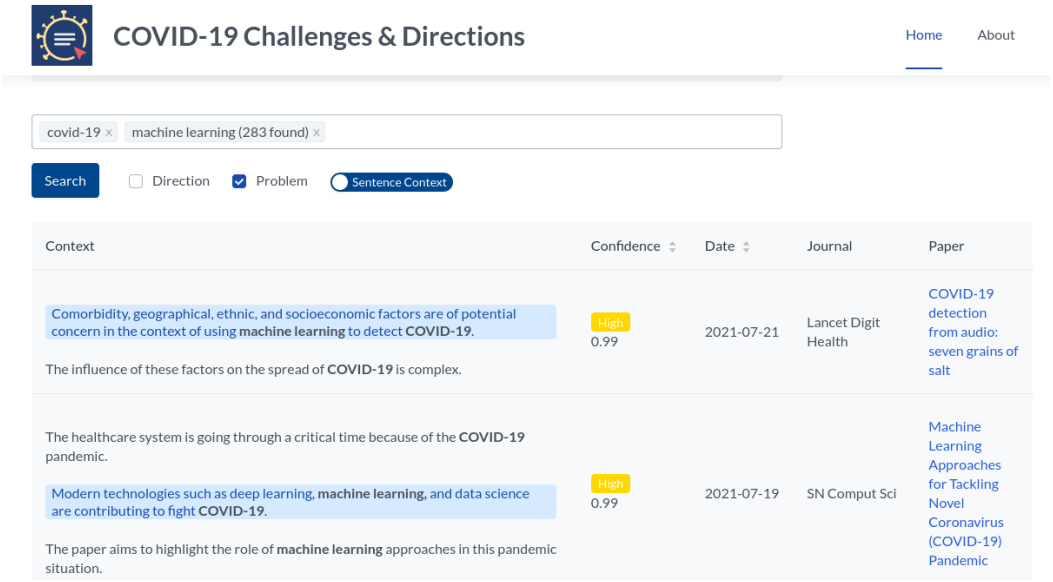

Figure 2.5 Search-engine for scientific challenges and directions about COVID-19.

disease association (Lee et al., 2013; Van Mulligen et al., 2012; Wu et al., 2019), protein-protein interaction (Bunescu et al., 2005), and chemical-protein interaction (Krallinger et al., 2017). Sentence in (1) expresses an *Advice relation* between the two drugs e_1 and e_2 in the DDI 2013 corpus (Segura-Bedmar et al., 2013).

(1) Coagulation test should be monitored when **[warfarin]**$_{e_1}$ or its derivatives and **[enoxacin]**$_{e_2}$ are given concomitantly. **(Relation:** `Advice`**)**

Annotating such domain-specific relations requires experts knowledge. However, few datasets are manually annotated by experts; hence, they contain small number of instances such as AiMed dataset for protein-protein interaction (Bunescu et al., 2005), EU-ADR for gene–disease association (Van Mulligen et al., 2012), CoMAGC dataset comprising gene-cancer associations on prostate, breast, and ovarian cancers (Lee et al., 2013), and CDR for chemical-disease interactions (Li et al., 2016).

Most of biomedical datasets for relation extraction are annotated using distant supervision relying on knowledge bases such as the Comparative Toxicogenomics Database (Davis et al., 2021), DisGeNet database (Piñero et al., 2016), and Unified Medical Language System metathesaurus (Bodenreider, 2004). Among them, we cite the GDA, a dataset for gene-disease associations (Wu et al., 2019), BioRel a large-scale dataset between a variety of entity types, including clinical drugs, pharmacologic substance, organic chemical, disease or syndrome, biologically active substance, molecular function, food, organ or tissue function,

Table 2.2 An Overview of biomedical RE datasets. DS: distantly supervised.

DATASET	Rel. level	#Inst.	#Rel.	#Entities	Lang.	Annot.
i2b2	sentence	6,310	11	8,296	EN	gold
AiMed	sentence	1,101	2	4,141	EN	gold
ChemProt	sentence	10,031	6	15,739	EN	gold
ADE	sentence	7,100	2	5,063 drugs 5,776 adv. eff. 231 dosages	EN	gold
DrugProt	sentence	17,288	13	89,529	EN	gold
BioRel	sentence	534,406	125	69,513	EN	DS
TBGA	sentence/doc.	218,973	4	3,635 genes 1,904 diseases	EN	DS
SciERC	document	4,716	7	8,089	EN	gold

and neoplastic process (Xing et al., 2020). We finally cite TBGA, a large-scale gene-disease association dataset (Marchesin and Silvello, 2022).

All the previous datasets target relations that occur at the sentence level. To overcome this limitation and account for entities at the document level, the BioRED dataset was proposed, which contains a set of 600 PubMed abstracts (Luo et al., 2022). More recently, Tiktinsky et al. (2022) proposed an expert annotated dataset between drugs from scientific literature, covering, for the first time, variable length *n*-ary relations, where the *n* referring to the number of entities involved in the relation is variable.

Table 2.2 gives a non-exhaustive list of BioRE datasets. We can see that the majority of them are in English,and manually annotated by experts. Furthermore, in comparison to those in the scientific domain, these datasets are larger (up to 534,406 instances for BioRel dataset). The number of targeted relation types is relatively large, ranging between 2 and 125.

2.2.2 Main Approaches

The first proposed models relied on co-occurrence or rule-based approaches (Blaschke et al., 1999; Herrero-Zazo et al., 2013). For example, Wong (2000) used a set of rules defined in the form of regular expressions over words or POS to extract protein-protein interactions from abstracts in MEDLINE journal. Šarić et al. (2006) extended the set of rules to extract a gene-protein networks while capturing the linguistic structures of these relations. Fundel et al. (2007) used dependency parse trees to extract informative paths between entity mentions that are protein names. These paths should contain just the relevant terms describing the relation between the given pair of proteins. Then, a set of rules is used to identify the type of relation between these entities. For example, the rule *effector-relation-effectee* is used to select *activation* relations from the extracted dependency paths between proteins A and B.

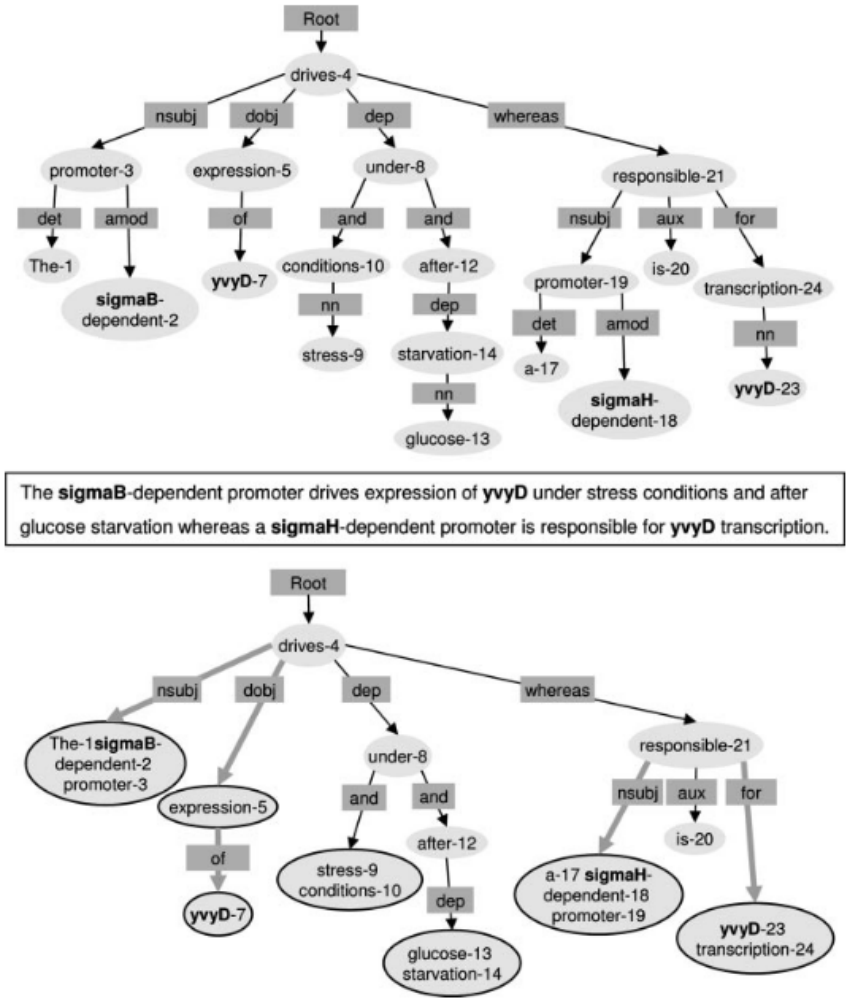

Figure 2.6 Rule based BioRE using dependency trees as proposed by (Fundel et al., 2007).

Figure 2.6 illustrates an example of this extraction. The upper panel represents the dependency parse tree as derived from the Stanford Lexicalized Parser, showing words (ellipses) assigned with word positions (numbers appended to words), dependencies (edges pointing from the head of a dependency to the dependent word), dependency types (rectangles) and the head of the sentence (Root). The lower panel describes the corresponding chunk dependency tree that groups the words into noun phrase chunks (framed ellipses). Words marked in bold indicate gene/protein names, thick gray edges indicate paths that are extracted by *effector-relation-effectee* rule.

As designing these patterns requires heavy human effort, Xu et al. (2013) automated pattern were extracted following a bootstrapping approach that uses known entity pairs from a knowledge base to extract patterns from scientific articles. These patterns later enable the

extraction of new entity pairs. Based on the idea that the entities which frequently appear together are more likely to be related in some way, a co-occurrence statistic method was used in (Chen et al., 2008) to calculate and evaluate the degree of association between disease and relevant drugs from clinical narratives and biomedical literature.

Machine learning methods have emerged to automate the extraction process as the amount of annotated data has grown through biomedical text mining competitions such as BioCreative (Arighi et al., 2011; Krallinger et al., 2008) and BioNLP (Deléger et al., 2016; Nédellec et al., 2013). First methods relied on manually selecting the most productive relation representations, either by using a set of handcrafted features extracted from different parsers (e.g., POS, syntactic, dependency) (Bundschus et al., 2008; Chowdhary et al., 2009; Craven et al., 1999; Kim et al., 2015), or by using rich structural representations such as syntactic parse trees or dependency parse trees (Airola et al., 2008; Özgür et al., 2008; Warikoo et al., 2018), to train SVM or Naive-Bayes linear classifiers. We can cite as an example Craven et al. (1999) who represented drug-protein interactions by using bag of words (BOW), relational dependencies, and external KB-based entities as features for a Naive Bayes classifier. Chun et al. (2006) recognized relations between prostate cancer terms and relevant gene terms from MEDLINE abstracts by training a maximum entropy-based model. It then classifies the identified interactions into six important topics as initially defined by human genetics experts and oncologists to facilitate the use of the extraction system. The following sentence (cf. (2)) illustrates such a topic relation where the topic *pharmacology* is about drugs used to heal cancer, their compositions, uses, and effects.

(2) Objective: To assess the involvement of calcitonin gene-related peptide ([**CGRP**]$_{e_1}$) in the occurrence of hot flashes in men after castration for treatment of **[prostate cancer]**$_{e_2}$, we investigated the effects of CGRP on skin temperature in surgically and medically castrated male rats. **(Relation:** `pharmacology topic interaction`**)**

To automate the extraction of features, deep learning architectures have been widely used in BioRE, in particular CNN (Björne and Salakoski, 2018; Quan et al., 2016) and RNN (Lim et al., 2018; Zhang et al., 2018b; Zheng et al., 2017). These models are trained on a combination of lexical embedding vectors of words computed on one or multiple sources (Björne and Salakoski, 2018; Quan et al., 2016), and on structural embeddings such as POS, SDP, relative distance, etc. (Björne and Salakoski, 2018; Lim et al., 2018; Zhang et al., 2018a,b; Zheng et al., 2017).

Other works combine the advantages of both architectures by proposing hybrid models (Mitra et al., 2020; Sun et al., 2019; Zhang et al., 2018a). For example, Mitra et al. (2020) conceptually combined different deep neural network models (CNN, RNN, MLP) to generate

different views of the input data and learn several abstract features. Jettakul et al. (2019) proposed a hybrid model between a RNN and a CNN. RNN is used to extract full-sentence features from long and complicated sentences, whereas CNN captures SDP features that are shorter, more valuable, and more concise.

As for generic RE (see Section 1.4 in Chapter 1), attention has also been a quite effective mechanism in capturing relevant features in BioRE (Ahmed et al., 2019; Asada et al., 2017; Jettakul et al., 2019). Jettakul et al. (2019) incorporated several kinds of attention mechanisms into their model: Additive attention, Entity-Oriented attention, and Multi-Head attention. Park et al. (2020) used a full dependency tree as input and proposed a novel attention-based pruning strategy that assigns attention weights to each edge via a self-attention mechanism to represent the strength of relatedness between nodes to efficiently use syntactic information of the input sentence while ignoring irrelevant information.

Transformers were also widely used. There have been several attempts to adapt BERT to biomedical corpora. Beltagy et al. (2019), in particular, employed 1.14M publications randomly selected from Semantic Scholar to fine-tune BERT and build SciBERT. There are 18% computer science publications and 82% biomedical papers in the used corpus. In BioRE, SciBERT produced results equivalent to state-of-the-art models. Lee et al. (2020) proposed BioBERT, a pre-trained language representation model trained on biomedical domain corpora (e.g., PubMed articles). BioBERT could outperform the state-of-the-art models in biomedical relation extraction by 3.49 F1 score.

Domain knowledge about biomedical entities can be present in external knowledge bases. Given the importance of entity pairs in the identification of the relation holding between them, some works integrated this external knowledge into the RE model. Lai et al. (2021) presented KECI, a model that integrates a span-based graph with a knowledge graph holding relevant background information for the entities mentioned in the text via an attention mechanism. Asada et al. (2021) presented a method to effectively use information from an external drug database as well as information from large-scale plain text for drug-drug interaction extraction. The model encoded both drug descriptions and information about their molecular structures using SciBERT. This encoded knowledge is used to enrich the input sentence representation and classify the target drug pair into a specific drug-drug interaction type.

To overcome the gap caused by a lack of domain expert annotated data for extremely particular types of biomedical relation, available annotated data for other specialized biological interactions was used. To extract gene-disease relation from a scarce dataset, Nourani and Reshadat (2020) first trained a model on a similar-domain large corpus, then used the lowest layers of this model as a feature extractor. Furthermore, Legrand et al. (2021) has stressed the importance of syntax in RE transfer learning. Their study enhanced the extraction of

interactions from sparse data by relying on knowledge transfer from three bigger datasets, designed to extract various types of biomedical or general domain relations using a TreeLSTM model that considers syntactic aspects.

Finally, some works focus on the extraction of more complex relations between more than two biomedical entities. For example, Zhao et al. (2020) proposed a novel cross-sentence n-ary relation extraction method that utilizes the multi-head attention and knowledge representation learned from a knowledge graph.

2.2.3 Applications

BioRE systems, for example, may be used to find drug-drug interactions in scholarly journals which is an important part of post-market drug safety surveillance, also known as pharmacovigilance (Zhao et al., 2021). Drug-drug interaction detection assists researchers and professionals in identifying and mitigating the unfavorable repercussions of mixing two or more medications administered to a patient. This task also contributes to the upkeep of existing drug databases, ensuring their high coverage and frequent updating (Asada et al., 2017).

The availability of up-to-date illness profiles may be beneficial in understanding disease features (e.g., treatment or symptoms and how they may change over time). Disease-related information may be retrieved and incorporated into the patient record via scientific articles and clinical narratives. They can be used in a variety of applications such as decision support (for example, recommending treatments), quality assurance (for example, inter- and intra-institutional review), clinical information needs (for example, answering clinical questions), information retrieval (for example, classifying documents), and data mining (e.g., hypothesis discovery) (Chen et al., 2008).

During the COVID-19 pandemic, many works have leveraged BioRE models to build useful applications that facilitate knowledge acquisition from the COVID-19 literature. For example, Tran et al. (2021) proposed CovRelex,[4] a scientific paper retrieval system targeting biomedical entities and relations from COVID-19 related scientific papers (cf. Figure 2.7).

Hope et al. (2021), on the other hand, was interested in extracting functional relations between concepts from different disciplines including bio-medicine, to enable the exploitation of interdisciplinary research papers about COVID-19 and build a KB. Figure 2.8 depicts the first result obtained from the search query about the two entities *Vitamin D* and *COVID-19* ran over the constructed KB.

[4]https://www.jaist.ac.jp/is/labs/nguyen-lab/systems/covrelex/home/

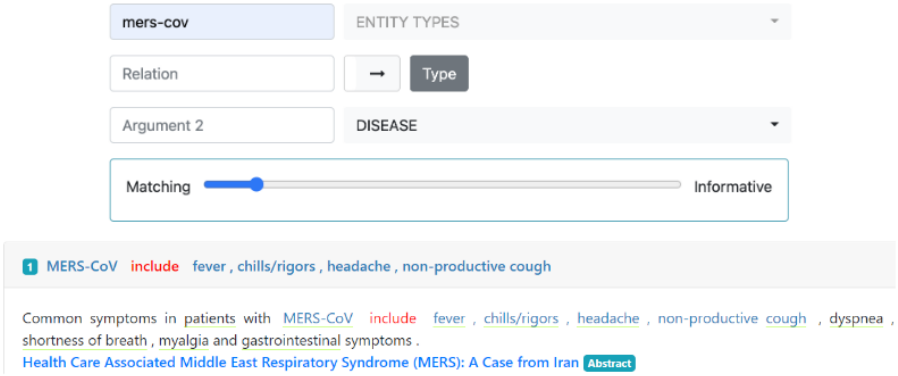

Figure 2.7 An example of a query for *(mers-cov, any-relation, DISEASE)* in CovRelex (Tran et al., 2021).

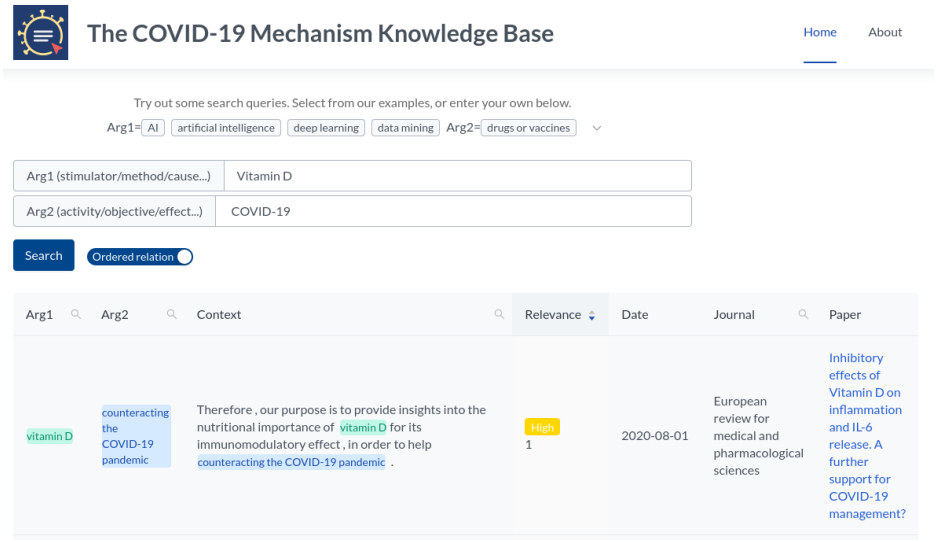

Figure 2.8 The COVID-19 mechanism knowledge base results for the search query (*Vitamin D, COVID-19*.

Finally, BioRE is also used to extract protein-protein interactions, which allows researchers to infer the biological activities of unknown proteins based on the proteins they interact with.It is critical for understanding complicated illness mechanisms and developing effective remedies.

2.3 RE in the Financial and Business domains

To monitor a financial institution's customers, or to track a company's competitors activities, the availability of an automated business relations extraction system is a valuable asset for

professionals to process and structure online as much information as possible. Business relation extraction refers to the automatic detection of semantic relationships between business and financial entities in plain text including named entities of type: *organization, person, product, currencies* or financial and economy concepts such as *management, transaction, stock exchange, etc.*

In comparison to biomedical and scientific domains, RE in financial and business domains has received far less attention in the literature. Early works (Costantino et al., 1996a) confirmed that analyzing qualitative data in financial articles from online news or company announcements is as important as processing quantitative data and numbers. Costantino et al. (1996b) extracted three main groups of relevant financial activities that can influence the decisions of players in the financial market. These activities concern either the life of the company such as *bankruptcy*, the restructuring of a company's activities such as *new products*, or general macroeconomics information such as *currency movements*.

Figure 2.9 displays a graph representation of an example of business relations where nodes refer to entities and edges to business relations between them. The text excerpt in (3) (entities are underlined while relations are in bold.) used to generate this graph is extracted from the Wikipedia page of the company *Meta Platforms, Inc.*.

(3) *Meta Platforms, Inc., doing business as Meta and formerly named Facebook, Inc., and The Facebook, Inc., is an American multinational technology conglomerate based in Menlo Park, California. The company* ***owns*** *Facebook, Instagram, and WhatsApp, among other products and services. It* ***has also acquired*** *Oculus, Giphy, Mapillary, Kustomer, Presize and has a 9.99% stake in Jio Platforms. In March 2020, the Office of the Australian Information Commissioner (OAIC)* ***sued*** *Facebook, for significant and persistent infringements of the rule on privacy involving the Cambridge Analytica fiasco.*[5]

[5]Extracted from https://en.wikipedia.org/wiki/Meta_Platforms. Accessed on 29.08.2022

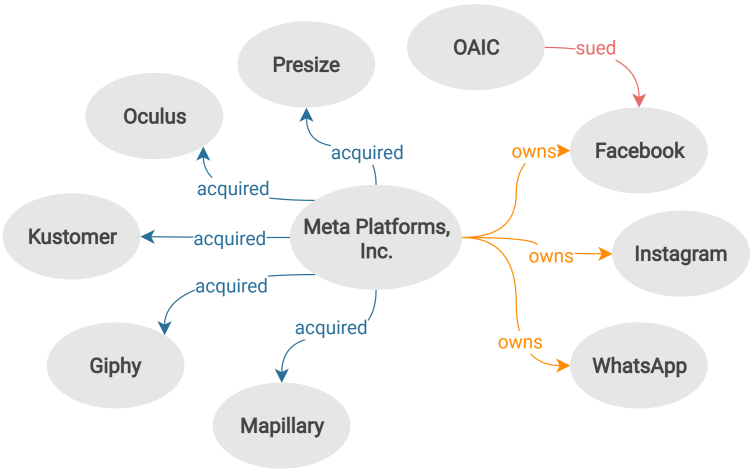

Figure 2.9 Graph representation of business relations extracted from Wikipedia.

The literature has primarily addressed two types of information extraction tasks: relation extraction and event extraction. Event extraction aims at identifying event triggers, their arguments (which can be companies, firms, date, monetary value, etc.), and the event type (Jacobs and Hoste, 2021; Jacobs et al., 2018; Lefever and Hoste, 2016; Qian et al., 2019; Wang et al., 2021b; Xingyue et al., 2021). We focus in this section on relation extraction linking two business entities in sentences, as this is the primary focus of this thesis. In the reminder of this manuscript, the terms *financial* and *business* domains are used interchangeably to refer to professional activities carried out by organizations that involve or do not involve monetary transactions.

2.3.1 Datasets

The key textual data sources for this task include financial newspapers and magazines, company announcements, earnings call transcripts, government websites and databases, and so on. Tweets from communication media users, focusing on economy and financial market in general, were also an additional source of data in some cases (Reyes et al., 2021). Wikipedia was sometimes disregarded because it is primarily concerned with encyclopedic information and less with economic news and announcements (Jabbari et al., 2020).

Different types of relations have been targeted in business textual data. We group them into two categories, according to the type of arguments involved in the relation: relations between financial entities, and relations between generic types of named entities. We detail each category below.

Relations Involving Financial Entities

We consider two main groups. The first one considers relations involving generic concepts such as: `ISA_SUBCLASS` *(team management, management)*. For example, Vela and Declerck (2009) used data from economical news articles from the German newspaper Wirtschaftswoche [6] to extract financial concepts and relations between them. He clearly stated the kind of ontological resource to be extracted from financial documents. `HAS` and `ISA_SUBCLASS` are two examples of such relations.

The second group is about relation linking fine-grained financial entities such as `HAS_VALUE` *(KPI, monetary_value)* where the first argument refers to *key-performance-indicator* that is a quantifiable measurement used to gauge a company's overall long-term performance,[7] and `HAS` *(Company, Asset:financial)* where the second argument refers to liquid asset in finance (e.g., cash, stocks, mutual funds, and bank deposits).[8] Hillebrand et al. (2022) extracted and linked key performance indicators (KPIs), e.g., "revenue" or "interest expenses", of companies from real-world German financial documents. Their task consisted in identifying KPI, their current year and previous year values, and in identifying the increase and decrease of these values. A dataset of 500 manually annotated financial documents containing a total of 15,394 sentences was sourced from the Bundesanzeiger, a platform hosted by the German department of Justice where companies publish their legally mandated documents.

Lately, Jabbari et al. (2020) proposed a compliance-related concepts and relations extraction dataset to help financial intuitions better understand their customers and rigorously monitor their financial activities. There are 6,736 named entities and 1,754 annotated relations in the dataset. It was created by compiling forty daily French financial newspapers and annotating them with a fine-grained ontology. The authors defined 26 fine-grained entity types, such as *Organization:Association* and *Asset:FinancialAsset*, as well as a set of 20 relation types, such as `ownedBy` to refer to the organization that owns an asset, or `hasCondemned` to refer to the party who was sentenced to prison (cf. Figure 2.10 for proposed entity types and Figure 2.11 for the relation types considered). Some of these relationships were designed for a specific application domain. Thus, they are not always adaptable to new ones.

[6] https://www.wiwo.de/
[7] https://www.investopedia.com/terms/k/kpi.asp
[8] https://www.investopedia.com/terms/f/financialasset.asp

Entity Type	**Definition**	**Examples (French)**
Person	Physical Persons	*Emmanuel Macron, Carlos Ghosn*
Organization:Association	Unions, clubs, and NGOs. It was not used due to its juridical ambiguity.	*CGT, Greanpeace*
Organization:Company	Private or public corporation	*Total S.A., Airbus*
Organization:GPE	National/International geopolitical entities	*gouvernement français, UE*
Organization:Media	Press and broadcasts	*Le Figaro, BBC*
Location:WorldRegion	World regions and continents	*Asie, moyen orient*
Location:Country	World countries	*France, USA*
Location:LocalRegion	Local regions, states and provinces	*Californie, Normandie*
Location:City	Cities, towns and urban areas	*Grand Londre, New York*
Role	Professional roles and positions	*PDG, patron*
Currency	Currencies and their symbols	*Euro, CHF, $*
Asset:FinancialAsset	Non-physical assets	*actions, obligations*
Asset:TangibleAsset	Physical assets and goods	*immobilier, véhicules*
Asset:MoneyAmount	Monetary amount without its currency	*dix millions, 10,35*
Document	Offical documents, diplomas, etc.	*passeport, contract*
Financing	Financings and investments	*financement, investissement*
Merger	Consolidation of two or more companies into one company	*fusion, opération M&A*
Demerger	Converse of a Merger	*scission*
Acquisition	Transfer of ownership of a company (Acquiree) to another (Acquirer)	*acquisition, rachat*
Activity	Economical sector of activity	*aéronautique, énergie*
IPO	Initial Public Offering	*introduction en bourse*
Penalty:Fine	Financial punishments	*amende, contravention*
Penalty:Imprisonment	Prison sentence	*peine de deux ans, condamnation*
Penalty:Sanction	Embargo and sanctions	*sanctioné, interdit*
Shareholding	This category was finally not used in annotations	*actionnarit*
BusinessDeal	Commercial transactions, i.e. payments and purchases	*transaction, paiement*

Figure 2.10 Entity types according to (Jabbari et al., 2020)'s proposed financial ontology.

Relation	Definition	Argument 1	Argument 2
hasRole	Occupation of a position in an organization	Person	Role
roleDepartment	Attachment of a position to an organization	Role	Organization
hasIncomingParty	Entering Parties	Merger/Demerger	Company/Organization
hasOutgoingParty	Outcome party of a merger/demerger	Merger/Demerger	Company/Organization
hasAcquirer	Initiator of an acquisition	Acquisition	Company
hasAcquiree	Subject of an acquisition	Acquisition	Company
hasActivity	Participation of an organization in a sector of activity	Organization	Activity
hasHQ	Having headquarters in a location	Organization	Location
hasNationality	Having citizenship of a country	Person	Country
isLocated	Localisation of an entity	*	Location
owns	Ownership of assets	Person/Organization	Asset
ownedBy	Converse of ownership relation	Asset	Organization/Person
hasCurrency	Corresponding currency of a money value	MoneyAmount	Currency
hasObject	Contract object of a transaction	BusinessDeal/Financing	Asset/MoneyAmount
hasCreditor	beneficiary of a transaction	BusinessDeal/Financing	Activity/Organization/Location
hasDebtor	Initiator of a transaction	BusinessDeal/Financing	Person/Organization
hasIPOCompany	Company going public in the IPO	IPO	Company
hasIPOMarket	IPO's launch market	IPO	Organization/Location
hasCondemned	The condemend party of a penalty sentence	Penalty	Person/Organization
hasIssuingAuthority	The authority issuing the penalty	Penalty	Organization

Figure 2.11 Relation types with possible involved entity types according to (Jabbari et al., 2020)'s proposed financial ontology.

More recently, the extraction of financial relations from Chinese news was also investigated in (Wu et al., 2020). A dataset of 55,032 instances was collected accounting for 14 relation types between three types of entities including *companies, financial and security institutions*.

Relations Involving Generic Named Entities

According to Zhao et al. (2010), these relations can be either:

- *Inner-Organizational* (Inner-ORG) for relationships involving a company and its components including persons, locations, statistics, and time. For example, the relation `company-manager` is an Inner-ORG relation between an organization and person.
- *Inter-Organizational* (Inter-ORG) for relationships involving two entities of type *Organization* such as: *startups*, *companies*, *non-profit organizations*, *universities*, etc. The relation `company-partner` is an example of this type.

The first one concerns relations involving different generic types of named entities such as *persons (PER), organizations (ORG), locations (LOC)*, etc. as in `EMPLOYER_OF`*(person, organization)*, `HEADQUARTER` *(organization, locations)*. These types of relations are the most studied in the literature. The collected datasets were either manually annotated by domain

experts or built relying on a distantly supervised approach using a KB. Most of the KB used are generic-purpose KBs, such as DBPedia, Wikidata, or Freebase where either a list of financial relations or an expert-built financial ontology are used to select the training data.

For the distantly supervised datasets, Aljamel et al. (2015) constructed a knowledge map for the financial domain focusing on generic named entities such as organizations, people, and locations, and the relations between them. This ontology is mapped to publicly available knowledge bases, such as DBPedia and Freebase, and is used to automatically annotate a set of sentences collected from 7,193 documents of online financial news sources including the BBC, Reuters, and Yahoo Finance RSS Feeds. More recently, FinRED was proposed (Sharma et al., 2022). It is a relation extraction dataset curated from financial news and earning call transcripts containing relations from the finance domain. The dataset was created using a distant supervision method with a subset of the Wikidata KB as a source of supervision, containing manually filtered 29 financial relations. The generated corpus consists of 7,775 sentences covering 29 types of relations between named entities of type ORG, PER, LOC, such as `owned_by`, `founded_by`, `parent_organization` or nominal like: `product/material_produced`.

Had et al. (2009) stated that a simple co-occurrence of entities in a single sentence is insufficient to conclude that they are positively linked with a relation (distant supervision assumption); thus an annotation procedure is required. Their work monitored economic information on the internet by tracking economic relations between companies in German online news, particularly the merger relationship. By crawling the web for information on the 30 DAX-indexed German companies, a corpus of 1,698 sentences with 3,602 relation candidates was manually created. 2,930 of these relation candidates are negative (there is no merger-relation), while 672 are true merger-relations. In (Plachouras and Leidner, 2015), experts manually annotated a set of 200 records in the World-Check database to evaluate the extraction of regulatory fine fact instances paid by a company. The task entails identifying dates, monetary amounts, causes, and regulators in the sampled records.

Reyes et al. (2021) extracted all potential semantic relations expressed using one or more tokens in business articles. To this end, they manually annotated a corpus with a total of 4,641 records, expressing relations at the sentence level. Conversely, De Los Reyes et al. (2021) only considered the identification of business relation between different types of named entities, without classifying them into more detailed business relation types. They manually annotated a dataset of 3,288 records to determine whether two companies in a sentence are related. The dataset contains 1,485 instances (45%) that are positive tuples, meaning they have a relationship between the highlighted entities, and 1,803 instances (55%)

that are negative tuples, meaning there is no relationship between the entities.

On the other hand, from a market intelligence and competitive point of view, Zhao et al. (2010) classified Inter-ORG between entities of type *ORG* into four types of relations, which are `cooperation` relation, `invest` relation, `sales` relation, and `supply` relation. The Inter-ORG is the primary focus of this dissertation.

Generally, these relations are marginally present in KBs, such as DBpedia, (Auer et al., 2007) where relations like `Subsidiary` and `Ownership_of` can be found. In addition, some of these relations are annotated in generic relation datasets with fairly low frequencies, such as `Employment/Membership/Subsidiary` in the ACE 2004 dataset (Mitchell et al., 2005), and `org:subsidiaries`, `org:shareholders` or `org:parents` in TACRED dataset (around 453 , 144, and 444 instances respectively) (Zhang et al., 2017b).

Yamamoto et al. (2017) thought that the two key relations, `cooperation` and `competition`, are the most crucial ones to extract between firms for a macro and micro overview of the industry structure for a specific domain. They collected 4,661 news articles written in English from online news sites that focus on the semiconductor industry field, dated from March 2009 to March 2015. First a list of 65 companies that each appear over 100 times in articles was extracted. Then 427 articles were manually selected in term of relations. The labeled data were used as seed knowledge for news articles in distant supervision including 46 `cooperative` relations and 27 `competitive` relations. In addition, lists of keywords per relation were extracted including 15 words for `cooperative` and 17 words for `competitive` to be used as rules of distant supervision. The final dataset contains 15,599 relation candidates with 464 *cooperative* relations and 732 `competitive` relations.On the other hand, Braun et al. (2018) was interested in extracting `owns`, `funds`, and `cooperation` relation types to describe the constitution of smart city ecosystem. They only collected 41 news articles and blog posts in German to construct an evaluation dataset. The manual annotation of these articles resulted in the following distribution of relations: 15 instances for `owns`, 14 for `cooperation`, and 12 for `funds` relation.

Besides, Zuo et al. (2017) focused on the extraction of `ownership of` between companies, where 359,459 articles were selected from NYTimes corpus containing entity pairs occurring in "Technology" and "Business" categories. The first 100 most frequent pairs are manually annotated and used as a starting seed for their extraction model.

Finally, Yan et al. (2019) manually annotated 1,000 instances of relation between companies [9] from corporate news reported by the chinese online media platforms.

[9]Targeted relation types are not specified in their paper.

Summary

Table 2.3 summarizes the main characteristics of the previously cited datasets. We consider in this table different comparison criteria, as follows:

- Relation level (Rel. level): sentence (sent.), or document (doc.);
- Number of annotated relation instances (#Inst.), with the number of business relation instances (#Biz.);
- Number of relation types considered (#Rel.);
- The language of the dataset (Lang.);
- The annotation method: either manual annotation (gold), or distantly supervised annotation (DS);
- Examples of entity types included in the dataset (Example ent. type);
- Examples of relation types considered (Example rel. type);
- If the dataset is available online or not (Av.?).

From the table, we can note that datasets in different languages have been proposed including: English (EN), Chinese (ZH), French (FR), German (DE), and Portuguese (PT). The number of business relations considered vary between 2 and 29 types, with no unified annotation schema. The datasets contain up to 55k instances, where most of them however are not available for the research community.

2.3.2 Main Approaches

As we focus on Inter-ORG relations in this dissertation, this section will be dedicated to it.

Different approaches have been used to extract business relations from text. Zhao et al. (2010) presented a method for extracting relations between people and organizations that leverages structural forms of conveying a person's position within an organization in text such as colon (:), HTML column tags (<td>), and so on. Braun et al. (2018) employed machine learning approaches to produce dependency parsing trees, then created a set of rules that exploit both syntactic dependencies between words from these trees and lexical traits from a list of terms related to the relations to extract. Using *nsubj*[10] and *nsubjpass*[11] links in

[10]The dependency type nsubj marks nominal subjects of a clause. Subjects are direct dependents of the main predicate of the clause, which may be a verb, noun, or adjective.

[11]A passive nominal subject is a noun phrase which is the syntactic subject of a passive clause.

Table 2.3 Overview of datasets for business relations extraction.

DATASET	Rel. level	#Inst. (#Biz.)	#Rel.	Lang.	Annot.	Example ent. type	Example rel. type	Av.?
(Sharma et al., 2022)	sent.	7,775	29	EN	DS	Generic concepts Organization Currency Monetary value	Product Manufacturer Industry Position held Original broadcaster Owned by Founded by Distribution format Headquarters location Stock exchange	Y
(Jabbari et al., 2020)	sent.	1,754	20	FR	gold	Person Penalty:Sanction BusinessDeal Financing	HasRole HasIncomingParty HasCreditor HasCondemner	N
(Wu et al., 2020)	sent.	55,032	14	ZH	rules	Company Bank Securities institution	Sue Debt Warrant Asset Swap Entrust Finance Pledge	N
(Hillebrand et al., 2022)	sent.	15,394	7	DE	gold	KPI Current year monetary value	Relation identification	N
(Aljamel et al., 2015)	sent.	21,582 (1,399)	8	EN	DS	Organization Person Location	EmployerOf FounderOf KeyPersonIn LocatedIn	N
(Had et al., 2009)	sent.	3,602 (672)	2	DE	gold	Organization	Merger	N
(De Los Reyes et al., 2021)	sent.	4,641	* ‡	PT	gold	Organization Person Location	Has just close a contract with In partnership with	Y
(Reyes et al., 2021)	sent.	3,288	2	PT	gold	Organization Person Location	Relation identification	N
(Yamamoto et al., 2017)	sent.	15,599 (1,178)	3	ZH	DS	Organization	Competition	N
(Braun et al., 2018)	sent.	41	3	DE	gold	Organization	Funds Owns	N
(Yan et al., 2019)	sent.	1,000	/	ZH	gold	Organization	/	N

‡All possible verbal expressions are extracted.

the dependency tree, their system was also able to determine the direction of the extracted relations.

Pattern construction being time-consuming, Zuo et al. (2017) suggested a semi-supervised strategy that takes only a small number of manually specified company pairs to efficiently extract new ones that belong to the same target relationship. Their model is based on Snowball (Agichtein and Gravano, 2000), a system that uses a seed of related entity pairs to uncover relation patterns from their context, then uses these patterns to extract new entity-pairs that are added to the beginning seed The pattern extraction phase is refined by employing a key-extraction method, which aids in the removal of irrelevant context around the company pairs. Lau and Zhang (2011) used a set of relationship indicators to construct a statistical inference approach for mining `cooperative` and `competitive` relations from news. The authors began by selecting a collection of relations indicators, which were then augmented by synonym linkages from WordNet. Relation patterns are defined from sentences collected from the web that contain a pair of entities and a relation indication. The entity pair is given the relation type with the most patterns based on the number of derived patterns per relation type.

Feature-based methods were also explored. For example, Yamamoto et al. (2017) trained DeepDive (Zhang, 2015), a machine learning system based on Markov Logic Network, on a set of handcrafted features including n-gram, POS tags, named entities tags around companies, and the number of words between the two companies. They achieved a precision of 67% for `cooperative` relations and 81% for `competitive` relations.

Business relation extraction has recently benefited from neural architectures performances. Most of proposed models leveraged lexical and syntactic representations including word embedding, relative positions, POS, entity type, dependency type, etc. of the input sentence to learn feature vectors using neural architecture like Bidirectional Gated Recurrent Unit (Yan et al., 2019; Yang et al., 2020a). The experimental results demonstrated the effectiveness of such models. Collovini et al. (2020) integrated prior NER and RE Portuguese methods proposed in the literature into a unified framework for discovering organizational relationships. Their RE module is based on the RelP system (de Abreu and Vieira, 2017), which was designed for open relation extraction in Portuguese literature. A CRF model is trained to identify relation descriptors among other words based on a set of extracted attributes such as POS tags and syntactic tags.

Recently, the BERT model has been utilized to find semantic relations between entities in financial and economic texts without the usage of any other lexical-semantic resources (De Los Reyes et al., 2021; Reyes et al., 2021).

Other studies pre-trained the BERT model on large financial data to adapt it to financial related tasks. Four new models have been proposed, all named FinBERT and are targeting English language. Araci (2019) was the first to offer FinBERT as a pre-trained domain-adapted BERT by retraining it in multitask fashion on two datasets: TRC2-financial, [12] which includes 46,143 documents from a set of Reuters news stories with more than 29M words and approximately 400K sentences, and Financial Phrasebank consisting of 4845 English sentences selected randomly from financial news found on LexisNexis database.[13] During testing, they observed a considerable improvement in results reaching 15% increase in classification task accuracy. The next adaptation of BERT was proposed by DeSola et al. (2019) who trained it on 497 million words of 10-K files from 1998 to 1999 and 2017 to 2019, and it outperformed BERT on the masked LM and next sequence prediction tasks.[14]

Yang et al. (2020b) collected a corpus of three types of data based on financial and business communications of companies: 10-K and 10-Q reports, earnings call transcripts, and analyst reports totaling 4.9 billion word tokens. They demonstrate that their model considerably outperforms BERT in three sentiment analysis tasks. Lastly, Liu et al. (2021c) trained their FinBERT on three financial corpora: 13 million financial news (15 GB) and financial articles (9 GB) from Financial Web, totaling 6.38 billion words; financial articles from Yahoo! Finance (19 GB), totaling 4.71 billion words, and question-answer pairs about financial issues from Reddit (5 GB), totaling 1.62 billion words. During testing, their model outperformed BERT on all financial tasks in terms of accuracy, precision, and recall. More recently, Xia et al. (2022) proposed a framework that combines both BiLSTM and language models to represent long financial text documents via sequential chunking.

Most of these adaptations produced the state-of-the-art results on financial tasks, including Sentiment Analysis and Question Answering. However, they were not assessed on business relations extraction because there is no widely used benchmark dataset in the literature for this task.

2.3.3 Applications

Business relation extraction has proven to be essential in a broad range of business applications. The extracted relations are typically structured into company networks, which provide further information about the market it shapes.

For example, these networks have been used to obtain competitive intelligence from the web by extracting information about rivals and finding their relationships with the

[12] https://trec.nist.gov/data/reuters/reuters.html
[13] https://www.researchgate.net/publication/251231364
[14] https://github.com/psnonis/FinBERT

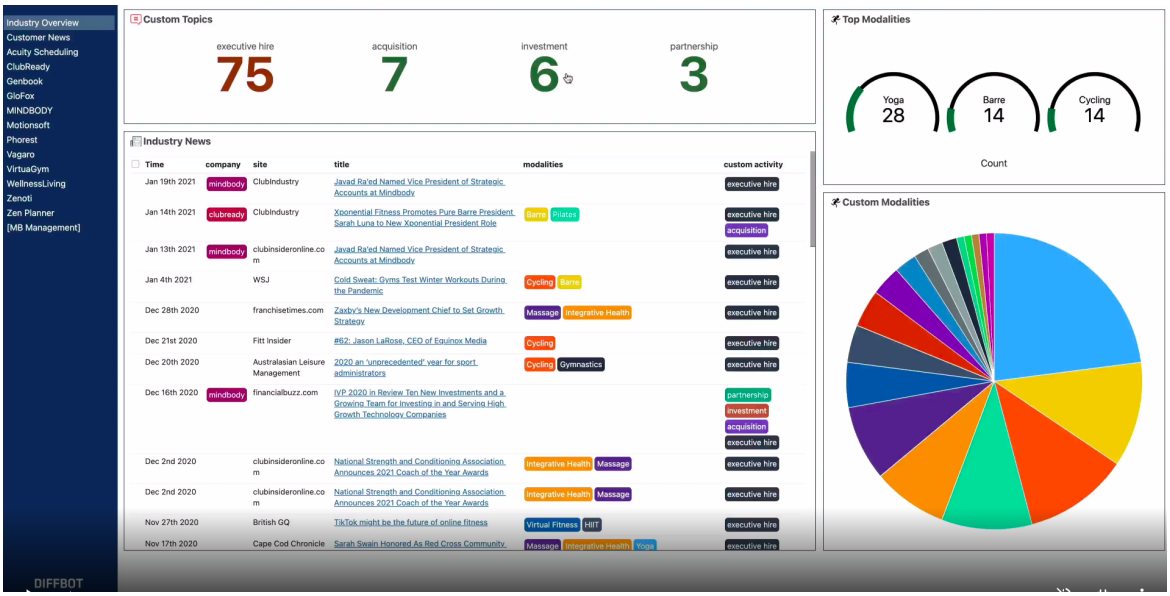

Figure 2.12 Fitness and well-being market news monitoring using DiffBot.

ecosystem they evolve in, such as clients and suppliers (Collovini et al., 2020; Zhao et al., 2010). Figure 2.12 displays a brand monitoring solution by analysing news online using Diffbot software.[15] The extracted knowledge demonstrates to managers how to position their company as competitive in the market. Analyzing a segment of companies network also helped professionals to analyze the structure and the ecosystem of a certain industry, such as smart cities related industry (Braun et al., 2018; Yamamoto et al., 2017).

In particular, financial institutions must now understand the entire financial market, particularly the network of enterprises in which they invest (Schwenkler and Zheng, 2019; Yan et al., 2019). Figure 2.13 depicts the network of firms active in zero-carbon and low-carbon batteries ecosystem generated by the GEOTREND[16] platform, that will be detailed in Chapter 6.

The network structure of the enterprises allows for the examination of various financial scenarios, such as the impact of business bankruptcy on other market participants in the network. In this situation, the linkages between individual market participants can be used to predict which companies and how much are affected by bankruptcy (Repke and Krestel, 2021). As a result, such business relation extraction systems can help financial organizations make risk management decisions for pre-lending, lending, and post-lending operations.

[15] https://www.diffbot.com/solutions/news-monitoring/
[16] https://www.geotrend.fr/fr/

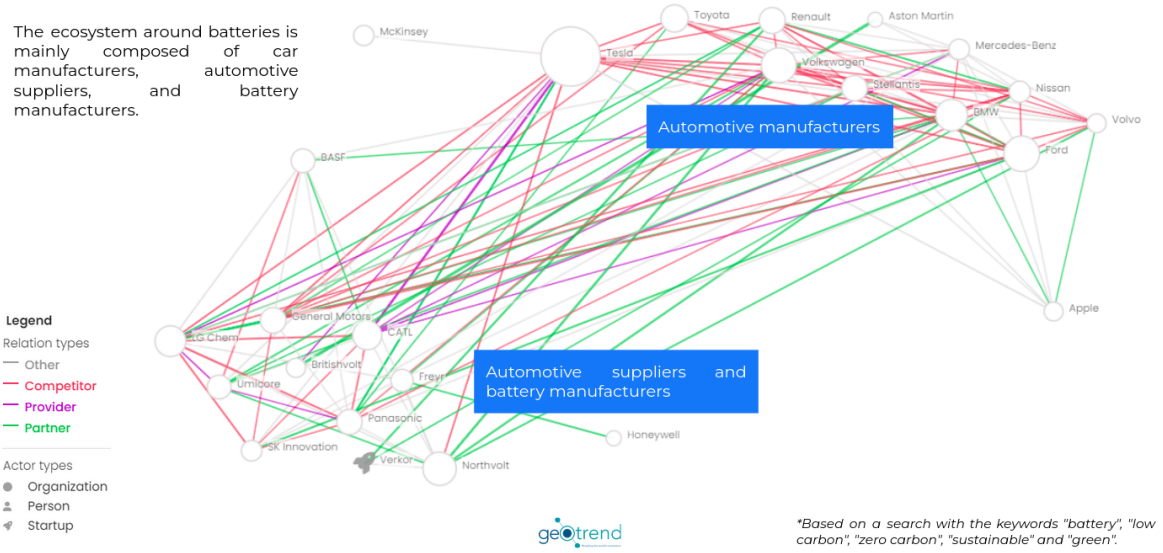

Figure 2.13 Enterprise network analysis for zero-carbon and low-carbon batteries ecosystem.

2.4 Conclusion

The goal of this chapter was to provide an overview of three domain-specific relations extracted from textual data: scientific, biomedical, and business. For each domain, we presented the proposed datasets and models as well as their domain of applications. Overall, the following findings may be drawn from these three domain-specific RE:

- **Dataset construction.** Manually annotating large amounts of data necessitates requires human expertise and a large amount of resources. As a result, knowledge bases have been used to generate large training datasets under distant supervision. However, this method has two major drawbacks:
  - The distant supervision assumption, which states that "if two entities participate in a relation, any sentence containing those two entities may express that relation" (Mintz et al., 2009), is too strong and may generate a large number of false positive instances, lowering the quality of the generated data;
  - The availability of knowledge bases for specific domains is not always possible because their creation needs a significant amount of human effort and expertise, and some domains may receive less attention from the research community and thus are less likely to be targeted.
- **Data languages.** The majority of the proposed works and datasets are in English, with minor efforts in German, French, Portuguese, and Chinese. Almost no work has been

done to extract multilingual domain-specific relations because multilingual datasets are not available in the literature, as far as we know.

- **Proposed models.** When compared to other supervised approaches, neural-based RE models performed better for the three domain-specific relations. In particular, transformer-based models achieved very well. When supplemented with external knowledge regarding entity pairs, these models significantly boost RE performances for scientific and biological domains.These latter models, however, were not investigated for inter-organizational business relationships.

We focus in this dissertation on business relation extraction, which is one of the least studied relations in the literature. Most of the works share three main limitations:

- The proposed models are based on simple architectures, ignoring external knowledge about involved entities;
- The datasets focus on one language at a time;
- The proposed models are evaluated on different datasets, which are either small in size or not freely available to the research community, making comparison between different proposed works difficult.

Focusing particularly on Inter-ORG relations, the chapters that follow describe some potential solutions to these limitations, including:

1. A unified characterization for business relations between organizations focusing on five relations: INVESTMENT, COOPERATION, SALE-PURCHASE, COMPETITION, and LEGAL PROCEEDINGS;
2. BIZREL, the first manually annotated multilingual dataset annotated according to this characterization and considering four languages: *French*, *Spanish*, *English*, and *Chinese*;
3. A set of neural-based experiments to validate the proposed dataset;
4. A new approach for incorporating external entity knowledge into the business relation extraction systems.

In the following chapter, we will begin to present our first and second contributions.

Chapter 3

BIZREL: A Multilingual Business Relations Dataset

This chapter describes BIZREL, the first multilingual (French, English, Spanish, and Chinese) dataset for automatic extraction of binary business relations involving organizations from the web. We start by going over the resource's construction process, including how data were collected (cf. Section 3.1), how business relations characterization was defined (cf. Section 3.2.1), then, we describe the annotation process (cf. Section 3.2.2) and finally quantitative results about the constructed dataset (cf. Section 3.2.3).

This dataset is used to train several monolingual and cross-lingual deep learning models to detect these relations in texts (cf. Section 3.3). Our results are encouraging, demonstrating the effectiveness of such a resource for both research and business communities.

3.1 Data Collection

Our corpus has been collected from the open web including multiple sources of data such as news articles, companies websites, and Wikipedia articles. Figure 3.1 depicts the general process for data collection. It is composed of three main steps:

Figure 3.1 Data Collection Process

Table 3.1 Top 18 generic and specific keywords used to collect EN data.

GENERIC KEYWORDS	SPECIFIC KEYWORDS
news	hi-tech
press	autonomous car
newsreel	publication
information	e-commerce
economist	agriculture digital platform
market	cloud computing
intelligence	EV cars versus Gasoline
venture	telecom operator
leader	autonomous delivery
technology	5G infrastructure
competitor	biotechnology
business	Electric battery
investment	battery recycling
partnership	climate change
purchase	global warming
partner	system on chip
contract	virtual reality
research	e-cigarette

- **Document Collection.** A set of business domain's related keywords is constructed. We begin with a starting seed containing specific keywords such as *3D printing, flying cars, food delivery, etc.*, then we expand it using generic expansion keywords such as *technology, market, business, etc.*, to refine our search and orient the obtained results to be more about markets news and information. The set of keywords does not contain any named entity to prevent biases toward any named entity in our data. The selected key-words are then used to request the two search engines Google and Bing through their search APIs. The language of the collected data is specified as a parameter. Only the first top 1,000 web pages are selected. We consider textual contents from various sources (online news, company websites, industry reports, etc.) and formats (web page, PDF, word), while excluding those retrieved from social media, e-commerce, and code versioning websites. Table 3.1 displays the top 18 generic and specific keywords in terms of the number of documents returned. [1] The collected documents are cleaned then segmented into sentences, with duplicates removed.

[1]The complete list of keywords used cannot be shared due to confidentiality concerns.

- **Named Entity Recognition.** Named entities of type *Organization* are extracted from the identified sentences using the two taggers spaCy[2] and StanfordNLP.[3] After identifying all named entities of type *Organization* (henceforth EO), each sentence is repeated as many times as the different EO pairs it contains. This is done to account for all possible relationships between entities in a single sentence. The four extracted EOs are underlined in the sentence (1). As a result, six EOs pairs are present, and the sentence is repeated six times to account for all possible relations between these EOs pairs.

  (1) Amazon's recent investment in British food delivery company Deliveroo late last week is being widely hailed as a "shot across the bow" of Uber Eats, Grubhub, and a bevy of other mobile food delivery market players, but it is much more than that.

- **Relation Candidate Selection.** Relation instances are selected according to two rules:
  - only sentences containing entities recognized by both taggers are retained, to prevent error propagation and guarantee the quality of selected named-entities.
  - sentences whose words are at least 95% of type ORGANIZATION are discarded, this mainly concerns enumerations of organizations.

This procedure resulted in a total of 25,469 sentences for French, English, Spanish, and Chinese. Table 3.2 describes sentences distribution and complexity per language. To measure the complexity of business relations and their syntactic richness, we compute the minimum, maximum, and average count of words, verbs, and entities per relation instance and per language (*Nb. word_per_sentence*, *Nb. verb_per_sentence*, and *Nb. entity_per_sentence* respectively), the ratio of unique entity-pairs in the dataset (*RatioU. e_pairs*), and the ratio of unique entity pairs in the dataset (*RatioU. entity_pairs*). Sentences in our dataset contain on average from 5 to 7 *EOs*, therefore, potentially a maximum of 10 to 21 relations could occur in a single sentence between different *EOs* pairs. In addition, sentences are complex containing in average 2 verbs and the context surrounding a given relation instance is 39 tokens on average for English data (41, 34, 50 for French, Spanish, and Chinese respectively).

Moreover, 77% of *EOs* pairs in English data (53%, 42%, 52% in French, Spanish, and Chinese respectively) are unique reflecting entity pairs disparity in the dataset. Overall, these measures confirm the diversity and complexity of business relations expressed in our

[2]https://spacy.io/
[3]https://stanfordnlp.github.io/CoreNLP/ner.html

Table 3.2 Statistics about relation candidates complexity.

STAT. ↓ / LANG. →	**EN**	**FR**	**ES**	**ZH**
Nb. sentence	10,034	10,033	3,085	2,316
Nb. word_per_sentence	min. = 4 avg.= 39 max. = 352	min. = 5 avg.= 41 max. = 258	min. = 6 avg.= 34 max. = 213	min. = 10 avg.= 50 max. = 233
Nb. verb_per_sentence	min. = 0 avg.= 2 max. = 24	min. = 0 avg.= 2 max. = 21	min. = 0 avg.= 2 max. = 20	min. = 0 avg.= 5 max. = 46
Nb. entity_per_sentence	min. = 2 avg.= 6 max. = 84	min. = 2 avg.= 5 max. = 34	min. = 2 avg.= 6 max. = 33	min. = 2 avg.= 7 max. = 37
RatioU. entity_pairs (%)	77	53	42	52

dataset. This is more salient for the Chinese language, where the average number of verbs per sentence is the most important.

3.2 Data Annotation

3.2.1 Characterizing Business Relations

As stated in Section 2.3.1, Chapter 2, we focus in this dissertation on business relations expressed between generic entities in sentences. More specifically, our main interest is on business interactions between organizations such as *companies*, *startups*, *non-profit organizations*, *governmental entities*, *universities*, etc.

First, we consider four types of Inter-ORG business relationships as proposed by the initial work by Zhao et al. (2010) who : defined an ontology for business relations (cf. Figure 3.2) :

- COOPERATION referring to the contracted cooperation between two companies,
- INVEST referring to companies that buy some stocks of other companies as a future investment,
- SALES referring to the customers of a company,
- and finally SUPPLY referring to the suppliers of a company.

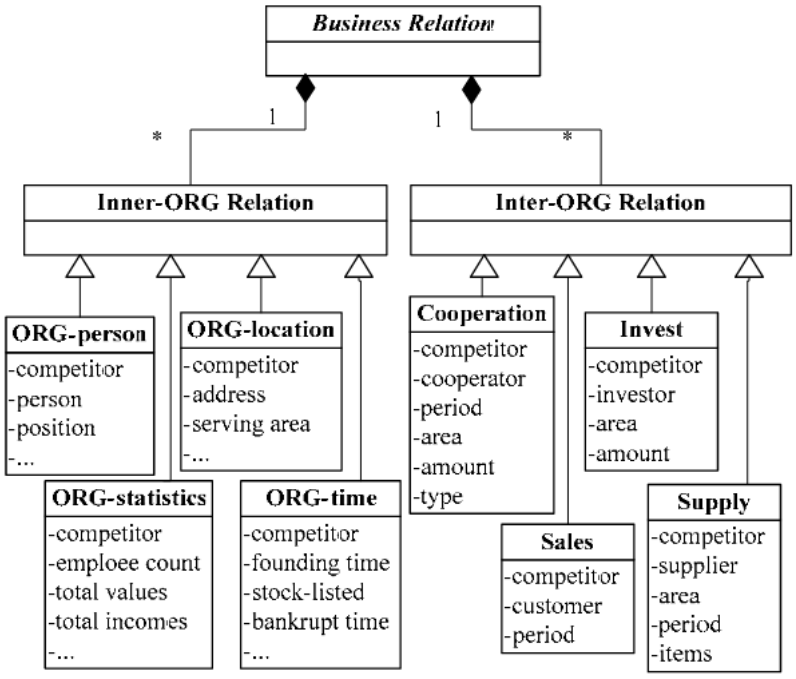

Figure 3.2 The ontology of business relations as defined by (Zhao et al., 2010)

These relationships can be time-bound, for example, any partner can end the partnership and effectively bring the business to a close.[4] We argue that including the time component in representing business relations would make the task more difficult because an additional step would be required to verify the validity of an extracted relation in time. We decide to only use lexical and semantic patterns to identify these relationships, so the extraction is time-independent.

Besides, a SALE relation refers to the immediate purchase of a good or a service in consideration of a cash amount paid by the purchaser, whereas a SUPPLY relation refers to a temporal provision of certain quantities of goods or services over a period of time in consideration of a price or wage paid by the purchaser.[5] Given the semantic similarity between the two relations (cf. Examples (2) and (3)) and since they both involve a seller/provider and a purchaser, we only focus in our work on one relation that we call SALE-PURCHASE that refers to both SALE and SUPPLY relations.

(2) Trading in dollars between two European companies may seem incongruous. Yet, this is the norm in the aeronautics industry, except for a few rare exceptions, which are mostly ignored. But when the euro keeps falling against the greenback, Air France-KLM

[4] https://business.vic.gov.au/business-information/exit-your-business/dissolve-a-business-partnership
[5] https://www.alsaadiadvocates.com/media/difference-between-sale-and-supply-contract

can be pleased that it has managed to buy its latest aircraft directly from Airbus in euros.[6]

(3) Pratt & Whitney and General Electric have a joint venture, Engine Alliance selling a range of engines for aircraft such as the Airbus 380 of Airbus.[7]

The definition of business relations used by (Zhao et al., 2010) was entity-centric, with the goal of acquiring intelligence about competitors of a given business entity from the web. Following the works done by (Lau and Zhang, 2011; Yamamoto et al., 2017), we first add a new business relation type, COMPETITION, to the pool of four business relations defined earlier to model competitiveness. Then, both papers consider another COMPETITION sub-relation to refer to LEGAL-PROCEEDINGS between two competing business entities. Here, we separate these two types because they are semantically distinct (cf. Examples (4) and (5)), resulting in a final characterization of five business relations.

(4) Some key vendors of the global green packaging market are Amcor, Berry Plastics, BASF, DuPont, Printpack, Inc., Innovia Films Ltd, Bemis Company, Tetra Laval, and Ball Corporation among others. [8]

(5) Vans suing Primark for selling £8 'intentional copies' of its £55 trainers. [9]

As stated in Section 1.1.3, Chapter 1, two target entities cited in the same sentence may have no semantic relationship. In addition, given the pre-defined set of business relations of interest, any other semantic relationship between target entities that does not belong to it is not intended to be extracted. To account for the aforementioned scenarios, we include OTHERS relation type in our relation characterization.

Figure 3.3 displays the final characterization of business relations in our dataset. Relations definitions along with their examples will be presented in the remain of this section. It is important to note that the orientation of the business relationship to extract is ignored in our characterization, i.e., $R(S, EO_1, EO_2) = R(S, EO_2, EO_1)$, with EO_i being named entities of type ORGANIZATION and R being the relation between them expressed in the sentence S.

Our relations are defined below, along with examples taken from our annotated datasets (French, Spanish, and Chinese instances are provided together with their English translations).

[6] Source of the translated text using DeepL.

[7] https://airpowerasia.com/2020/12/20/major-aircraft-turbofan-engine-manufacturers

[8] https://www.globenewswire.com/news-release/2017/11/08/1177736/0/en/Green-Packaging-Market-Size-to-Grow-US-218-50-Billion-by-2021-Zion-Market-Research.html

[9] https://www.thesun.co.uk/fabulous/8089618/vans-suing-primark-trainers-copy/

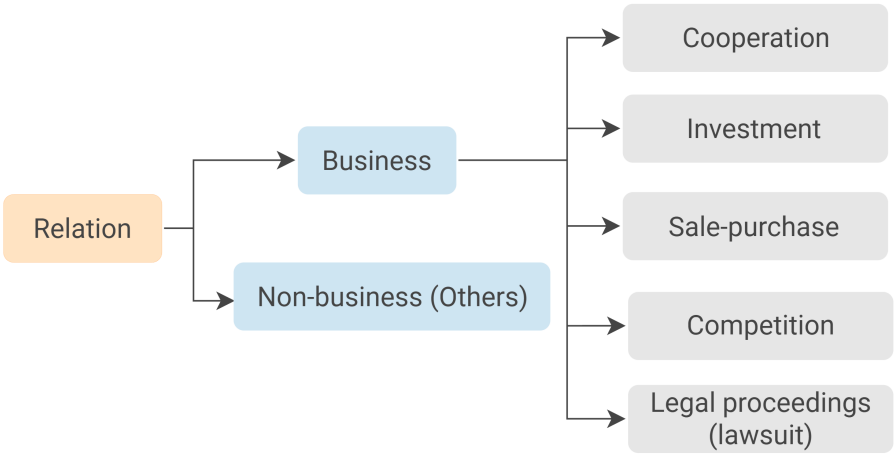

Figure 3.3 BIZREL business relations characterization.

In each example, the two entities involved in the relation are underlined.

- **INVESTMENT:** an *EO* is a subsidiary of another *EO*, or *EO* holds (all or part) of the shares of another *EO*, which means it either acquires it or invest in it.

  In sentences (6), (7), (8), (9), and (10), EO_2 is a part of EO_2. The relation is explicitly expressed using lexical terms such as *its branch*, *a subsidiary of*, and *owned by*, in sentences (6), (7) and (8), respectively, or expressed implicitly using punctuation: "*( )*" in sentence (9).

  (6) 据路透中文网23日报道，[诺基亚]$_{EO_1}$ 表示，计划对旗下法国分支 [阿尔卡特- 朗讯]$_{EO_2}$ 裁撤1233个岗位，相当于该部门总员工数的三分之一。

  (According to a Reuters Chinese website on the 23rd, [Nokia]$_{EO_1}$ stated that it plans to abolish 1,233 positions in its French branch [Alcatel-Lucent]$_{EO_2}$, which is equivalent to one-third of the department's total employees.)

  (7) Developed and delivered by The Climate Corporation, a subsidiary of [Bayer]$_{EO_1}$ and a leader in digital innovation for agriculture, [FieldView]$_{EO_2}$ is the most complete digital farming platform on the market, Bayer says.

  (8) The FAA is also working with PrecisionHawk and [BNSF Railway]$_{EO_1}$ (owned by [Berkshire Hathaway]$_{EO_2}$) to test commercial BVLOS drones — but under very controlled conditions, far away from other people.

(9) [Snecma]$_{EO_1}$ ([Safran]$_{EO_2}$) signed an agreement with AFI KLM E&M to carry out LEAP engine development tests.

(10) Seven of [Unilever]$_{EO_1}$'s top ten brands – Dove, [Knorr]$_{EO_2}$, Omo/Persil, Rexona/Sure, Lipton, Hellmann's and Wall's ice cream – are all Sustainable Living Brands.

In sentences (11), (12), and (13), EO_1 is investing in EO_2 or owns/acquired stocks in it.

(11) [Amazon]$_{EO_1}$'s recent investment in British food delivery company [Deliveroo]$_{EO_2}$ late last week is being widely hailed as a "shot across the bow" of Uber Eats, Grubhub, and a bevy of other mobile food delivery market players, but it is much more than that.

(12) After receiving approval from the State Council of China, it was announced that on 2 September 2007, [Singapore Airlines]$_{EO_1}$ and Temasek Holdings (holding company which owns 55% of Singapore Airlines) would jointly acquire shares of [China Eastern Airlines]$_{EO_2}$.

(13) On 6 April 2006, [BAE Systems]$_{EO_1}$ planned to sell its 20% share in [Airbus]$_{EO_2}$, then "conservatively valued" at €3.5 billion (US$4.17 billion).

$EO1$ acquired EO_2, is expressed in sentences (14) and (15) in two different ways.

(14) [Uber]$_{EO_1}$, the US ride-hailing company, purchased Middle Eastern rival [Careem]$_{EO_2}$ for $3.1 billion in its biggest acquisition to date.

(15) Société Générale to sell [Société Générale Serbia]$_{EO_1}$ to [OTP Bank]$_{EO_2}$.

- **COMPETITION:** a competition/rivalry between two EOs providing the same goods or services, or wanting to access the same relatively small market. It is important to note that two EOs using the same technology does not necessarily mean that they are in competition. In sentences (16), (17), (18), the two target entities EO_1 and EO_2 are operating in the same market (*biopharmaceutical industry*,*self-driving car market*, *fintech companies and banks*, respectively), thus they are in a competition.

(16) Shaw has held positions of increasing influence and authority across the bio-pharmaceutical industry over three decades, including leadership positions at

[Eli Lilly and Company]$_{EO_1}$, Johnson & Johnson and [Novartis]$_{EO_2}$.

(17) The players in the self-driving car market are diverse: traditional car manufacturers like Nissan, Audi and [Mercedes]$_{EO_1}$, and new companies such a [Tesla]$_{EO_2}$, Google's Waymo and Uber, are all competing to develop the first fully autonomous self-driving car.

(18) Since then, the company has quickly grown thanks to its online verification technology, which is now used by fintech companies and banks, including [Monzo]$_{EO_1}$ and [Revolut]$_{EO_2}$.

In sentence (19) the two target entities EO_1 and EO_2 are providing the same goods (*Cellular handset*); thus they are in a competition.

(19) Cellular handset manufacturers such as [Nokia]$_{EO_1}$, Samsung, and [Motorola]$_{EO_2}$ rushed to introduce Internet capable smartphones following the success of Apple's iPhone.

In sentence (20) the use of the lexical pattern "EO_1 *rival of* EO_2" is a way to explicitly express a competition.

(20) Boeing et l'avionneur brésilien [Embraer]$_{EO_1}$, rival de [Bombardier]$_{EO_2}$ sur les avions régionaux, ont annoncé discuter sur un éventuel rapprochement de leurs activités.
(Boeing and the brazilian aircraft manufacturer [Embraer]$_{EO_1}$, [Bombardier]$_{EO_2}$'s regional aircraft rival, have announced discussions on a possible merger of their activities.)

- **COOPERATION:** a contractual cooperation between two *EO*s, or when two *EO*s work together on the same project. In sentence (21), EO_1 and EO_2 are partners as they have worked together to launch a new product.

(21) Mobile provider [Turkcell]$_{EO_1}$ and [Garanti Bank]$_{EO_2}$ have launched an NFC trial for mobile payments, involving a contactless MasterCard PayPass credit card application stored on an NFC-enabled SIM card provided by G&D.

In sentences (22), (23), (24), (25), (26), and (27), signing a cooperation agreement between EO_1 and EO_2 is expressed using different lexical patterns (expressions are in *italic* form).

(22) [Xiaomi]$_{EO_1}$ y [Nokia]$_{EO_2}$ firman acuerdos de cooperación comercial.
([Xiaomi]$_{EO_1}$ and [Nokia]$_{EO_2}$ sign commercial cooperation agreements.)

(23) [China Southern Airlines]$_{EO_1}$ extends [Thales]$_{EO_2}$ *partnership with* Avant IFE selection

(24) The US $263.5m *deal, led by* a consortium of [Bain Capital]$_{EO_1}$ and [Goldman Sachs]$_{EO_2}$ for Carver Korea, returned more than six times the invested capital after just a year, analysts at Korea Economic Daily found.

(25) Chrysler has partnered with Google, [Volvo]$_{EO_1}$] *will work with* [Nvidia]$_{EO_2}$ and Autoliv, GM invested in Lyft and acquired Cruise.

(26) [Airbus]$_{EO_1}$ and [The Climate Corporation]$_{EO_2}$ *join forces* to empower farmers with reliable satellite imagery.

(27) Depuis le 25 novembre 2017, 32 associations et startups, 400.000 citoyen.nes, la Fondation [Kering]$_{EO_1}$, [Facebook]$_{EO_2}$ et la Région Île-de-France *ont travaillé ensemble* avec Make.org pour élaborer le premier plan de actions de la société civile contre les violences faites aux femmes.
(Since November 25th, 2017, 32 associations and startups, 400,000 citizens, the [Kering]$_{EO_1}$ Foundation,[Facebook]$_{EO_2}$, and the Île-de-France region *have worked together with* Make.org to develop the first civil society action plan against violence against women.)

- **LEGAL PROCEEDINGS:** one EO launches a legal proceeding against another EO. In sentences (28), (29),(30), and (31), EO_1 has filed a lawsuit against EO_2. This is expressed using different verbal expressions such as *claimed*, *sued*, *filled an order against*, and *brought him against*.

(28) [Oracle]$_{EO_1}$ (ORCL, Tech30) claimed that [Google]$_{EO_2}$ violated its copyrights and patents by using the APIs in Android.

(29) [Oracle]$_{EO_1}$ bought Sun in 2010 and sued [Google]$_{EO_2}$ later that year.

(30) [J.C. Penney]$_{EO_1}$ has filed a temporary restraining order against [Sephora]$_{EO_2}$.

(31) Grégoire Triet a représenté [Shionogi]$_{EO_1}$ dans une action en contrefaçon de brevet portant sur un médicament contre le VIH, qui l'a opposé à [Merck]$_{EO_2}$ et ses filiales.
(Grégoire Triet represented [Shionogi]$_{EO_1}$ in a patent infringement action relating to an HIV drug, which brought him against [Merck]$_{EO_2}$ and its subsidiaries.)

This relations type is also expressed using direct nominal expressions (cf. sentences (32), (33)), or metaphoric expressions like in sentence (34).

(32) Defendants' fraud was alleged to be contained in affidavits and statements made during the pendency of litigation between [Lubrizol]$_{EO_1}$ and [Exxon]$_{EO_2}$ in New Jersey federal district court.

(33) A jury is currently deliberating a landmark court case between [Google]$_{EO_1}$ and [Oracle]$_{EO_2}$ over Android's use of Java APIs.

(34) The contentious court battle between Google's [Waymo]$_{EO_1}$ and [Uber]$_{EO_2}$.

- **SALE-PURCHASE:** one EO is a client of another EO, or supplies it with goods or services. As the relation is not oriented, it can be expressed in both directions using different lexical expressions. In sentences (35) and (36), EO_1 is providing a service to EO_2, whereas in sentence (37), EO_1 is buying a good from EO_2.

(35) Even more than Volusion, [Squarespace]$_{EO_1}$ offers a cheaper e-commerce solution to [Shopify]$_{EO_2}$.

(36) [Mobileye]$_{EO_1}$ is one of [Baidu]$_{EO_2}$'s (BIDU) suppliers for the Apollo autonomous driving project.

(37) When a company such as [Exxon]$_{EO_1}$ needed an additive it did not manufacture itself, it preferred to buy the additive from [Lubrizol]$_{EO_2}$ rather than a rival oil company.

The lexical expression "signing a deal" can also be used to express this relation type (cf. Example (38)).

(38) [Pratt]$_{EO_1}$ signed a deal this year with British Airways owner [IAG SA]$_{EO_2}$ to supply engines for 47 of the Airbus planes.

- **OTHERS**: If none of the previously described relations are expressed between the tagged entity pair, or if other types of relations out of this list are expressed, the relation should be OTHERS. In sentence (39), there is no business relation expressed between the two underlined EOs, Airbus and Adient, even if there are other business relation in the sentence expressed between other pairs of entities.

  (39) While [Airbus]$_{EO_1}$ partners with Audi, Boeing is cozying to [Adient]$_{EO_2}$, Mercedes- Benz, and even General Motors.

  Other examples of this relation are described below, along with reasons why it is annotated as a such.

  (40) Massive companies like [Boeing]$_{EO_1}$ and [Zimmer Biomet]$_{EO_2}$, a medical device manufacturer, are increasingly using 3-D printers to redesign products and parts to make them lighter and more efficient.
  Why? *They are not in the same industry, even if they both use 3-D printers; Using 3D printers is different from selling 3D printers.*

  (41) NCCER would like to thank the following organizations for their generous financial support and prize donations for this year's carpentry competition: Associated Builders and Contractors, The Associated General Contractors of America, Bahco, Bechtel, Build Your Future, Calculated Industries, Cianbro, ClarkDietrich Building Systems, Crossland Construction Company, Day & Zimmerman, DEWALT, Fluor, The Haskell Company, Irwin Tools, ISN, Kiewit, Klein Tools, Malco, Marek, McCarthy Holdings, Inc., Milwaukee Tools, Morton Buildings, Inc., Nabholz, NEF NAWIC Education Foundation, North American Crane Bureau, Pearson, Prov, S&B Engineers and Constructors, Smith Level Company, [Snap-on]$_{EO_1}$, Stiletto Tool Company, Sundt, and [Yates Construction]$_{EO_2}$.
  Why? *Both entities are donating to the same event, which does not necessarily mean that they are competitors or partners. Thus it is* OTHERS *relation.*

  (42) Some of the Exa's customers include BMW, [Delphi]$_{EO_1}$, Denso, Fiat Chrysler, Ford, Hino, Honda, Hyundai, Jaguar Land Rover, [Kenworth]$_{EO_2}$, Komatsu, MAN, Nissan, Peterbilt, Peugeot, Renault, Scania, Toyota, Volkswagen, and Volvo Trucks.
  Why? *If two companies are customers of the same third company, that doesn't*

necessarily mean that they are in the same industry, or linked by any other business relation defined earlier.

(43) Shira Goodman, the former CEO of Framingham office supply retailer [Staples]$_{EO_1}$, has been elected to the board of directors of Los Angeles real estate giant [CBRE Group]$_{EO_2}$.
Why? *An employee transfer from one EO to another, is not a business relation.*

(44) Ten French entities were among the world's 100 most innovative organizations in 2016: three research centers (CNRS, CEA, IFP Energies Nouvelles) and seven companies (Alstom, [Arkema]$_{EO_1}$, [Safran]$_{EO_2}$, Saint-Gobain, Thales, Total, and Valeo).
Why? *The two tagged EO are sponsoring the same event.*

3.2.2 Annotation Procedure

The collected sentences were manually annotated by non-expert native speakers via the collaborative annotation platform *Isahit*.[10] Given a sentence S, and a set of entity pairs composed of non overlapping entities $\{(EO_1,\ EO_2), EO_i \in S\}$, the annotation consists in assigning one relation R per entity pair among the five business relations: INVESTMENT, CO-OPERATION, COMPETITION, SALE-PURCHASE, LEGAL-PROCEEDINGS, and one negative relation OTHERS.

Essentially, two rules are defined for annotators to follow. First, we note that many relation types can hold between a given entity pair in the real world. In this case, ACE annotation principles should be followed Doddington et al. (2004) and annotators are asked to only consider explicit mentions of relations in the current sentence without any additional external knowledge. For example, in (45), although the underlined EO can be linked by COMPETITION (it is well known that they share the same automotive market), the final annotation is OTHERS because there are no linguistic signals for a COMPETITION relationship.

(45) Present in the city of Wuhan, PSA, Renault and even Valeo had to close their sites in the containment zone while awaiting the green light from the Chinese authorities to resume their activities.

[10] https://isahit.com/en/

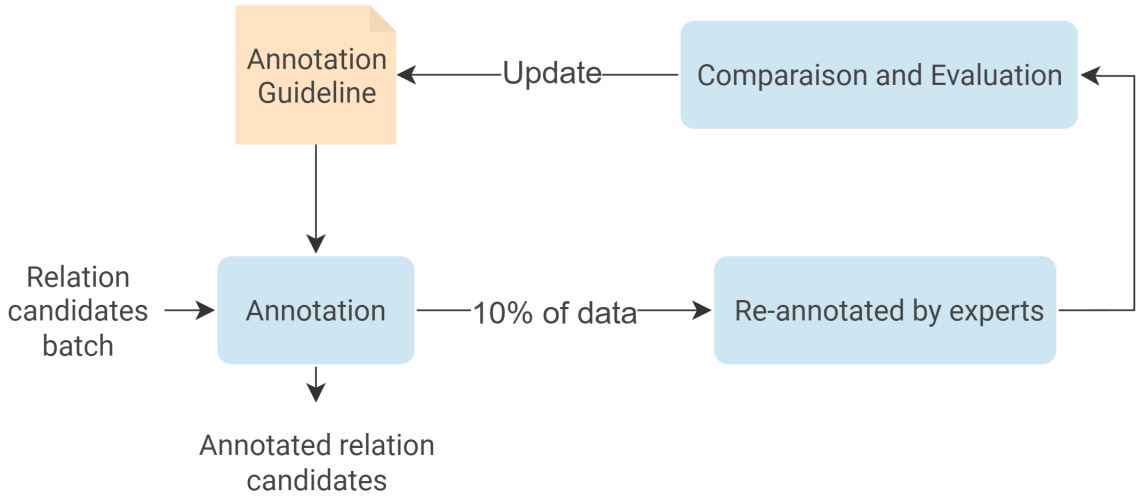

Figure 3.4 Iterative Annotation Procedure.

Second, the annotators have to make sure to choose a relation based on the semantic link between the two tagged *EO*s and not based on the general semantic meaning of the sentence, nor on other *EO*$'$*s* connections in the sentence. In Example (42), the annotated relation between the target *EO*s is OTHERS, while a relation of type SALE-PURCHASE is expressed between other pairs of *EO*s in the sentence.

We first start by annotating *English* and *French* datasets, then *Spanish* and *Chinese*. The annotation was made in batches, each containing $2k$ instances. For each batch from *English* or *French* data, 10% of the annotated data is re-annotated by experts. This helped to assess the quality of the annotations and improve annotation guidelines (cf. Figure 3.4).

Over 1k of re-annotated instances, the average Cohen Kappa (Cohen, 1960) between the annotators and the experts is 0.766 for English data and 0.685 for French data, which are strong agreements given the complexity of the task. We, therefore, use the same annotation procedure and guidelines for *Spanish* and *Chinese* data.

We considered expert re-annotations to be the most accurate. We present some reasons for annotators' inconsistencies while augmenting the experts' choice of the retained relation type below.

- Instead of focusing on the semantic relationship between the two target entities, annotators rely on the general meaning of the sentence, or/and relations between other pairs of entities (cf. Example (46)), or on their background knowledge about the target entities (cf. Example (47));

(46) The British drug-maker AstraZeneca has teamed up with Ali Health, a subsidiary of Alibaba, with the aim to grow the drug market in China and to help patients find and keep using the correct medicine with the help of smart health services and AI.
Annotation: INVESTMENT, **Validation**: COOPERATION
Explanation: The INVESTMENT is rather present between the two *EO*s *Ali Health* and *Alibaba*. However, the two target entities are *AstraZeneca* and *Alibaba*, who are in an indirect COOPERATION.

(47) Le deuxième pari gagnant est celui d'avoir misé sur les start-ups du numérique, en créant 20.000 m2 de pépinières publiques qui, ajoutées à l'arrivée massive d'incubateurs privés (Airbus, Orange, Crédit Agricole, At Home) fait de Toulouse la 2e ville de France en termes de création de start-up.
(The second winning bet is to have bet on digital start-ups, by creating 20,000 m2 of public incubators which, added to the massive arrival of private incubators (Airbus, Orange, Crédit Agricole, At Home) makes Toulouse the 2nd city in France in terms of start-up creation.)
Annotation: OTHERS, **Validation**: COMPETITION
Explanation: *Contradictory to the background knowledge* -> The two *EO* are providing the same services, which is opening start-up incubators.

- Some sentences' descriptions of the relationship between the target entities lack clarity and expressiveness, and more context or information about them is required (cf. Examples (48) and (49));

(48) Or ce sont presque toujours les mêmes acteurs qui remplissent le rôle de chef de file : EADS, BAE Systems (Royaume Uni), DRS Technologies and la Raytheon Corp.(USA), LG Electronics (Corée du Sud), Thales (France) et une flopée de sous-traitants internationaux (dont beaucoup sont israéliens).
(However, it is almost always the same players who fill the role of leader: EADS, BAE Systems (United Kingdom), DRS Technologies and the Raytheon Corp.(USA), LG Electronics (South Korea), Thales (France) and a host of international subcontractors (many of whom are Israeli))
Annotation: COMPETITION, **Validation**: OTHERS
Explanation: There is no specified market/business activity for the term "leader".

(49) Several telecommunication companies that have partnered with Ericsson to develop 5G technologies include Telstra (Australia), KT (South Korea), Turkcell (Turkey), SoftBank (Japan), SK Telecom (South Korea), LG Uplus (South Korea), NTT DOCOMO (Japan), KDDI (Japan), MTS (Russia), China Mobile (China), TeliaSonera (Sweden), Telefónica (Spain), Vodafone Group (U.K.), Singtel (Singapore), Verizon (U.S.), T-Mobile (U.S.), China Unicom (China), and Deutsche Telekom (Germany).
Annotation: COMPETITION, **Validation**: OTHERS
Explanation: It is not clearly stated that these companies belong to the same market segment. The only information is that they worked with *Ericsson* on the same technology, they can be from different market segment and contribute differently with *Ericsson*.

- Semantic and structural characteristics of the sentence can be a source of confusion for the annotator (cf. Examples (50), (51));

(50) Pour garantir l'objectivité des réponses auprès de l'ensemble des parties prenantes du groupe, Danone Way fait l'objet d'audits réalisés depuis 2002 par un organisme externe (KPMG depuis 2007).
(To guarantee the objectivity of the answers given to all the group's stakeholders, Danone Way has been audited since 2002 by an external organization (KPMG since 2007)).
Annotation: OTHERS, **Validation**: SALE-PURCHASE
Explanation: The relation that states that *KPMG* is a service provider for *Danone Way* is expressed indirectly using brackets.

(51) The banks participating in the instant cross-border payments trial are National Australia Bank, Australia and New Zealand Banking Group, Bangkok Bank, Bank of China, China Construction Bank, Commonwealth Bank, DBS, Industrial and Commercial Bank of China, Kasikornbank, Siam Commercial Bank, Standard Chartered and United Overseas Bank.
Annotation: LEGAL-PROCEEDINGS, **Validation**: OTHERS
Explanation: The semantic meaning of the word "trial" in this example is not referring to lawsuit or court cases, it rather refers to evaluation and testing procedures of the instant cross-border payments.

Table 3.3 BIZREL dataset distribution per relation type.

LANG ↓ / REL. →	Inv.	Com.	Coo.	Leg.	Sal.	Oth.	**#Total**
EN	331	1,971	738	58	292	6,644	10,034
FR	315	1,755	854	59	268	6,782	10,033
ES	62	1,067	99	81	54	1,722	3,085
ZH	86	729	329	8	26	1,075	2,316
#Total	749	5,522	2,083	206	640	15,224	25,469

3.2.3 Quantitative Results

We present here general statistics about the annotated dataset. Table 3.3 shows the total number of annotated relations per language. From the table, we can observe that our dataset is imbalanced and that OTHERS is dominant for all languages (66% for English, 68% for French, 56% for Spanish, and 46% for Chinese). The distribution of business relations is similar across languages, the most frequent ones being COMPETITION, followed by COOPERATION, then INVESTMENT. We finally observe that SALE-PURCHASE and LEGAL-PROCEEDINGS are under-represented for all languages.

To analyze the variety of relation instances in our dataset, ratio of unique sentences (RatioU. sentences), number of instances per duplicated sentences set (Nb. inst_per_dup_sent), number of unique relation types per duplicated sentences set (Nb. Utype_rel_per_dup_sent), and ratio of unique entity pairs per relation type (RatioU. entity_pairs_per_rel) are presented in Table 3.4. Indeed, a single sentence containing more than two *EOs* can be duplicated to account for the many potential relationships expressed in it between different pairs of entities during annotation. Our dataset contains 78.7% of distinct sentences in English (i.e., 21.3% of them are duplicated) (62.4%, 51.2%, and 94.3% distinct sentences in French, Spanish, and Chinese, respectively) revealing the dataset's context variety.

One sentence may be duplicated a maximum of 12 times for English data (34 for French, 16 for Spanish, and 10 for Chinese), while expressing a maximum of 3 different types of relations (max of 3 types for all languages). Furthermore, the ratio of unique entity pairs per relation type is relatively low, indicating the possibility of an over-fitting problem on target entity-pairs when learning relation types features. This is more salient for the relation LEGAL-PROCEEDING (for English, French, and Spanish), INVESTMENT (for English, French, and Spanish), COMPETITION (for Spanish and Chinese), and OTHERS (for Spanish).

Table 3.4 Statistics about BIZREL dataset relation types diversity.

STAT. ↓ / LANG. →	**EN**	**FR**	**ES**	**ZH**
RatioU. sentences (%)	78.7	62.4	51.2	94.3
Ratio. dup_sentences (%)	21.3	37.6	48.8	5.7
Nb. inst_per_dup_sent	min. = 2 avg.= 2.4 max. = 12	min. = 2 avg.= 2.6 max. = 34	min. = 2 avg.=4.7 max. = 16	min. = 2 avg.= 2.6 max. = 10
Nb. Utype_rel_per_dup_sent	min. = 1 max. = 3	min. = 1 max. = 3	min. = 1 max. = 3	min. = 1 max. = 3
RatioU. entity_pairs_per_rel (%)	Inv.= 61.3 Com.= 82.2 Coo.= 85.0 Leg.= 55.2 Sal.= 92.1 Oth.= 82.8	Inv.= 59.4 Com.=67.7 Coo.= 66.8 Leg.= 36.2 Sal. = 72.0 Oth.= 59.4	Inv.= 46.8 Com.= 47.7 Coo.= 62.6 Leg.= 65.4 Sal.= 84.0 Oth.= 51.5	Inv.= 73.3 Com.= 52.1 Coo.= 59.4 Leg.= 75.0 Sal.= 73.1 Oth.= 70.1

We further investigate the lexical features of business relations by plotting word clouds per relation type for the largest datasets, French and English (cf. Figures 3.5, 3.7, 3.9, 3.8, 3.6, and 3.10). We can see that each relation type has its set of lexical terms that refer to the semantic of the relationship. For example, the terms: *lawsuit, alleging, case, claim, sued, dispute, infringement, court, complaint, settlement* represent the business relation LEGAL-PROCEEDING in the English dataset. Semantically related terms are also used in the French dataset to express this same type of relation: *accused, lawsuit, filed, complaint, court, violation, theft, offence justice, cheated* (translation of the terms: *accuse, procès, intente, porte plainte, tribunal, violation, vol, infraction justice, triché* using DeepL).[11] On the other hand, the lexical features of the relation OTHERS are not very representative of any specific semantic relation, since it can cover many relation types under this type.

Given that the business relations are actions taken by both or one of the target EOs, we assume that the expression of these relations is closely correlated with the verbs used in the sentence. As a result, we've included the top 10 verbs per relation type and per dataset (English and French in particular) in Table 3.5. We exclude auxiliary verbs *to be* and *to have*, and the verb *to do* from the list (*être, avoir* and *faire* for French dataset).Overall, we can notice that there are two types of verbs per relation type: generic verbs such as: *include, say, drive, work* that are present in many relation types; and verbs that reflect the semantic meaning of the relation type such as *acquire, buy, own, sell, invest*, and *hold*, for

[11] https://www.deepl.com/fr/translator

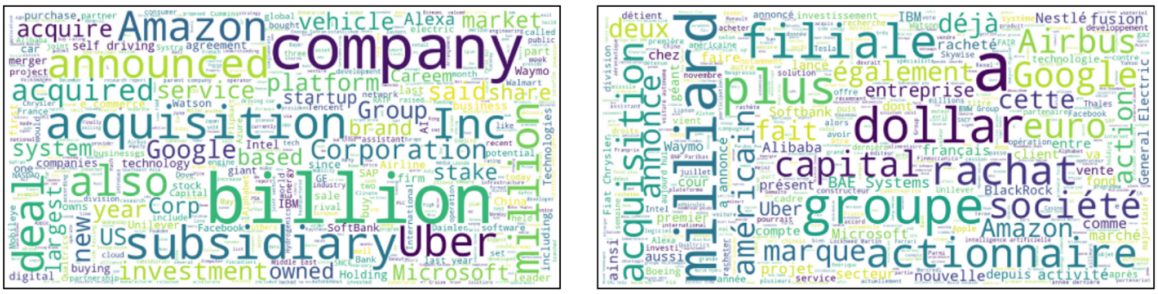

Figure 3.5 Word cloud of INVESTMENT relation for French and English BIZREL.

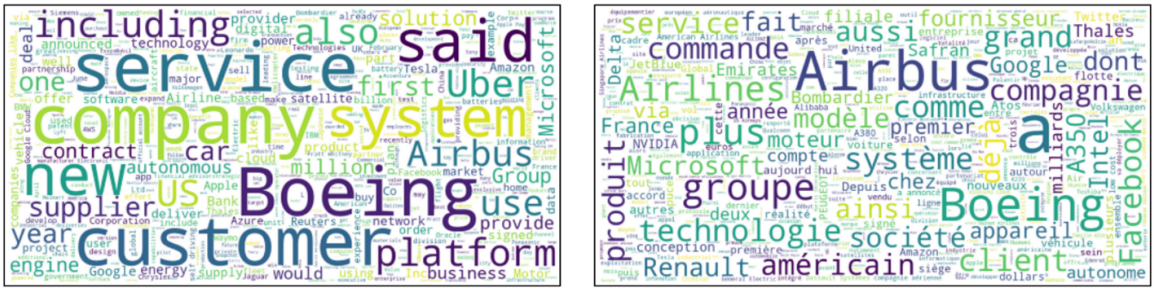

Figure 3.6 Word cloud of SALE-PURCHASE relation for French and English BIZREL.

INVESTMENT relation, or *use, provide, sell,* and *buy* for SALE-PURCHASE relation. The same pattern is noticed for the French dataset.

3.3 Pilot Experiments

We detail here the experiments we carried out on our multilingual dataset BIZREL. We first start by presenting the monolingual (Section 3.3.1) and cross-lingual experimental settings

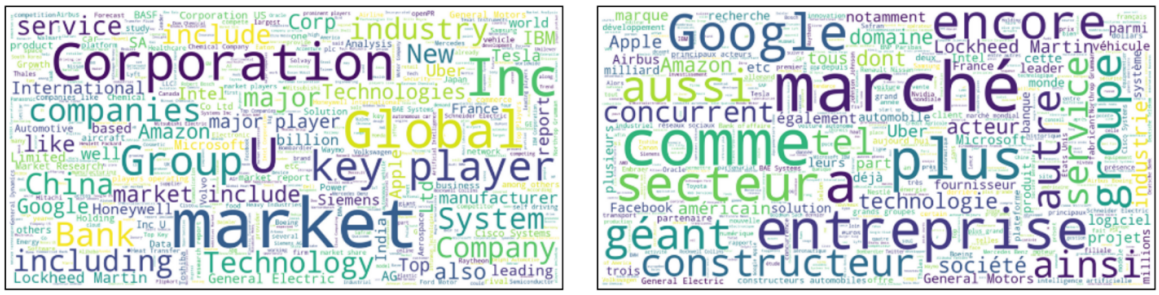

Figure 3.7 Word cloud of COMPETITION relation for French and English BIZREL.

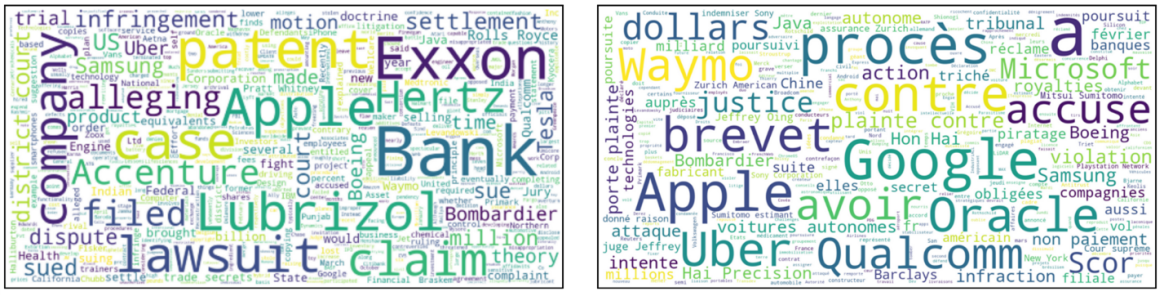

Figure 3.8 Word cloud of LEGAL-PROCEEDING relation for French and English BIZREL.

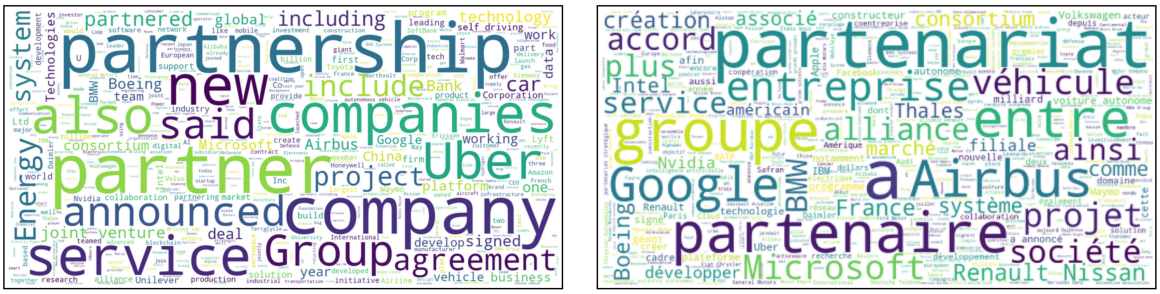

Figure 3.9 Word cloud of COOPERATION relation for French and English BIZREL.

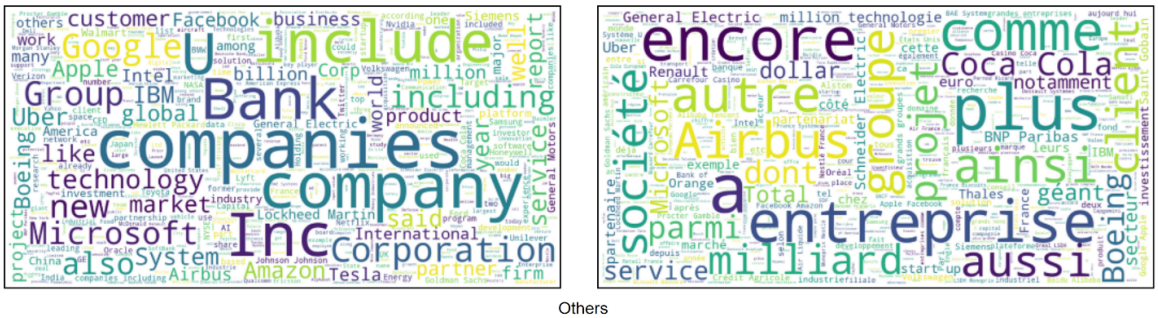

Figure 3.10 Word cloud of OTHERS relation for French and English BIZREL.

Table 3.5 Top 10 verbs per relation type in English (**EN**) and French (**FR**) BIZREL dataset.

LANG.	RELATION TYPE	TOP 10 VERBS
EN	INVESTMENT	acquire, buy, own, announce, include, drive, say, sell, hold, invest
	COMPETITION	include, lead, operate, drive, compete, provide, profile, develop, work, base
	COOPERATION	include, work, partner, develop, say, announce, sign, lead, drive, build
	LEGAL-PROCEEDING	sue, allege, fill, patent, sell, explain, complete, make, develop, drive
	SALE-PURCHASE	include, use, provide, say, deliver, buy, announce, sell, make, sign
	OTHERS	include, say, lead, work, use, make, base, take, see, announce
FR	INVESTMENT	racheter, annoncer, détenir, acquérir, pouvoir, investir, venir, assister, lancer, acheter
	COMPETITION	pouvoir, suivre, permettre, proposer, travailler, lancer, intel, devoir, twitter, développer
	COOPERATION	annoncer, associer, signer, développer, allier, créer, lancer, réunir, travailler, mettre
	LEGAL-PROCEEDING	accuse, poursuivre, estimer, porte, suprême, donner, obliger, indemniser, copier, condamner
	SALE-PURCHASE	annoncer, twitter, fournir, vendre, utiliser, développer, déployer, permettre, intégrer, aller
	OTHERS	pouvoir, annoncer, mettre, utiliser, twitter, signer, créer, aller, développer, travailler

Table 3.6 Train/test split per language for BIZREL dataset.

DATASET ↓ / LANG. →	EN	FR	ES	ZH
TRAIN	8,528	8,528	2,622	1,968
TEST	1,506	1,505	463	348

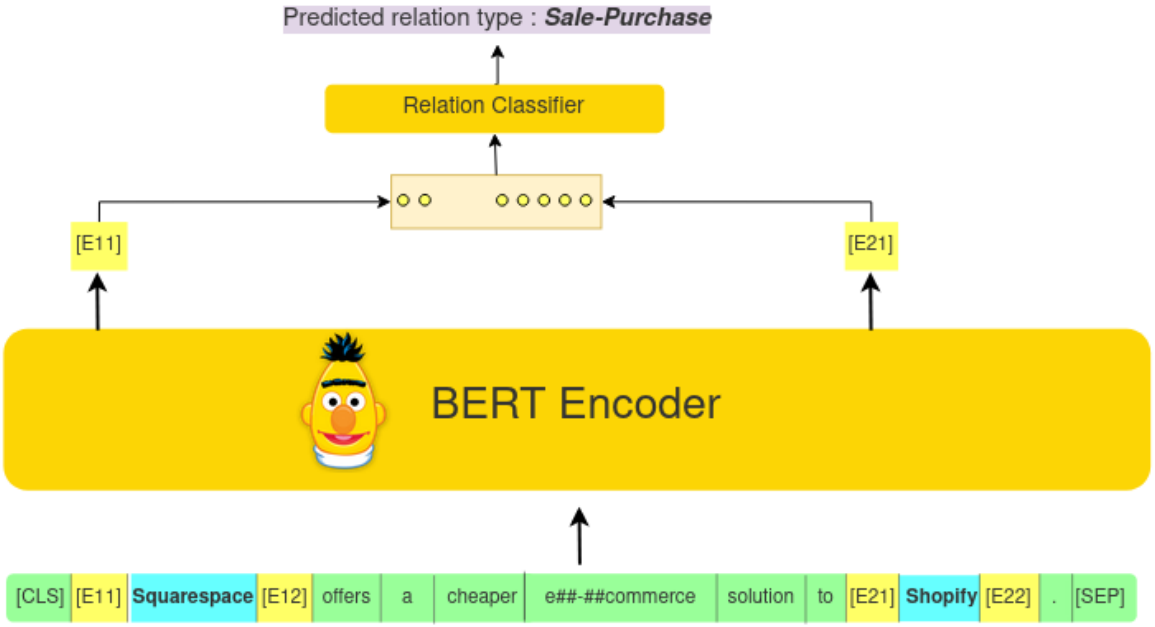

Figure 3.11 RE model as proposed by (Zhou and Chen, 2021).

(Section 3.3.2), then give our results (Section 3.3.3). We end this section with an error analysis showing main causes of misclassification (Section 3.3.4).

For all experiments, we rely on Zhou and Chen (2021)'s architecture for RE that identify entities at the input level using specific markets, and uses their representations as generated by a pre-trained language model to represent the relation instance. This architecture obtained the best scores on TACRED dataset while casting the task of RE into a multi-class classification problem. Figure 3.11 depicts the overall architecture .[12] For each language, 85% of the dataset is used to train the models, while the remaining 15% is used to evaluate the trained models' performance (cf. Table 3.6). We use a stratified split to keep the distribution of relation types unified in both train and test sets.

[12]Other models achieved better scores than (Zhou and Chen, 2021) on TACRED, however, most of them transform the RE task to non-classification tasks.

Table 3.7 Hyperparameters values in the monolingual experiments.

HYPERPARAMETER	VALUE
train_batch_size	64 (16 for monolingual FR)
test_batch_size	64
num_epochs	5
max_seq_length	400
learning_rate	5e-5
adam_epsilon	1e-6
warmup_ratio	1e-1

3.3.1 Monolingual Experiments

We rely on monolingual pre-trained language models for each language, which are pre-trained on large non-annotated data using a multi-layer bidirectional Transformer encoder (Vaswani et al., 2017) (cf. Section 1.4.5, Chapter 1) that uses a multi-head self-attention mechanism to model dependencies between tokens regardless of their distances. We use English and Chinese BERT (Devlin et al., 2019) for English and Chinese data, FlauBERT (Le et al., 2020) for French, and Beto (Cañete et al., 2020) for Spanish. All the models use 12 layers of 768 dimensions and 12 heads of attention. Each model is fine-tuned on language-specific train/test datasets using the hyperparameters in Table 3.7. We refer to this setting as (S_0) and these models are considered as strong baselines.

3.3.2 Cross-lingual Experiments

We conduct a set of experiments using the multilingual pre-trained language model mBERT,[13] a variant of BERT. mBERT is composed of 12 layers of 768 dimensions and 12 heads of attention. It is pre-trained on the concatenation of monolingual Wikipedia corpora from 104 languages. Despite being pre-trained without an explicit objective for multilingual sentence representation, mBERT can perform cross-lingual transfer on downstream tasks while fine-tuned on an annotated data of a source language with none or few annotated data in target languages (Karthikeyan et al., 2019; Wu and Dredze, 2019, 2020).Here, we fine-tune mBERT on our BIZREL multilingual dataset and consider different settings to evaluate the model ability to perform cross-lingual business relation extraction.

[13] https://github.com/google-research/bert/blob/master/multilingual.md

Let $L \in \{EN, FR, ES, ZH\}$ be the set the four languages in BIZREL. Let $T = \bigcup_i t_i$ be the dataset composed of training instances t_i from one or several source languages, $i \in L$, and let $E = \bigcup_j e_j$, the dataset composed of test instances e_j from one target language $j \in L$. We propose four experimental settings, each one involves training and testing mBERT model on different subsets of T and E, as follows:

- (S_1) **Transfer between all-language-pairs.** The model is trained on one language and tested on another, i.e., $T = \{t_i\}$, and $E = \{e_j\}$. Note that when $i = j$, this setting is similar to (S_0) but relies on multilingual contextual embeddings instead of monolingual ones. This setting aims to evaluate the cross-lingual transfer between pairs of languages.

- (S_2) **Zero-shot transfer.** Train on all languages except a given target language and test on that target, i.e., $T = \bigcup_i t_i$ and $E = \{e_j\}$ with $i \neq j$. This allows to evaluate the generalization power across-languages when training data is missing for a specific language. In addition, to measure the impact of the unseen target language during training on the overall performances of already seen languages, we further test our models on other source languages. Hence, *zero_EN*, stands for $T = \{t_{FR}, t_{ES}, t_{ZH}\}$ and $E = \{e_{EN}\}$, in addition, we evaluate the performances by testing on $E = \{e_{FR}\}$, $E = \{e_{ES}\}$, and $E = \{e_{ZH}\}$.

- (S_3) **Richly labeled transfer.** The distribution of relations across languages in BIZREL is imbalanced, with a higher frequency of French and English instances. To evaluate the impact of size on the cross-lingual experiments, we split the dataset into richly labeled (French and English) vs. poorly labeled languages (Chinese and Spanish) and either : (i) Train on $T_{richL} = \{t_{EN}, t_{FR}\}$ then evaluate on $E = \{e_j\}$ with $j \in \{ES, ZH\}$, or (ii) Train on $T_{poorL} = \{t_{ES}, t_{ZH}\}$ then evaluate on $E = \{e_k\}$ with $k \in \{FR, EN\}$.

- (S_4) **All-joint transfer.** In this last stetting, the model is trained on all the languages at the same time, and tested on one target language already seen during training, i.e., $T_{all} = \{t_{EN}, t_{FR}, t_{ES}, t_{ZH}\}$, and $E = \{e_j\}$, $j \in L$.

3.3.3 Results

Results of the monolingual and cross-lingual experiments are reported in Table 3.8, in terms of macro precision, recall, and F-score.

Overall, we can observe that models trained on multilingual data outperform their monolingual counterparts for all the languages, except for *ZH* where the Chinese BERT achieves the best with an F-score of 74.3%.

Table 3.8 Monolingual and cross-lingual models results per language. Best performing models in each (S_i) setting are in bold, while the best model for each language is underlined. ‡: Baselines models.

	Lang.	*EN*			*FR*			*ES*			*ZH*		
Sett.	**Models**	P	R	F	P	R	F	P	R	F	P	R	F
S_0 ‡	*Monolg.*	67.7	71.9	69.5	72.2	66.8	69.0	74.4	72.5	73.1	**75.8**	**73.2**	**74.3**
S_1	*EN*	66.8	**72.4**	69.1	67.8	51.9	57.3	72.2	57.3	62.3	41.6	32.4	34.8
	FR	62.6	57.5	59.6	69.0	63.4	65.8	**78.3**	67.0	70.3	39.3	30.9	31.4
	ES	54.1	57.5	54.2	58.8	51.1	53.6	**77.1**	**76.8**	**76.8**	39.0	43.3	38.4
	ZH	49.6	32.3	35.5	50.7	29.4	32.7	54.5	36.4	40.7	62.9	72.2	66.0
S_2	*zero_EN*	60.9	63.6	61.3	72.1	65.6	68.0	**83.3**	**86.3**	**84.6**	72.7	59.2	62.2
	zero_FR	66.3	70.2	67.8	68.3	55.6	60.1	**83.6**	**79.1**	**81.0**	60.0	60.6	60.3
	zero_ES	65.1	69.7	67.1	**73.1**	65.2	68.3	**79.6**	71.0	**73.9**	60.4	60.2	60.3
	zero_ZH	66.9	70.8	68.3	**74.4**	**67.0**	**69.8**	**80.5**	**77.7**	**78.7**	60.3	52.0	54.2
S_3	*richL*	66.5	**75.0**	**70.2**	71.9	**67.5**	**69.4**	**77.8**	66.5	71.1	42.5	36.9	38.4
	poorL	58.6	56.7	56.9	61.5	50.7	54.8	**75.5**	**73.8**	**73.2**	53.7	54.3	54.0
S_4	*all*	**67.8**	**72.9**	**69.9**	**74.4**	**68.8**	**70.8**	**79.3**	**80.4**	**79.7**	73.8	64.1	65.1

Compared to (S_0), transfer between all-language-pairs (i.e., the (S_1) setting) using multilingual embeddings was less productive, except for *ES* where all the scores increased (e.g., +3.7% F1). As expected, this decrease is, however, less important when the test concerns the same language. For example, -3.2% F1 when $T = \{t_{FR}\}$ and $E = \{e_{FR}\}$, while -15.4% F1 when $T = \{t_{FR}\}$ and $E = \{e_{ES}\}$. We can also conclude that language transfer from *EN*, *FR*, or *ES* to *ZH* is very poor (F1 $< 50\%$) while transfer to *ES* is feasible.

Regarding (S_2), the zero-shot transfer configuration, we note that excluding a target language from the training set was not conclusive, except for *ES*, where *zero_ES* can outperform monolingual *ES* (+0.8% F1). Similarly, excluding *ZH*, helps to boost performances of the model when evaluated on *EN*, or *FR*, while excluding *EN* or *FR* yield better results on *ES*.

Training m-BERT on *richly labeled* data boosted the results when tested on those data (see for example, +0.7% when $E = \{e_{EN}\}$ and +0.4% when $E = \{e_{FR}\}$. However, the results were lower when compared to the baselines (e.g., -2% F1 for *ES*). On the other hand, training on *poorly labeled* data has weak transfer power compared to richly labeled data.

Finally, *all-joint transfer* that combines all languages during training was the best, beating all monolingual baselines. This is more salient for *FR* where we achieve the highest F-score of 70.8%. One reason behind that could be that one relation can be expressed using similar syntactic patterns across languages, which can augment artificially relation instances for one language.

Table 3.9 Monolingual ($_m$) and best multilingual models ($_b$) F1-score per relation type and per language. Best results of each language are in bold.

	Inv.	Com.	Coo.	Leg.	Sal.	Oth.
EN$_m$	66.0	78.5	**67.2**	**77.8**	40.5	**86.7**
EN$_b$	**67.2**	**78.7**	63.9	**77.8**	**47.2**	86.3
FR$_m$	53.9	**72.4**	**68.2**	**80.0**	52.5	**87.3**
FR$_b$	**65.2**	71.7	67.7	76.9	**56.8**	86.6
ES$_m$	50.0	**86.5**	86.7	95.7	30.8	**88.7**
ES$_b$	**80.0**	84.0	**93.3**	**100**	**62.5**	88.0
ZH$_m$	**69.6**	**94.5**	**89.8**	**100**	**0.0**	**91.8**
ZH$_b$	-	-	-	-	-	-

Here again, the results when testing on the Chinese test set were not conclusive. This is probably due to the difference in script writing between *ZH* and the other languages: *EN*, *FR*, and *ES*. Thus, including these languages during training won't improve results on *ZH*. Furthermore, one possible explanation to the very good results obtained on the other languages when including *ZH* during training, may come from the named entities that are often written in English in our Chinese dataset.

A closer look into the results per class for monolingual models and best performing multilingual models per language (cf. Table 3.9)[14] shows that, in general, the relation types with the best F-score, for all languages, are the ones with more training data (COMPETITION, COOPERATION). LEGAL PROCEEDINGS has high F-scores, which can be due to the similarity and little variations of relation instance patterns because of the few examples we have. Conversely, under-represented relation types (INVESTMENT, LEGAL PROCEEDINGS, SALE-PURCHASE) gained an improvement over baseline models for many languages when training on more than one language.

3.3.4 Error Analysis

We performed a detailed error analysis on the best performing models for each language (cf. Table 3.8) in order to gain insights into the main shortcomings of the current approach. We can notice the following main sources of errors.

– One sentence-many relations. This concerns sentences containing more than one relation between different entity pairs, as in (52) and (53). In these examples, only the relation

[14] In this table, the line ZH$_b$ is empty since the monolingual model was the best.

linking the two *EO* underlined has to be identified. Our best model predicts COOPERATION (EO_2,EO_3) in (52) and (53), whereas the ground-truth annotation is INVESTMENT(EO_2,EO_3) in (52) and OTHERS(EO_2,EO_3) in (53). Note that a COOPERATION relation actually exists between EO_1 and EO_2 in (52), and between EO_1 and EO_3 in (53).

(52) [Microtronic]$_{EO_1}$ présentera ses solutions de paiement sans contact, en partenariat avec [Swisscom]$_{EO_2}$ (groupe [Vodafone]$_{EO_3}$), une solution de porte-monnaie électronique hébergée sur un téléphone portable et utilisant la technologie NFC pour communiquer.

([Microtronic]$_{EO_1}$ will present its contactless payment solutions, in partnership with [Swisscom]$_{EO_2}$ (group [Vodafone]$_{EO_3}$), a door-to-door solution electronic money hosted on a mobile phone and using NFC technology to communicate.

(53) 2017 年3 月，[Mobileye]$_{EO_1}$ 被[英特尔]$_{EO_2}$ 在2017年以153 亿美元收购，此前这家以色列视觉公司也是[特斯拉]$_{EO_3}$ 的合作伙伴，正是双方联手才有了初代的Autopilot。

(In March 2017, [Mobileye]$_{EO_1}$ was acquired by [Intel]$_{EO_2}$ in 2017 for 15.3 billion U.S. dollars. This Israeli vision company was also partner of [Tesla]$_{EO_3}$, to have the first generation of Autopilot.)

– Use of generic lexical clues. In (54), the lexical clue "*de*" (of) is generally used to express the relation type *Investment* referring to a subsidiary link between two organizations in French language. However, in this example, it does not. Our model misclassifies this sentence as INVESTMENT(EO_1, EO_2) whereas the ground-truth annotation is OTHERS (EO_1, EO_2). Moreover, the clue "*por detrás de*" (behind of) is used in (55) to express a comparison between EO_1 and EO_2 about sponsoring Fifa, whereas it can be used to express the business relation COMPETITION in other contexts. This sentence is, therefore, misclassified as COMPETITION(EO_1, EO_2) whereas the ground-truth annotation is OTHERS (EO_1, EO_2).

(54) Si [Google]$_{EO_1}$ est sorti de [Stanford]$_{EO_2}$, il y a aussi des startups françaises connues qui sont nées au sein d'incubateurs des écoles.

(If [Google]$_{EO_1}$ came out of [Stanford]$_{EO_2}$, there are also well-known French startups that were born within school incubators)

(55) [Hyundai]$_{EO_1}$ es el tercer patrocinador más antiguo de la Fifa, por detrás de Coca-Cola y [Adidas]$_{EO_2}$.

([Hyundai]$_{EO_1}$ is the third-oldest sponsor of Fifa, behind Coca-Cola and [Adidas]$_{EO_2}$.)

– Indirectly expressed relations. In (56), the expression "*has issued Autonomous Vehicle Testing Permits*" triggers a COMPETITION relation between EO_1 and EO_2. However, the model predicts OTHERS.

(56) Wheego and Valeo now join the likes of Google, Tesla, GM Cruise and Ford on the list of companies the Californian DMV has issued Autonomous Vehicle Testing Permits to, as well as [Volkswagen]$_{EO_1}$, Mercedes-Benz, [Delphi Automotive]$_{EO_2}$ and Bosch.

3.4 Conclusion

In this chapter, we presented the first multilingual corpus annotated for business relation extraction, BIZREL. It is composed of about 25,469 sentences in four languages (*French*, *Spanish*, *English*, and *Chinese*), annotated according to a novel unified characterization for *Inter-Organizational* relations composed of five important relations: INVESTMENT, COOPERATION, SALE, SUPPLY, COMPETITION, and LEGAL PROCEEDINGS. We experimented multilingual relation extraction with monolingual models, then with various cross-lingual transfer settings ranging from zero-shot to joint transfer. The best results are obtained with m-BERT trained on all-joint datasets. This work has been published as a long paper in the LREC 2022 conference (Khaldi et al., 2022c).

The following chapters investigate different learning strategies and incorporate various sources of knowledge into BERT to account for business relations specificities to improve its capabilities in extracting them.

Chapter 4

Fighting Data Imbalance for Business Relations

Business Relation Extraction between organizations is a challenging task that suffers from data imbalance due to the over-representation of negative relations (also known as *No-relation* or *Others*) compared to positive relations that corresponds to the taxonomy of relations of interest. This chapter proposes novel solutions to tackle this problem, relying on *knowledge distillation*, *multitask learning*, and *data augmentation*. When evaluated on our dataset, the results suggest that the proposed approaches improve the overall performances, beating state-of-the art solutions for data imbalance.

This chapter will begin by discussing various approaches (including data and model level approaches) to dealing with data imbalance in NLP with a special attention to those used in the context of RE in particular, (cf. Section 4.1). Then our proposed solutions in the context of business relation extraction are detailed in Sections 4.2 and 4.3. Section 4.4 reports our results on the English BIZREL. We finally end this chapter by a portability study showing how our proposed approaches can handle data imbalance in another domain specific RE, focusing on the biomedical domain (cf. Section 4.5).

4.1 Data Imbalance Solutions in NLP

In general, supervised approaches to RE consider this task as a multi-class classification problem where each class corresponds to a predefined relation type (cf. Section 1.1.3, Chapter 1). In addition to the set of *positive relations* (henceforth R^+) which corresponds to the taxonomy of relations of interest (like hypernymy, meronymy, and cause-effect relationships), most popular datasets manually annotated either for generic (e.g., SemEval-2010 Task 8

(Hendrickx et al., 2010), TACRED (Zhang et al., 2017b)) or domain-specific relations (e.g., ChemProt (Krallinger et al., 2017), BioRel (Xing et al., 2020), our dataset BIZREL) include a *negative relation* (henceforth R^-) to account either for the absence of a relation between two target entities (see NO-RELATION in TACRED), or any other types of relations not present in the annotation scheme (see OTHERS in SemEval-2010 and BizRel). NRs share two main characteristics:

- They are often over-represented, making R^+ hard to predict due to the highly imbalanced nature of the problem (see the ratio of R^- in Table 4.1).
- They have irregular and unstable linguistic realizations because they include all possible relations that are not considered in the pre-defined annotation schema (see sentences in (1) and (2) from BIZREL where different non-business relations such as *list of innovative companies*, or *employee's transfer from company A to company B* are expressed).

(1) Shira Goodman, the former CEO of Framingham office supply retailer [**Staples**]$_{EO_1}$, has been elected to the board of directors of Los Angeles real estate giant [**CBRE Group**]$_{EO_2}$.

(2) Ten French entities were among the world's 100 most innovative organizations in 2016: three research centers (CNRS, CEA, IFP Energies Nouvelles) and seven companies (Alstom, [**Arkema**]$_{EO_1}$, [**Safran**]$_{EO_2}$, Saint-Gobain, Thales, Total, and Valeo).

In addition, these patterns can be very close to the ones used to express R^+. In Example (3) taken from BIZREL, a R^- is annotated between EO_1 and EO_3 while a R^+ of type COOPERATION exists between EO_1 and EO_2. We can notice that both entity pairs follow the syntactic pattern "EO_1 *partners with* EO_2".

(3) While [**Airbus**]$_{EO_1}$ partners with [**Audi**]$_{EO_2}$, Boeing is cozying to [**Adient**]$_{EO_3}$, Mercedes-Benz, and even General Motors.

Several strategies have been proposed in the literature to account for NR: discard them during training (Doddington et al., 2004), ignore them at the evaluation stage focusing only on the performances of PR as done in most RE shared tasks (Hendrickx et al., 2010; Zhang et al., 2017b), or include them during training by treating all relations equally (Wu and He,

Table 4.1 R^- in existing generic and domain-specific datasets. ‡: We report stats. of the processed dataset by Lim and Kang (2018).

Dataset	# Sent.	#Rel.	% R^-
TACRED (Zhang et al., 2017b)	106,264	42	79.5
SemEval 2010 (Hendrickx et al., 2010)	10,717	19	17.4
BioRel (Xing et al., 2020)	533,560	125	50
ChemProt ‡ (Krallinger et al., 2017)	47,872	5	63.4
i2b2 2010 (Uzuner et al., 2011)	63,934	8	85.3
SemEval 2013 DDI (Segura-Bedmar et al., 2013)	31,927	4	84.4
BizRel (our dataset)	**10,034**	**6**	**63**

2019; Zhou and Chen, 2021). However, in a real-world scenario, these strategies fail to deal with the sparsity of R^+ and the characteristics of R^-.

On the other hand, in the NLP literature, many solutions have been proposed to deal with data imbalance at two levels: *data* and *model*. *Data level* approaches, in particular, directly address the data imbalance problem by either undersampling majority classes, or oversampling minority classes by generating new instances similar to the ones belonging to them. *Model level* solutions, on the other hand, assist the model in learning more significant features from minority classes using specific model architectures or by adapting the loss functions of the model to penalize minority classes misclassifications.

These two approaches are presented in the following sections, while highlighting their use in the context of RE.

4.1.1 Data Level Approaches

The SOTA presented here about data level approaches in NLP was partially taken from (Chiril, 2021). We however extend it and adapt it with related work on data augmentation techniques for relation extraction.

Down-sampling (undersampling) the Majority Class

The amount of data needed depends on the application and generally, the more easily distinguishable the positive class is from the negative class, the less data is needed.

Undersampling is based on the idea that the dominant class has many redundant instances, and as such, a set of majority class instances can be discarded. Many different undersampling techniques exist depending on whether the method selects:

- *which instances from the majority class should be kept* (e.g., Condensed Nearest Neighbors (Hart, 1968), Near Miss (Mani and Zhang, 2003));
- *which instances from the majority class should be deleted* (e.g., random undersampling, Edited Nearest Neighbors (Wilson, 1972)), Tomek Links (Tomek, 1976));
- *a combination of which instances from the majority class should be kept and deleted* (e.g., One-Sided Selection (Kubat et al., 1997), Neighborhood Cleaning Rule (Laurikkala, 2001)).

However, these strategies do not use all the available information (i.e., all the annotated instances), which may lead to information loss. As such, undersampling is often a solution of little interest, rarely implemented, except in scenarios with large and complex datasets, a case in which preparing/exploring the data and building pilot models is too expensive.

(Chiril (2021), p.117)

Oversampling (upsampling) the Minority Class

The drawback of undersampling could be overcome by oversampling the minority class by adding additional instances (to the minority class) and forcing the model to focus on the least represented examples.

Several approaches can be applied for obtaining new instances:

- *Random oversampling*, one of the earliest proposed methods, consists in randomly duplicating instances in the minority class. This method was shown to be an effective solution to the imbalance problem (Branco et al., 2015). However, this strategy may lead to model overfitting.
- *Collecting more data (finding a new data source).*
- Applying *data generation* techniques for generating slightly modified (or new) instances (from the already existing data) which will share the label of the original class of the instance from which they have been generated. Although common in Computer Vision, additional challenges are raised in NLP, as one needs to find semantically invariant transformations.

(Chiril (2021), p.117-118)

Data Augmentation Techniques

We provide in Table 4.2 an overview of the main existing NLP techniques. We discuss them below.

Table 4.2 Main NLP techniques for data augmentation.

DATA AUGMENTATION TECHNIQUE		METHODOLOGY
Back-translation		- translate an instance (from the source language) to another language before translating it back into the source language (Yu et al., 2018)
Lexical substitution	Thesaurus-based substitution	- replace a random word with one of its synonyms as given by a thesaurus (Mueller and Thyagarajan, 2016; Su et al., 2021a; Wei and Zou, 2019; Zhang et al., 2015b)
	Word-embeddings substitution	- replace a word with one of its nearest neighbors in the embedding space (Wang and Yang, 2015)
	Masked Language Model	- using transformer Masked Language Model predictions for replacing and inserting tokens in the previously masked portion of the text (Baek and Choi, 2022; Garg and Ramakrishnan, 2020)
	Tf/IDf based substitution	- replace uninformative words (i.e., the words having the lowest Tf/IDf scores) with other non-keywords (Xie et al., 2020)
Surface transformations (contractions and expansions)		- transform verbal forms from contraction to expansion (and vice versa) (Coulombe, 2018)
Syntax trees transformations		- the dependency tree of the original sentence is first generated, then transformed by using grammar rules (Coulombe, 2018; Xu et al., 2016)
Noise injection	Random insertion	- insert a random synonym of a non stop word in a random position in the sentence (Wei and Zou, 2019)
	Random swap	- swap the position of two random words in the sentence (Su et al., 2021a)
	Random deletion	- randomly remove each word in the sentence with a probability p (Su et al., 2021a)
	Blank noising	- randomly replace a word in the sentence with a placeholder token (Xie et al., 2017)
	Spelling error injection	- inject spelling errors to a random word in the sentence
	Sentence shuffling	- shuffle the sentences of an instance
Mixup	wordMixup/senMixup	- generate new instances by linearly interpolating word/sentence embeddings (Guo et al., 2019a)
Generative methods	PLMs	- finetune a PLM and generate new instances by using the class label and a few initial words as cue for the model (Anaby-Tavor et al., 2020; Papanikolaou and Pierleoni, 2020)

> **Back-translation** (cf. Figure 4.1) is a technique based on paraphrasing that relies on translating an instance (from the source language) to another language before translating it back into the source language (Yu et al., 2018). The major advantage of employing this method is that the overall semantics of the sentence are maintained, while bringing more syntactical diversity to the newly generated data.
>
> (Chiril, 2021, p.118)

Yu et al. (2020) employed multiple neural machine translation systems to generate possible paraphrases for each sentence in an existing RE dataset via back-translation. However,

because word alignment information is unavailable, target entity tokens cannot be restored after back-translation. To solve this issue, a contextual similarity based method was proposed to align entities of paraphrased instances with the ones in the original sentence.

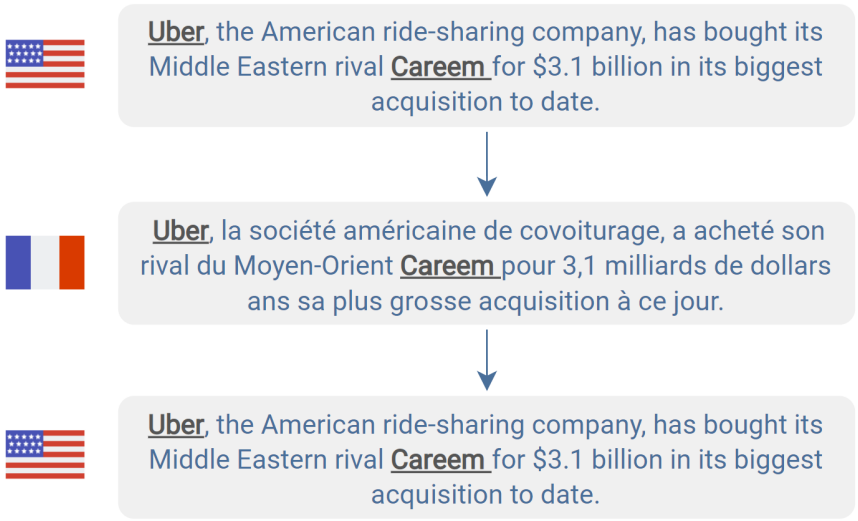

Figure 4.1 Data augmentation through back translation.

> There are also some techniques relying on replacing some words in the text while preserving its meaning (**lexical substitution**) for generating additional data:
>
> - replacing random words with one of their synonyms as given by a thesaurus (e.g., WordNet (Miller, 1995), BabelNet (Navigli and Ponzetto, 2012), ConceptNet (Speer et al., 2017)) (Mueller and Thyagarajan, 2016; Wei and Zou, 2019; Zhang et al., 2015b).
> - leveraging pre-trained word embeddings for selecting the nearest neighbors in the embedding space as replacement for some words in the text (Wang and Yang, 2015).
> - leveraging BERT (or other transformer models) Masked Language Model (MLM) predictions for replacing and inserting tokens in the previously masked portion of the text (Garg and Ramakrishnan, 2020).
>
> (Chiril, 2021, p.118-119)

To preserve target entities in a relation instance while using MLM to generate new samples, Baek and Choi (2022) repeatedly masked all tokens except target entities, and replaced them

randomly with one of the top-k most likely words candidates, to prevent the generation of common phrases and repetitive texts (Fan et al., 2018) (cf. Figure 4.2).

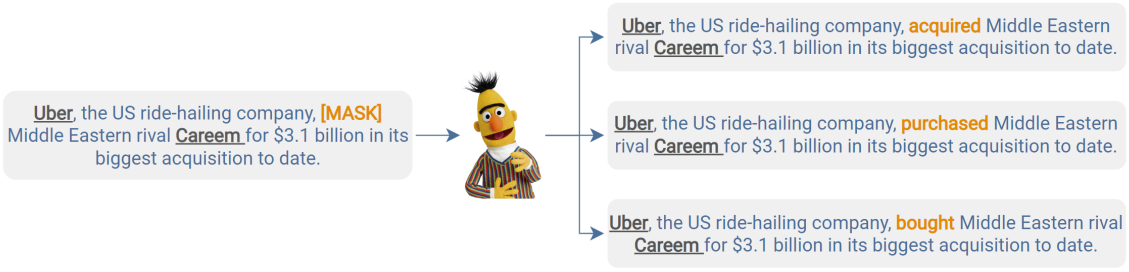

Figure 4.2 Data augmentation through BERT masked language model.

Another augmentation technique relies on **surface transformations**, semantically invariant transformations that are language-dependent and which rely on contractions and expansions (cf. Figure 4.3). To preserve the semantic invariance, Coulombe (2018) proposes to allow ambiguous contractions but avoid ambiguous expansions that can lead to misinterpretations.

Figure 4.3 Example of contraction and expansion. The word concerned is underlined.

(Chiril, 2021, p.121)

Coulombe (2018) proposes a second strategy using **syntax trees transformations**, where the dependency tree of the original sentence is first generated, then transformed by using grammar rules. Finally, the transformed dependency tree is used to generate a paraphrased sentence (e.g., the transformation from active voice to the passive voice of the sentence (and vice versa) is a semantically invariant transformation). Furthermore, Xu et al. (2016) reversed the direction of the SDP between target entities to generate new instances while ignoring the context out of the SDP (cf. Figure 4.4).

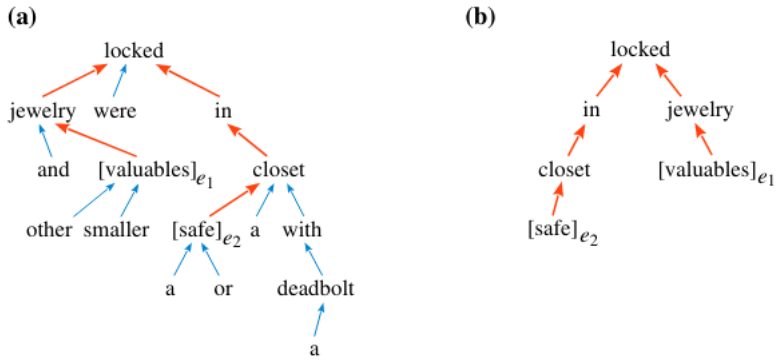

Figure 4.4 (a) The dependency parse tree corresponding to the sentence "Jewelry and other smaller [valuables]$_{e_1}$ were locked in a [safe]$_{e_2}$ or a closet with a deadbolt." Red arrows indicate the shortest dependency path between e_1 and e_2. (b) The augmented data sample.

Generating new instances through **noise injection** relies on duplicating instances and injecting noise into them. The added parasitic noise will not change the semantic of the new instance, but rather introduce several variations of the same sample, which will allow the model to better generalize when encountering instances having this kind of perturbations. Several noise injection techniques were proposed:

- *random insertion* relies on finding a random synonym for a random non-stop word in the sentence and inserting it in a random position (Wei and Zou, 2019);
- *random swap* relies on randomly choosing two words in the sentence and swapping their position (Wei and Zou, 2019);
- *random deletion* relies on randomly removing each word in the sentence with a probability p (Wei and Zou, 2019);
- *blank noising* is similar to the *random deletion* technique, but rather than deleting a word, it will replace it with a placeholder token (e.g., '__') (Xie et al., 2017);
- two other techniques (not referenced in literature) rely on *injecting spelling errors* (either to some random words in the sentence or by simulating typing errors, i.e., replacing some letters in a word by letters found close by on a keyboard) and *shuffling the sentences of an instance*.

(Chiril, 2021, p.121-122)

Some works for RE combined multiple noise injection strategies such as random delete and random swapping to generate new RE instances for low-resource languages (Moein et al., 2021). Akkasi and Moens (2021) randomly over-sampled instances from minority relations using a replacement strategy to generate a causality-extraction dataset with equal number of instances for different relation types. Smirnova et al. (2019), on the other hand, used instance duplication of minority relation types with a ratio larger than 1/5 between the least frequent relation and the most frequent one. Finally, Su et al. (2021a) considered that the SDP between the two target named entities captures the required knowledge for the relation expression. Therefore, the words on SDP are fixed during the data augmentation to prevent information loss while the other words are either deleted, swapped or replaced by their WordNet synonym (cf. Figure 4.5).

Original	We further show that @PROTEIN$ directly interacts with @PROTEIN$ and Rpn4.
After SR	We further show that @PROTEIN$ **straight** interacts with @PROTEIN$ and Rpn4.
After RS	**Further we** show that @PROTEIN$ directly interacts with @PROTEIN$ and Rpn4.
After RD	We further show that @PROTEIN$ interacts with @PROTEIN$ and Rpn4.

Figure 4.5 Data augmentation while fixing the SDP between target entities. The SDP between the two proteins is "@PROTEIN$ interacts @PROTEIN$" (underlined in the examples). The changed words are also marked with bold font.

Initially introduced by Zhang et al. (2017a), **Mixup** is an image augmentation technique where new instances are generated by linearly interpolating pixels of random image pairs. Contrary to other data augmentation techniques, the images can belong to different classes. Guo et al. (2019a) adapted this technique for NLP tasks and propose two strategies of Mixup on sentence classification:

- in *wordMixup* (cf. Figure 4.6) the interpolation is performed on word embeddings (i.e., the two instances are zero-padded to the same length and their word embeddings are interpolated);

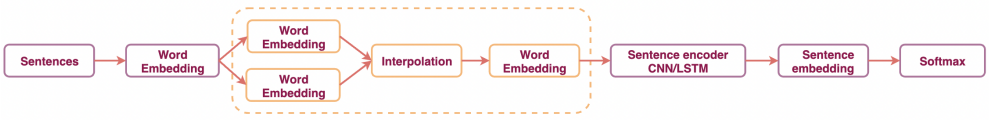

Figure 4.6 wordMixup technique (Guo et al., 2019a) (the added part to the standard sentence classification model is in the orange rectangle).

- in *senMixup* (cf. Figure 4.7) the interpolation is performed on sentence embeddings (i.e., the hidden embeddings for the two instances are generated by an encoder (e.g., CNN, LSTM) before being linearly interpolated).

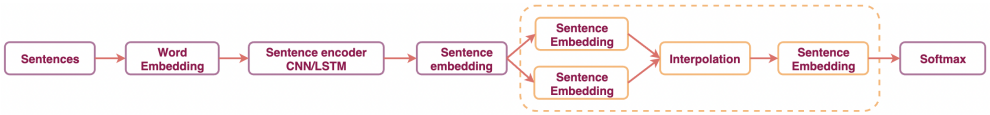

Figure 4.7 senMixup techniguo2019augmentingque (Guo et al., 2019a) (the added part to the standard sentence classification model is in the orange rectangle).

(Chiril, 2021, p.123)

Pre-trained **generative models** have also been to augment training data. To generate new instances, Papanikolaou and Pierleoni (2020) fine-tuned separate pre-trained GPT-2 models per relation type. Target entities were initially masked in the fine-tuning data with special masks to highlight their presence. Only generated instances with two special entity masks are kept in the augmented dataset (cf. Figure 4.8).

Dataset(relation type)	Generated sentences
CDR(Induce)	DISEASE was the most common adverse reaction (21 %) reported for DRUG, and occurred in approximately 50 % of patients .
DDI2013(Effect)	DRUGA may enhance the effects of alcohol, barbiturates, DRUGB, and other cns depressants.
DDI2013(Advise)	caution should be observed when DRUGA and DRUGB are coadministered.
DDI2013(Mechanism)	co-administration of DRUGA decreased the oral bioavailability (48%) of DRUGB, a substrate for cyp2d6.
ChemProt(Activate)	DRUG enhances PROTEIN sensitivity via activation of the pi3k / akt signaling pathway.
ChemProt(Inhibit)	DRUG, a novel orally bioavailable xanthine PROTEIN inhibitor,
ChemProt(Product)	the enzyme PROTEIN catalyzes the two-electron reduction of DRUG to produce acetyl groups.

Figure 4.8 Examples of generated sentences with fine-tuned GPT-2 models. Each model is fine-tuned on examples from the specific relation type (Papanikolaou and Pierleoni, 2020).

Eyal et al. (2021), on the other hand, generated a small set of lexically and structurally diverse instances per relation type then used them to collect new syntactically similar instances occurring in knowledge bases using on a syntactic search over syntactic-graphs search system (Shlain et al., 2020).

4.1.2 Model Level Approaches

We group model level approaches into two categories: those that *adapt the model's architecture*, and those that *adapt the model's loss function*. Table 4.3 summarizes the reviewed approaches to handle data imbalance at the model level.

Table 4.3 Model level approaches for data imbalance.

Model Level Approaches		Methodology
Model's architecture	Multi-task	Including RE related task as auxiliary tasks to be optimized along with the main task of RE (Lyu et al., 2020; Ye et al., 2019).
	Knowledge distillation	Train a teacher model on sentences augmented with ground-truth labels then use this model to generate soft-labels that supervise a student model (Song et al., 2021).
Model's loss function	Weighted cross-entropy	Weights are assigned to classes as costs of mislabeling that class during training.
	Ranking loss	Maximize the score assigned to the correct class and minimize the one of incorrect classes (Dos Santos et al., 2015).
	Focal loss	Down-weights the loss assigned to well-classified examples and focus on hard example (Lin et al., 2017).
	Adaptive scaling	Dynamically scales costs of instances of different classes during the training (Lin et al., 2018).
	Dice loss	Optimizes the number of correctly predicted positive instances compares to the total number of positive instances (Li et al., 2020b).

Adapting Model's Architecture

Model's architecture is adapted to transform the data imbalance problem into many balanced sub-problems, or to include new components that can learn more knowledge about the minority classes. We focus here on the methods proposed for RE.

Multitask learning. It is an effective method for improving the performance of a single task by incorporating other related tasks, making it easy to combine information from multiple resources. These architectures have shown to be effective for generic and specific relation extraction by learning additional implicit information from either generic auxiliary tasks (Wang and Hu, 2020; Zhou et al., 2019) (e.g., dependency parsing, recognizing textual entailment task from GLUE Benchmark), or from multiple domain related tasks (Yadav et al., 2020) (e.g., Protein-Protein Interaction, Drug- Drug interaction). Recently, a new auxiliary task of relation identification that consists in identifying R^+ from R^- is performed along with the main RE task to account for the semantic information of the R^- resulting in the improvement of generalization performance (Lyu et al., 2020; Ye et al., 2019). Furthermore, Yu et al. (2020) suggested using an augmented dataset and its original version to jointly train a RE model using two optimization objectives to handle data imbalance. Finally, Smirnova

et al. (2019) designed a dual-learning architecture where both tasks of relation identification and relation classification are learned, then their relation probabilities are merged together in a final RE classification layer.

Knowledge distillation. The main idea behind Knowledge distillation (henceforth KD) is to design a simple student model that mimics the behavior of a complex, more informed, or a large teacher model to achieve comparable results in performing a specific task. It has first been proposed for model compression task (Hinton et al., 2015).

KD has been recently proposed for RE. Zhang et al. (2020b) incorporates knowledge about type constraints between entities and R^+ into the teacher model, then use knowledge distillation to generate well-informed soft labels used to supervise a student model that can inherit this knowledge from its teacher. Song et al. (2021) integrated ground truth sentence-level identification information into the teacher network during training, then transfer it to the student by sharing the classification layer to counter data imbalance problem. KD has also been used to alleviate the interference of noise from relation annotations in distant supervision via label softening (Li et al., 2022) or by leveraging a small set of manually annotated data to generate soft-labels that supervise training on distantly supervised data (Tan et al., 2022a).

Adapting Loss Functions

In classification tasks, a loss function is a measure computed on the outputs of a given model to assess the quality of its predictions in comparison to the true labels while minimizing the model's errors. The majority of classifiers assume that the costs of misclassification are the same for all classes. This assumption is however false in the vast majority of real-world applications.

Considering a binary classification where *a cancer* is regarded as *positive* and *non-cancer* (healthy) as *negative* in medical diagnosis, then missing a cancer (the patient is actually *positive* but is classified as *negative*; thus it is also known as "false-negative") is much more serious and expensive than false-positive error (Thai-Nghe et al., 2010). Multiplying the loss of each example from a minority class by a certain factor referring to the cost of its misclassification is one proposed solution for this problem. In the following, we go over different loss functions that have been proposed in this context.

Generally, **cross-entropy loss** (CE) is the most commonly used loss function for NLP classification tasks. This loss aims to maximize the neural model's accuracy across all training instances. It is calculated following Equation 4.1, with y_i is the true label, and p_i

is the model's prediction. C denotes the number of classes and L the number of training instances.

$$\mathscr{L}_{CE} = -\sum_{l \in L} \sum_{i \in C} y_i^{(l)} log(p_i^{(l)}) \tag{4.1}$$

If the dataset is imbalanced, however, this loss is more likely to correctly classify instances of majority classes because doing so improves overall accuracy, which penalizes minority class learning. For example, to extract an uncommon class with a 1% occurrence in the dataset, a trivial classifier that never predicts that class is 99% accurate.

Weighted-cross-entropy (WCE) is a variant of CE that allows to assign weights to classes (cf. w_i in Equation 4.2), which can increase or decrease the cost of mislabeling a certain class during training. These weights are generally set to the proportion of the inverse of the number of instances per class.

$$\mathscr{L}_{WCE} = -\sum_{l \in L} w_i \sum_{i \in C} y_i^{(l)} log(p_i^{(l)}) \tag{4.2}$$

Focal loss (Lin et al., 2017) is another interesting loss function to handle class imbalance in training data. It was first used for dense object detection where the imbalance between background and foreground classes is extreme, then later applied to NLP problems (Huang et al., 2021a; Liu et al., 2021a; Tan et al., 2022a). It is a reshape of the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples and focus on hard example, which prevents the vast number of easy negatives from overwhelming the model during training. Practically, a modulating factor $(1-q_i)^{\gamma}$ is added to the cross entropy loss, with a tunable focusing parameter $\gamma \geq 0$ (cf. Equation 4.3).

$$\mathscr{L}_{FC} = -\sum_{l \in L} \sum_{i \in C} (1-p_i)^{\gamma} y_i^{(l)} log(p_i^{(l)}) \tag{4.3}$$

The focusing parameter γ is used to control the rate at which easy examples are down-weighted. Especially, when $\gamma = 0$, $\mathscr{L}_{FC}$ is equivalent to $\mathscr{L}_{CE}$. The effect of the modulating factor $(1-y_i)^{\gamma}$ is increased with γ. Then, when an example is misclassified and y_i is small, the modulating factor nearly tends to 1, and thus the loss is unaffected. As y_i increases to 1, the modulating factor nearly tends to 0 and the loss for well-classified examples is down-weighted.

Ranking loss has also been used to handle R^- characteristics in RE task. In (Dos Santos et al., 2015), instead of using a softmax function to generate class probabilities, a dot product

between the relation representations and embeddings of classes is used to calculate classes scores. The R^- is ignored at this stage, where no embedding vector is assigned to it. The loss is calculated following Equation 4.4.

$$\mathscr{L}_{RN} = -\sum_{l \in L} log(1 + exp(\gamma(m^+ - s_{c^+}^{(l)}))) + log(1 + exp(\gamma(m^- + s_{c^-}^{(l)}))) \tag{4.4}$$

Where s_{c^+} and s_{c^-} are respectively the scores for class labels c^+ (being the correct label), and c^- (being the wrong label), generated by the RE model. m^+ and m^- are margins and γ is a scaling factor that magnifies the difference between the score and the margin and helps to penalize more on the prediction errors. The first term on the right side of the equation decreases as the score s_{c^+} increases. The second term on the right side decreases as the score s_{c^-} decreases. Training a RE model consisted in giving scores greater than m^+ for the correct class and (negative) scores smaller than $-m^-$ for incorrect classes. c^- is set to the incorrect class with the highest score (cf. Equation 4.5).

$$c^- = argmax_{c \in C; c \neq c^+} s_c \tag{4.5}$$

Instead of using loss functions that optimize the accuracy of the model (i.e., cross-entropy), Lin et al. (2018) proposed **adaptive scaling**, an algorithm borrowed from economics based on the idea of marginal utility (Stigler, 1950) to directly optimize F1-measure and handle R^+ vs. R^- imbalance without introducing any additional hyperparameters. They claimed that because of the R^+ sparsity problem, the marginal utility of predicting one more positive instance differs from the marginal utility of predicting one more negative instance during the training of a RE model and can change dynamically. They proposed a dynamic cost-sensitive learning algorithm that adaptively scales costs of instances of different classes during the training procedure, allowing the loss function to optimize to be in accordance with the F1-measure.

Another F1-oriented loss function is the **dice loss** that has recently been used for data-imbalanced in NLP binary classification tasks (Li et al., 2020b). It is based on Sørensen–Dice coefficient (Dice, 1945; Sørensen, 1948) that is used to compare the similarity between two sets (cf. Equation 4.6). When trained the classifier, the two sets refer to positive instances predicted by the model, and the set of golden positive instances.

$$\mathscr{L}_{DC_{bin}} = 1 - \frac{2\sum_{i \in L} p_1^{(l)} y_1^{(l)} + \gamma}{\sum_{i \in L} p_1^{(l)^2} + \sum_{i \in L} y_1^{(l)^2} + \gamma} \tag{4.6}$$

This loss gives the same importance to false-positive and false-negative instance. Therefore, following the idea of focal loss, Li et al. (2020b) suggested adding a modulating factor to the dice loss to dynamically weight instances predictions and thus improve hard-examples learning (cf. Equation 4.7).

$$\mathscr{L}_{DC_{bin}} = 1 - \frac{2\sum_{i\in L}(1-p_1^{(l)})^{\alpha} p_1^{(l)} y_1^{(l)} + \gamma}{\sum_{i\in L} p_1^{(l)^2} + \sum_{i\in L} y_1^{(l)^2} + \gamma} \tag{4.7}$$

The multi-class dice loss is given by the following equation (Milletari et al., 2016).

$$\mathscr{L}_{DC_{multi}} = \frac{1}{N}\left[1 - 2\frac{\sum_{l\in L}\sum_{i\in C} y_i^{(l)} p_i^{(l)} + \gamma}{\sum_{l\in L}\sum_{i\in C}(y_i^{(l)} + p_i^{(l)}) + \gamma}\right] \tag{4.8}$$

4.2 Handling Business Relations Data Imbalance

We aim to answer this main research question:

– *How can we leverage on existing data-imbalance solutions to improve business relations extraction, without relying on additional manually annotated data?*

To this end, we propose the three following approaches:

– MT-RE. A multitask learning approach for improving RE that considers relation identification and classification as auxiliary tasks. Relation identification as an auxiliary task has been shown to improve RE performance (Lyu et al., 2020; Smirnova et al., 2019; Ye et al., 2019). We also consider the task of relation classification on a relatively balanced dataset where R^- is excluded from the training data. Our contribution consists in combining various auxiliary tasks from the same dataset to assist the model in learning more discriminative features between business vs. non-business instances, as well as between business instances only.

– SADA-RE. A Semantically-Aware Data Augmentation approach for RE based on similarity between relation instances representations. Despite the variety of data augmentation techniques (cf. Section 4.1.1), the new instances obtained through these methods may contain the same or similar words as the original instance but in a different order, which may result in generating instances that do not make sense to

humans. In addition, these methods do not guarantee that the new generated instances belong to the same class as the original ones. To avoid this, we propose a new approach for data augmentation based on sentence similarity. Following Chiril et al. (2021) who used a similar method using SentenceBERT to augment a gender stereotype detection dataset, we designed, for the first time as far as we know, the first similarity-based data augmentation approach for RE.

– BSLS-RE. A new knowledge distillation approach to account for R^- characteristics in imbalanced RE problem based on *Binary Soft Labels Supervision* . State-of-the-art results show that using soft-labels to supervise RE allows transferring more specific knowledge about relation types from a teacher model to a student model. We continue this effort here to inherit salient features distinguishing R^+ from R^- when training a RE model. Our work is close to (Song et al., 2021) but instead of adding more features to the teacher model, we rather train the teacher and student models on two different complementary tasks: binary relation identification (R^+ vs. R^-) and multi-class relation extraction. We assume that training a teacher model on binary relation identification helps to learn discriminative features that differentiate R^+ from R^-, on a less imbalanced dataset, since all R^+ are merged into one class. The student model can therefore inherit from the teacher's produced binary soft labels the salient learned features about R^+ and R^-, to mitigate R^- irregular patterns' problem.

Moreover, loss function selection is not a common practice, despite its importance, where suboptimal loss functions are frequently chosen, impacting the performance of the trained model. As a result, it is important to compare available loss functions and select the most appropriate and optimal ones for a given task. To this end, each of the proposed approaches above are optimized using state-of-the-art loss functions that handle data imbalance. We experiment in particular with the focal loss (Lin et al., 2017), the dice loss (Li et al., 2020b), the adaptive scaling (Lin et al., 2018), and the weighted cross-entropy.

We detail our contributions in the next section, then report their results when evaluated on the English portion of BizRel in Section 4.4.

4.2.1 Multitask Business Relation Extraction (MT-RE)

We aim to answer one main research question:

— What additional training objectives could help learning more specific features about R^+ and R^- ?

To this end, we propose MT-RE, a multitask learning approach shown in Figure 4.9.

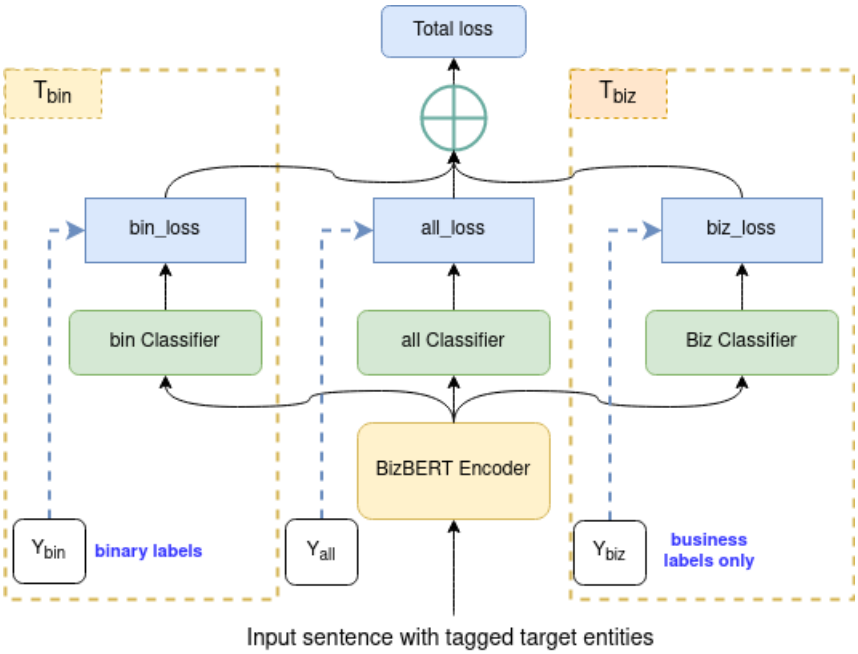

Figure 4.9 Multitask learning approach for BRE.

Our objective is to assign to a relation instance noted $i = (S, e_1, e_2)$, where S is the sentence, e_1 and e_2 are target entities, one relation type r from a set of predefined relations R. We consider three variants of R:

– bin = { *business*, *Others*}.

– biz = { *Invest., Compet., Cooperat., Legal., Sale.*}.

– all = { *Invest., Compet., Cooperat., Legal., Sale., Others*}.

Let T_R be a relation classification task, where R is the set of pre-defined relations to consider in this task. Our main task is T_{all} performing business relation classification while accounting for the negative relation OTHERS given its importance in end-user systems. We consider two different auxiliary tasks to be learned jointly with the main task:

- T_{bin} a binary relation classification task (business vs. non-business) that helps the main task to learn more generic features about business relations and discriminates them from non-business ones (OTHERS),
- T_{biz} a multi-class business relation classification task that learns more specific features about business relations while discarding the noisy negative relation OTHERS which has irregular patterns.

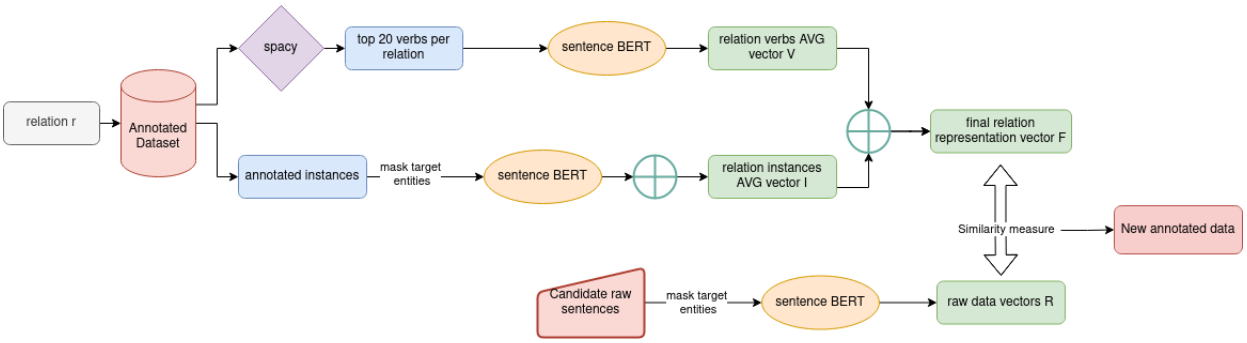

Figure 4.10 Semantically aware Data Augmentation for Relation Extraction (SADA-RE)

All these three tasks have a shared sentence encoder, and each one has its classifier accounting for the number of relations to consider: 2 for T_{bin}, 5 for T_{biz}, and 6 for T_{all}.

We experiment with three configurations aiming to improve the main classification task by considering three different combinations of auxiliary tasks: MT-RE$_{all+bin}$, MT-RE$_{all+biz}$ and MT-RE$_{all+biz+bin}$. The models are trained using the cross-entropy loss, where one loss is calculated per task in each configuration. The total loss equation is presented in Equation 4.9.

$$\mathscr{L}_T = \alpha.\mathscr{L}_{all} + \beta.\mathscr{L}_{bin} + \gamma.\mathscr{L}_{biz} \tag{4.9}$$

Where $\mathscr{L}_i$ is the cross-entropy loss associated to the task i, α, β , and γ are loss weights, set to zero when the task associated to the loss is not considered.

4.2.2 Semantically-Aware Data Augmentation for BRE (SADA-RE)

We aim to answer one main research question:

— Is sentence similarity an effective data augmentation strategy for BRE?

To this end, we propose SADA-RE, a new Semantically-Aware Data Augmentation approach for BRE (cf. Figure 4.10) based on the similarity between sentences and relations representation to augment the positive instances. We test our approach on the English portion of our BIZREL dataset.

We use SentenceBERT, a modification of BERT that generates semantic sentence embeddings that can be compared using cosine-similarity (Reimers and Gurevych, 2019), to extend positive instances of our dataset from a non-annotated textual data collected from the web by requesting search engines API using domain activity keywords such as *tourism and*

Covid-19, *aerospace technologies*, etc. The collected documents are split into sentences and two named entities of type organization are identified per sentence.[1] We got a total of 6,800 sentences.

The data augmentation is performed as follows:

- We first extract the top-20 most frequent verbs per positive relation from the original dataset. We assume that verbs can be a good semantic descriptor of a semantic relation linking two named-entities (cf. Section 3.2.3, Chapter 3).
- We therefore use the generated lists of verbs per relation to generate semantically aware representation vectors for each relation type. Verbs per relation are concatenated to form a sentence that is fed into SentenceBERT to generate the verb-based relation vector.
- For each relation, its verb-based representation is then combined with the averaged representations of instances per relation type, to produce the final semantically aware relation representation.
- These relation representations are then used to retrieve the most similar instances from the unlabeled data.

A similarity threshold is fixed per relation type to only select the most semantically close sentences and reduce noise in the retrieved instances. To avoid that one sentence is returned for more than one relation type, the conflicting instances are assigned to the relation with the highest similarity measure. Note that the two targeted entities in both annotated and non-annotated dataset are masked and marked using special markers to highlight their positions in the sentence.

This approach results in an augmentation of 34% of the initial dataset (only the positive relations have been augmented). Table 4.4 gives some examples of augmented instances per relation type, along with the similarity score (Simi.) to the relation vector representation. We can notice that long contexts are used to express the relation types in the augmented dataset.

The generated data is combined with the original training data and used to train a RE model. Table 4.5 summarizes the distribution of R^+ and R^- in the original dataset BIZREL^O and the augmented dataset BIZREL^{Aug}.

[1] We follow the same procedure used to collect sentences to annotated for BIZREL dataset. For more details, see Section 3.1, Chapter 3

Table 4.4 Examples of instances per relation type from the of data augmentation.

REL.	EXAMPLE	SIMI.
Inv.	RCA, with its own RCA Astro Electronics satellite construction business, identified a role for itself as a satellite owner/operator.	0.55
Com.	Some of the companies competing in the Collaborative Robots Market are ABB, Robert Bosch, KuKa Ag, Aubo Robotics, Fanuc, Rethink Robotics, Precise Automation, Inc., Universal Robots, Yasakawa Electric Corporation, TECHMAN Robots and Kawasaki Heavy Industries, Ltd., among others.	0.56
Coo.	To further streamline the process, Bayer in 2012 established a partnership with EcoVadis, a leading supplier of collaboration platforms with which companies can assess the sustainability performance of their suppliers.	0.57
Leg.	The Mexican arm of drinks company Anheuser-Busch InBev accused U.S. firm Constellation Brands in a lawsuit filed on Monday of breaching a deal on the use of the Corona brand name by applying it to a product other than beer.	0.54
Sal.	Vietjet agreed in February to buy an additional 100 Boeing 737 Max airplanes, on top of the 100 it had already ordered.	0.51

Table 4.5 Results of data augmentation on R^+ and R^-.

DATASET →	BIZREL^{O}		BIZREL^{Aug}	
RELATIONS →	R^+	R^-	R^+	R^-
Nb. Inst.	3,390	6,644	4,543 (+34%)	6,644

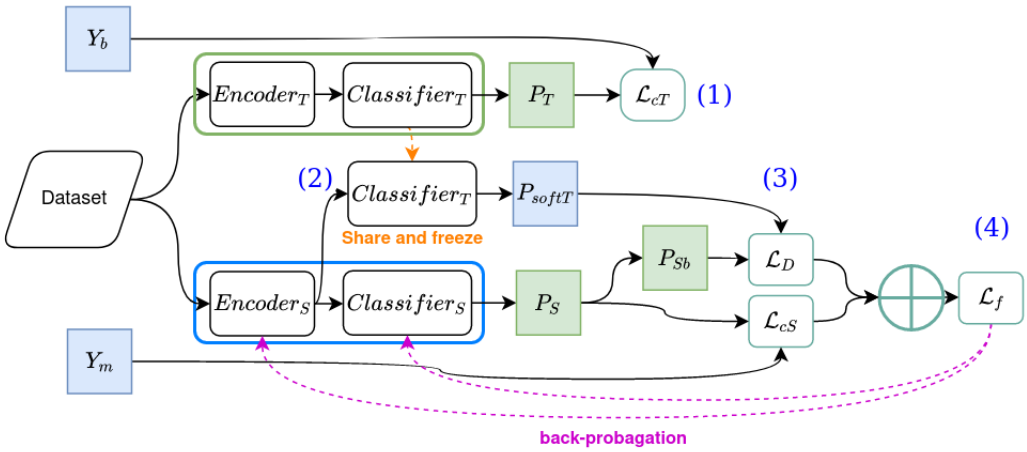

Figure 4.11 Binary soft-labels supervision architecture for Business Relation Extraction. (1) Teacher training, (2) Teacher classifier freezing and sharing, (3) Student training through knowledge distillation, (4) Final loss to train the student.

4.2.3 A Binary Soft-labels Supervision for Multi-class BRE (BSLS-RE)

We aim to answer one main research question:

— How can we adapt knowledge-distillation approach to better learn R^+ and handle R^- characteristics?

To this end, we propose *binary soft label supervision* approach for multi-class relation extraction (BSLS-RE) which is based on knowledge distillation where binary soft labels generated by a teacher model noted T are used to supervise the training of a student model noted S (cf. Figure 4.11). Following Zhou and Chen (2021), both S and T share the same architecture that has two main components:

(a) *a sentence encoder* noted $Encoder_i$ with $i \in \{S,T\}$ based on the pre-trained BERT model (Devlin et al., 2019) while using entity markers as sentence representation vectors,

(b) *a relation classifier* noted $Classifier_i$ composed of two linear layers followed by dropout layer then a softmax activation function.

An input sentence is first fed into $Encoder_i$, to get its contextual representations that are injected into $Classifier_i$ to predict the relation type. Let $P_i = (P_{i0}, ..., P_{in})$ the prediction probabilities generated by $Classifier_i$, with n being the number of relations to predict. Let P_{SoftT} the *soft labels*, i.e., the prediction probabilities generated by a pre-trained teacher

binary classifier $Classifier_T$ whose weights are frozen and shared with S. Finally, let Y_b and Y_m be respectively the binary and multi-class *hard labels* that encode the ground-truth labels as one hot vectors. These soft and hard labels are used by two different losses to optimize the models parameters through back-propagation: $\mathscr{L}_{cT}$ (resp. $\mathscr{L}_{cS}$) , the classification loss that minimizes the errors between P_T and Y_b (resp. P_S and Y_m). and $\mathscr{L}_D$, the distillation loss calculated between a binarised form of P_S and P_{SoftT}.

The distillation algorithm consists in the following steps:

(1) First, train T on binary relation identification (R^+ vs. R^-), while optimizing the teacher classification loss $\mathscr{L}_{cT}$.

(2) Then $Classifier_T$'s weights are frozen and shared with S.

(3) S is trained on multi-class RE and supervised by both Y_m and P_{SoftT}, while optimizing both the student classification loss $\mathscr{L}_{cS}$ and the distillation loss $\mathscr{L}_D$. To this end, P_S are first binarised into P_{Sb} following the equation in (4.10) where P_{S_0} refers to the prediction probability of R^- as given by $Classifier_S$.

$$P_{Sb} = (P_{S_0}, max(P_{S_1}, ..., P_{S_n})) \tag{4.10}$$

(4) The weighted sum of $\mathscr{L}_{cS}$ and $\mathscr{L}_D$ is the final loss $\mathscr{L}_f$ optimized to train the student model, α=0.6, $\beta = 0.4$, being loss weights.

$$\mathscr{L}_f = \alpha.\mathscr{L}_{cS} + \beta.\mathscr{L}_D \tag{4.11}$$

4.3 Experimental Settings and Baselines

4.3.1 Models Architecture

The three proposed approaches above (Section 4.2) are based on (Zhou and Chen, 2021) architecture for RE (cf. Figure 3.11, Section 3.3, Chapter 3). The sentence encoder that is used to generate a contextualized representation of the input sentence is initialize using three different pre-trained models to determine the impact of specialized vocabulary on the overall performance:

– **BERT** (Devlin et al., 2019). It is a transformer based PLM for English, trained on the BooksCorpus (800M tokens) (Zhu et al., 2015) and English Wikipedia (2,500M tokens) where only the text passages are used while ignoring lists, tables, and headers.

– **FinBERT** (Yang et al., 2020b). This is a transformer based PLM pre-trained on financial communication text to enhance financial NLP research and practice.[2] It is trained on the following three financial communication corpus: Corporate Reports 10-K & 10-Q (2.5B tokens), Earnings Call Transcripts (1.3B tokens), and Analyst Reports (1.1B tokens), giving a total corpora size of 4.9B tokens. FinBERT results in state-of-the-art performance on various financial NLP task, including sentiment analysis, ESG classification, forward-looking statement classification. We use two variants of this model:

  – **FinBERT**$_{finVocab}$ a variant trained from scratch for 1M iterations using a new financial vocabulary.

  – **FinBERT**$_{genVocab}$ a variant initialized from the original `bert-base-cased` model, and is further pre-trained on the financial corpora for 250K iterations at a smaller learning rate of $2e-5$.

4.3.2 Baselines

We compare our models against four baseline used to tackle data imbalance in RE: augmentation of the training data (DA), multitask architecture (MLT), optimizing using an adapted loss (ALS), and knowledge distillation (KD) via soft labels. We describe below each of these configurations.

1- Shortest dependency path data augmentation (DA$_{SDP}$) (Su et al., 2021a). As the shortest dependency path is assumed to capture the required information to express a relation between two target entities (Bunescu and Mooney, 2005), the augmentation consists in extracting tokens located in this path, fixing them, then the rest of tokens are randomly transformed by: synonyms replacement, random swapping, and random deletion. In our experiment, this method augments the positive instances by 300%.

2- Multitask architecture (MT-RE$_{all+bin}$): This is a multitask RE model that performs both relation identification (R^+ vs R^-) and relation extraction (multi-class classification). The relation identification task is an auxiliary task designed to help the main task of multi-class relation classification learn more features about R^+ vs. R^- distinction. It has been proposed before in (Lyu et al., 2020; Ye et al., 2019) for generic-relation extraction.

[2]https://github.com/yya518/FinBERT

3- Adapted loss (ALS) : We rely on four adapted losses as follows (cf. Section 4.1.2 for losses definitions and equations):

- Weighted Cross Entropy loss (ALS_{WCE});
- Focal loss (ALS_{FC}) (Lin et al., 2017);
- Adaptive scaling (ALS_{AD}) (Lin et al., 2018);
- Dice loss (ALS_{DC}) (Li et al., 2020b).

4- Soft label supervision using knowledge distillation (KD_{SLS}): Soft labels generated by a teacher model trained on a multi-class RE task are used to supervise a student model performing the same task. We use the focal loss to train the teacher model to handle class-imbalance when generating soft labels. This is the standard KD following (Hinton et al., 2015), where the teacher and the student models perform the same task, while only the teacher classifier is distilled as in (Song et al., 2021). Note that our teacher model is simpler, as it does not include any additional features.

4.4 Results and Discussion

4.4.1 Results

Results of the baselines, SADA-RE, MT-RE, and BSLS-RE experiments on the English BIZREL dataset using three different sentence-encoders are reported in Table 4.6 in terms of macro precision, recall, and F-score. The distribution of instances in our dataset can be found in Section 3.2.3, Chapter 3.

We presents our results focusing on the performances of sentences encoders, as well as each of the data augmentation strategies we considered.

Sentence encoders. Overall, most models based on simple BERT with a generic vocabulary trained on English Wikipedia outperform FinBERT, which either uses a fine-tuned generic vocabulary on financial texts or a financial vocabulary trained from scratch.

Data augmentation models. Our similarity-aware model (SADA-RE) outperforms the one based on shortest dependency path (DA_{SDP}) when using BERT or $FinBERT_{finVocab}$ (+1.2 and +0.9 respectively). DA_{SDP}, when BERT or $FinBERT_{genVocab}$ are used, achieves the best precision, outperforming all other models.

Table 4.6 Experimental results on the English BIZREL dataset. Best results per S. ENCODER are in **bold**, and best results are underlined.

S. ENCODER →	BERT			FinBERT$_{genVocab}$			FinBERT$_{finVocab}$		
MODELS ↓	P	R	F	P	R	F	P	R	F
DA$_{SDP}$ (Su et al., 2021a)	**69.7**	67.8	68.2	**70.1**	66.4	67.9	69.0	66.3	67.2
SADA-RE	66.6	72.8	69.4	63.4	68.8	65.6	67.2	69.2	68.1
MT-RE$_{all+bin}$ (Lyu et al., 2020)	62.8	73.2	67.2	66.8	68.1	67.2	66.3	69.6	67.8
MT-RE$_{all+biz}$	65.0	70.6	67.1	63.4	69.8	66.1	66.7	66.5	65.6
MT-RE$_{all+bin+biz}$	66.4	73.6	69.5	66.0	69.9	67.8	67.4	69.4	68.2
ALS$_{CE}$	62.5	72.5	66.7	64.1	68.4	66.0	65.9	68.9	67.3
ALS$_{WCE}$	63.1	**75.1**	68.1	60.7	73.4	65.7	64.2	**72.9**	67.8
ALS$_{FC}$ (Lin et al., 2017)	65.9	71.7	68.5	63.8	70.1	66.4	67.7	69.4	68.4
ALS$_{DC}$ (Li et al., 2020b)	66.9	65.4	65.7	63.0	61.8	61.4	61.2	47.0	52.3
ALS$_{AD}$ (Lin et al., 2018)	62.6	70.9	66.0	64.2	71.9	67.2	64.7	69.4	66.8
KD$_{SLS}$ (Song et al., 2021)	63.9	70.9	67.0	64.8	70.9	67.7	68.3	69.2	68.7
BSLS-RE$_{CE}$	65.4	71.7	68.2	66.9	70.4	68.3	66.1	69.6	67.6
BSLS-RE$_{WCE}$	63.0	73.2	67.1	63.5	**73.7**	67.8	64.1	70.7	66.9
BSLS-RE$_{FC}$	66.1	75.0	**69.9**	68.2	71.0	**69.5**	67.4	69.8	68.3
BSLS-RE$_{DC}$	66.7	69.8	68.1	67.3	67.1	66.8	67.4	62.3	64.0
BSLS-RE$_{AD}$	66.6	69.8	67.6	65.9	70.2	67.6	**69.7**	69.1	**69.2**

Adapted loss functions. Overall, the best performing models in terms of F-score are the ones optimized using a *focal loss* (ALS$_{FC}$) (when using BERT and FinBERT$_{finVocab}$, F1 is 68.5% and 68.4% respectively). In particular, the *weighted cross entropy loss* (ALS$_{WCE}$), when employing BERT, has the highest recall (75.1%) of all models.

Multitask models. Both BERT and FinBERT$_{finVocab}$ exhibit the same aspects. First, considering T_{bin} as an auxiliary task in the multitask model (MT-RE$_{all+bin}$) could improve the model recall, yet low precision. Second, when the auxiliary task is T_{biz} (MT-RE$_{all+biz}$), the model achieves a better precision compared to T_{bin}, however, the recall is still low. Combining the two auxiliary tasks $T_{bin} + T_{biz}$ with the main task (MT-RE$_{all+bin+biz}$) offers a good compromise between precision and recall, achieving therefore better F1-score than the two other multitask configurations.

Knowledge distillation models. When comparing between knowledge distillation models, we can observe that our *binary soft labels* (BSLS-RE) are more efficient than KD$_{SLS}$, the *multi-class soft labels* state-of-the art (+2.9%, +1.8%, and +0.5 F-score for BERT, FinBERT$_{gebVocab}$, and FinBERT$_{finVocab}$ respectively).

In general, we can observe that the best performing models for all encoders are those based on *binary soft labels supervision* (BSLS-RE), with the one using BERT optimized

Figure 4.12 Confusion matrix to compare between business and non-business instance classification in our best model (BSLS-RE$_{FC}$) and the best baseline (ALS$_{FC}$).

using a focal loss (*FC*) is the best, achieving an F-score of 69.9%, outperforming therefore all BERT baselines (+1.4% over the best one which is ALS$_{FC}$). This is also the case for both FinBERT variants where models based on *binary soft labels supervision* are the most effective (69.5% for BSLS-RE$_{FC}$ using generic vocabulary, and 69.2% F1 for BSLS$_{AD}$ using financial vocabulary).

When experimenting BSLS-RE with different loss functions, we can notice that, for most of the experiments, BSLS-RE optimized using $loss_i$ outperforms the baseline model optimized using the same $loss_i$. For example, for BERT encoder, BSLS-RE$_{CE}$ scores higher than ALS$_{CE}$ (+1.5 % F-score), BSLS-RE$_{DC}$ is better than ALS$_{DC}$ (+2.4 % F-score), BSLS-RE$_{AD}$ outperforms ALS$_{AD}$ (+1.6 % F-score), and finally BSLS-RE$_{FC}$ outperforms ALS$_{FC}$ (+1.4 % F-score). The same pattern in noticed for both FinBERT$_{gebVocab}$ and FinBERT$_{finVocab}$.

4.4.2 Analysis

We further compare the performances of the best baseline (ALS$_{FC}$) with our best performing model (BSLS-RE$_{FC}$) for BERT sentence encoder.

Binary confusion matrix. Figure 4.12 gives a confusion matrix that shows the number of false/true positives/negatives between R$^+$ and R$^-$. We can see that BSLS-RE$_{FC}$ was able to reduce the number of false negative instances (from 157 to 152), and increase the true negative (from 840 to 845). We can also observe the impact of these changes on the recall, where our model achieves one of the best score. It was however not able to reduce misclassifications due to false positive, leading therefore to a decrease in the precision when compared to the best precision.

R$^+$ performances. A closer look into the results per class for the best baseline and best performing model (cf. Table 4.7) shows that our model can improve the performances of

Table 4.7 Best baseline (ALS_{FC}) and our best model (BSLS-RE_{FC}) F1-score per relation type. Best results of each relation are in bold.

	Inv.	Com.	Coo.	Leg.	Sal.	Oth.
ALS_{FC}	61.0	**78.8**	65.0	**77.8**	41.9	**86.6**
BSLS-RE_{FC}	**68.9**	77.2	**66.7**	73.7	**46.2**	**86.6**

most under-represented positive relations, namely: INVESTMENT, COOPERATION and SALE-PURCHASE that represent 3.3%, 7.3% and 2.9% of test set. R^- results remain stable and this was expected as our approach was specifically designed to handle under-represented R^+. A final interesting finding is that R^+ with less frequencies are the one that benefits the most from *binary soft labels*. For example, an improvement of +7.9 % (resp. +4.3 %) in terms of F1 is observed for under-represented relation INVESTMENT (resp. SALE-PURCHASE) over the best baseline.

Strength of BSLS-RE. In order to gain insights into the main strengths of the current approach when compared to the best baseline, we analyze well classified instances by BSLS-RE_{FC}, that ALS_{FC} fails to classify correctly. We notice that our approach can identify the R^- OTHERS in some cases where many relations are expressed between different target entities, unlike ALS_{FC} (See Example (4)).

(4) While there were few mega acquisitions/ mergers primarily Chinese players acquiring European and US robotics/ automation companies (Kuka AG by**[Midea Group]**$_{EO_1}$, Dematic by **[Kion Group]**$_{EO_2}$ and KraussMaffei Automation by ChemChina) and few others by US industry giants (Affeymetrix by ThermoFisher and Intelligrated by Honeywel), most acquisitions were in the sub $ 500 M range.
BSLS-RE_{FC} correct label : OTHERS,
ALS_{FC} wrong label: INVESTMENT.

In addition, our model can also distinguish between semantically close R^+ such as INVESTMENT, SALE-PURCHASE, and COOPERATION, that uses the same lexical cues to be expressed such as *signing agreement*, *entering into a contract*. In Example (5), the expression *entering into a contract* refers to *service-selling* contract rather than a COOPERATION relation.

(5) **[General Electric Corporation]**$_{EO_1}$ has entered into a five – year, $ 128,500 million contract with **[Electronic Data Systems]**$_{EO_2}$ (EDS) to handle the corporation's desktop computer procurement, service, and maintenance activities.

BSLS-RE$_{FC}$ correct label : SALE-PURCHASE,
ALS$_{FC}$ wrong label: COOPERATION

4.5 Portability to the Biomedical Domain

We study the portability of our approached by evaluating their performances on another domain specific imbalanced RE dataset, ChemProt (Krallinger et al., 2017), that focuses on chemical-protein interactions extraction. This dataset is composed of five R^+ and a R^- representing 75% of whole the dataset. Table 4.8 summarizes the distributions of relations per type and per train/dev/test datasets.

Table 4.8 Relation instance distribution per relation type (Rel.) and per dataset (train/dev/test).

REL.	SEMANTIC MEAN.	TRAIN	DEV	TEST
CPR:3	UpRegularor/Activator Indirect_UpRegularor	756	546	663
CPR:4	DownRegularor/Inhibitor Indirect_DownRegularor	2,227	1,091	1,655
CPR:5	Agonist/Agonist-Activator Agonist-Inhibitor	173	115	178
CPR:6	Antagonist	229	199	292
CPR:9	Substrate/Product_Of Substrate_Product_Of	727	457	642
false	negative relation	13,923	8,860	12,315

We experiment with our best performing model, namely BSLS-RE and the best baseline (i.e., ALS), while selecting the loss functions for which BSLS-RE outperforms ALS in Table 4.6.

We use BioBERT (Lee et al., 2020) pre-trained language model as a sentence encoder in our model, as it has shown to outperform BERT on the biomedical RE (Lee et al., 2020). BioBERT was first initialized with weights from BERT, which was pre-trained on general domain corpora (English Wikipedia and BooksCorpus). Then, BioBERT is pre-trained on biomedical domain corpora, including 4.5B words from PubMed abstracts and 13.5B words from PMC full-text articles.

Table 4.9 Experimental results on the ChemProt dataset. Best results are in bold.

MODEL	P	R	F1
ALS_{CE}	77.7	**75.1**	76.4
ALS_{FC}	78.4	73.6	75.9
ALS_{AD}	79.3	74.4	**76.8**
ALS_{DC}	78.0	66.7	72.0
BSLS-RE_{CE}	78.7	74.0	76.3
BSLS-RE_{FC}	**79.8**	73.1	76.3
BSLS-RE_{AD}	79.2	74.5	**76.8**
BSLS-RE_{DC}	73.4	73.6	73.5

The results of experiments are reported in Table 4.9 in terms of micro precision (P), micro recall (R), and micro F1-score (F1).[3] We notice that the BSLS-RE used to extract chemical-protein interaction achieves the highest score. When using *CE* loss, BSLS-RE has a slightly negative impact (-0.1%), but no impact when using *AD* loss. However, BSLS-RE optimized with *FC* and *DC* losses outperforms its baseline counterpart optimized with the same losses (+0.4% and +1.5%, for *FC* and *DC* losses, respectively), confirming partially the effectiveness of *binary soft label supervision* for imbalanced R^+ vs. R^- RE.

4.6 Conclusion

In this chapter, we proposed various solutions to the problem of R^+ vs. R^- imbalance and R^- irregular patterns. The first is based on *multitask learning*, in which the binary classification R^+ vs. R^- and the classification of R^+ are regarded as auxiliary tasks optimized along with the main task of RE. The second solution aims to *augment* R^+ by leveraging vector representations of relation types and using non-annotated sets of data to find instances that are most similar to them. The final solution is based on *binary soft-labels supervision* produced by knowledge distillation, where the teacher model trained on relation identification task transfer salient knowledge about R^+ vs. R^- to the student model that performs a RE task.

When evaluated on a business relation dataset, our approaches improved the overall performances outperforming strong state-of-the-art solution to handle R^+ vs. R^- imbalance. *Binary soft-labels supervision* was the most productive one and enhanced the detection of under-represented relations while reducing false negative misclassification rates. As a future direction for this approach, the teacher model can be enhanced with semantic or syntactic knowledge, making the produced soft-labels more informed and thus contributing to improv-

[3] Micro are the official measures for the ChemProt dataset

ing the student model's performance. We finally studied the portability of this approach to another domain-specific dataset. Our results partially confirm the effectiveness of binary soft label supervision for imbalanced R^+ vs. R^-. This work has been published as long papers in two workshops in AAAI 2021 and IJCAI 2022 conferences (Khaldi et al., 2022a,b)

In the following chapter, we continue presenting our contributions, focusing this time on the role of external knowledge about target entities, as given by structured resources such as KB, to improve a BERT-like model at various levels of representation.

Chapter 5

Multi-level Entity Enhanced RE

The use of pre-trained language models (PLM) (cf. Section 1.4.5, Chapter 1) has further improved the performances of RE task, where the representation of the input sentence accounts for the contextual meaning of each of its words including target entities by leveraging the power of multi-head self-attention mechanism. Recently, knowledge from external resources have been exploited to further improve entities representations in these models.

This chapter describes a business relation extraction system that combines contextualized representations from PLM with multiple levels of entity knowledge. We propose multiple neural architectures based on BERT, newly augmented with three complementary levels of knowledge about entities: generalization over entity type, pre-trained entity embeddings learned from two external knowledge graphs, and an entity-knowledge-aware attention mechanism. Our results show an improvement over many strong state-of-the-art models for relation extraction.

In this chapter, we first review the state-of-the-art about knowledge enhanced models proposed for RE in Section 5.1, and available resources for entity embeddings in Section 5.2. Section 5.3 describes the proposed neural architecture. Experimental settings and the baselines to which we compare our model are presented in Section 5.4. Section 5.5 reports the obtained results on the English BIZREL. We finally end this chapter by a portability study showing how our proposed approach can improve the extraction of business relations from French content by evaluating it on the French BIZREL (cf. Section 5.6).

5.1 Knowledge Enhanced RE models

5.1.1 Transformer-based Approaches

PLMs usually learn universal language representation from general-purpose large-scale text corpora, but lack domain-specific knowledge. Incorporating domain knowledge from external KBs into PLM has shown to be effective for many NLP tasks including RE where the external knowledge ranges from linguistic (Levine et al., 2020; Peters et al., 2019; Wang et al., 2020b), commonsense (Guan et al., 2020), encyclopedic (Liu et al., 2020; Peters et al., 2019; Wang et al., 2019; Zhang et al., 2019), to domain-specific knowledge (Liu et al., 2020).

In this dissertation, we refer to these models as *knowledge informed pre-trained language models* (*Kin*) while those that do not include any knowledge are referred to as *knowledge agnostic pre-trained language models* (*Kag*). We focus in this section on Kin, the reader can refer to Section 1.4.5 Chapter 1 for an overview of Kag.

Wei et al. (2021) provide a comprehensive survey about *Kin* models. We mainly focus on the ones that have been proposed for/evaluated on RE task. Roughly, these models have the following components:

(a) Source of knowledge. Encyclopedia KGs such as Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić and Krötzsch, 2014) have been the primary source of knowledge for these models, as they contain structured facts/world knowledge about entities in the form of triples (e1, R, e2) that can be used to perform entity-aware training for PLMs.

(b) Level of knowledge. Knowledge from KGs can be injected at different levels of granularity:

- *Entity level* where entity-aware objectives are introduced during PLM pre-training. Entity linking is a typical example which predicts the entity mention in text to entity in KG with a cross entropy loss or max-margin loss on the prediction (Peters et al., 2019; Poerner et al., 2020; Yamada et al., 2020b; Zhang et al., 2019);

- *Relation level* where links between entities in the KG - that are stored as relation triples- are exploited using different pre-training objectives such as link prediction in the KG (Liu et al., 2020), or augmenting sentence with the triplets from KG to transform it into a knowledge-rich sentence tree (Wang et al., 2019);

- *Sub-graph level* where the exploited information in a KG is expanded by considering a sub-graph involving many entities and relations. This structure is either injected using a GNN (Graph Neural Network, described in Section 1.4.3, Chapter 1) that

incorporates the learned features into the PLM (Su et al., 2021b), or it is transformed into a token sequence and appended to the input sentence (Sun et al., 2020b).

(c) Methods of knowledge injection. We mainly identify two different ways to augment PLM with external knowledge:

(1) Adding new entity-knowledge related objectives while retraining a language model or changing it core architecture: ERNIE (Zhang et al., 2019) and KnowBERT (Peters et al., 2019) adopt a same idea of using static entity embeddings separately learned from a KB while jointly training an entity linker and the language model. Conversely, KEPLER (Wang et al., 2019), a unified model based on RoBERTa (Liu et al., 2019), jointly learns knowledge embeddings from their descriptions and a language model resulting in one unified *Kin*, aligning therefore the factual knowledge and language representation into the same semantic space. Instead of only using textual data to learn entity knowledge, CoLAKE (Sun et al., 2020b) extracts the knowledge context of an entity from large-scale knowledge bases and integrates it with language context into a unified data structure named word-knowledge graph, to learn contextualized representations for both language and knowledge. Furthermore, to capture complex relationships between words and entities in PLM through self-attention mechanism, LUKE (Yamada et al., 2020b) treats both words and entities as independent tokens using as extended MLM objective that masks both words and entities while training the language model. The model also proposes an entity-aware self-attention mechanism that is an extension of the self-attention mechanism of the transformer, and considers the types of tokens when computing attention scores. On the other hand, K-BERT (Liu et al., 2020) explicitly injects related triples extracted from KG into the sentence to obtain an extended tree-form input for BERT during fine-tuning on downstream tasks. The BERT architecture is, however, adapted to control the visible area of each token, preventing changing the meaning of the original sentence due to too much knowledge injected from KG.

(2) Adapting the injected knowledge representations before injecting them directly to the PLM without requiring its retraining or changing it core architecture: This was adopted because adding new pre-training objectives to PLM is extremely computationally expensive and may result in catastrophic forgetting of distributional knowledge. E-BERT (Poerner et al., 2020) injects Wikipedia2Vec entity vectors as a source of knowledge into the PLM and aligns them with BERT native wordpiece vectors, then directly add them to its vocabulary without additional retraining. Following the same approach, K-Adapter (Wang et al., 2020b) relies on neural adapters that are plugged outside RoBERTa PLM injecting factual knowledge that comes from automatically aligned text-triples on Wikipedia and Wikidata and linguistic knowledge obtained from dependency parsing while supporting

continuous learning. Considering, the RE task as a prompt-tuning task, (Chen et al., 2022b) proposed KnowPrompt that injects entity-related and relation-related latent knowledge into prompt construction in the form of entity type tokens and relation embeddings. More recently, (Papaluca et al., 2022) combined independently pre-trained embeddings on a KG with RE-fine-tuned PLM via simple concatenation and obtained comparable results to state-of-the-art models on the Wikidata and NYT datasets while requiring very little computational power when compared to other *Kin*.

Table 5.1 gives an overview of the main existing *Kin* models for RE. Most of them require knowledge-aware pre-training of the PLM to inject factual knowledge into it, despite their high computational need to be constructed and the difficulty of their adaptation to inject new sources of knowledge. We note that all the models were designed for English, except for K-BERT that targets Chinese.

Table 5.1 Comparison between knowledge enhanced language models for RE, inspired from (Wei et al., 2021) and (Hu et al., 2022).

METHOD	Knowledge Aware pre-training	How is Knowledge Injected	Knowledge Source	Type of knowledge
ERNIE (Zhang et al., 2019)	Yes	entity prediction	Wikipedia/Wikidata	entity level
CokeBERT (Su et al., 2021b)	Yes	entity prediction	Wikipedia/Wikidata	sub-graph level
KnowBERT (Peters et al., 2019)	Yes	entity linking	WordNet, Wikipedia	entity level
KEPLER (Wang et al., 2019)	Yes	TransE scoring	Wikipedia/Wikidata	relation level
LUKE (Yamada et al., 2020b)	Yes	entity prediction	Wikipedia	entity level
CoLAKE (Sun et al., 2020b)	Yes	masked entity prediction	Wikipedia/Wikidata	sub-graph level
K-BERT (Liu et al., 2020)	No	finetuning on RE	WikiZh, WebtextZh, CN-DBpedia, HowNet, MedicalKG	relation level
K-Adapter (Wang et al., 2020b)	No	dependency relation	Wikipedia, Wikidata, Stanford Parser	relation level
E-BERT (Poerner et al., 2020)	No	entity/wordpiece alignment	Wikipedia2Vec	entity level
KB-embed-RE (Papaluca et al., 2022)	No	pre-trained KB embedding	Wikidata	entity level
KnowPrompt (Chen et al., 2022b)	No	finetuning on RE	entity type tokens relation embedding	entity level relation level

5.1.2 Other Neural Architectures

Other works inject entity knowledge into neural models through an entity knowledge attention mechanisms, which can obtain additional information about the input entity pairs from a

knowledge base (KB) and inject it into the relation instance representation through an attention mechanism. For example, Li et al. (2020c) proposed a model that acquires prior knowledge about genes and proteins from KBs (UniProt and BioModels) and determines their correlation to words in a sentence through attention mechanisms. The same idea has been adopted by Li et al. (2019) who proposed a dual CNN that uses a knowledge-based relation attention to mine the relationship representation of current entity pairs in a KB and also exploit information about other relationships involving these entities.

5.2 Entity Embedding from KBs

In addition to the availability of unstructured textual data, structured textual data in KB, such as Wikidata (Vrandečić and Krötzsch, 2014), Freebase (Bollacker et al., 2008), and BabelNet (Navigli and Ponzetto, 2012), has emerged in recent years.

There has always been a connection between KB resource and the RE task. KB was built first with RE models, then used to automatically generate training datasets for this task via distant supervision (Mintz et al., 2009; Zeng et al., 2015). It has also been shown that the relation triples stored in these resources can be used to further supervise and improve the overall performance of RE models (Ji et al., 2017; Yamada et al., 2020b; Zhang et al., 2019).

As previously stated, injecting factual knowledge about target entities into RE models has proven to be very effective. The majority of the works focused on dynamically learning representations of KB entities while simultaneously learning word representations (Bastos et al., 2021; Vashishth et al., 2018; Yamada et al., 2020b). Others attempted to use the information contained in static pre-trained representations of entities from a KB (Papaluca et al., 2022; Poerner et al., 2020). This section focuses on the second category and sheds light on the available entity-embeddings resources.

– Wikipedia2vec.[1] It was used by Poerner et al. (2020) as a source of entity knowledge for RE. It is based on the idea of jointly mapping words and entities into the same continuous d-dimensional vector space. These vectors are learned by exploiting both structural and textual data from Wikipedia using an extended word-skip-gram model (Mikolov et al., 2013a,b) by adding the link graph model and the anchor context model (Yamada et al., 2020a) (cf. Figure 5.1).

[1] https://wikipedia2vec.github.io/wikipedia2vec/

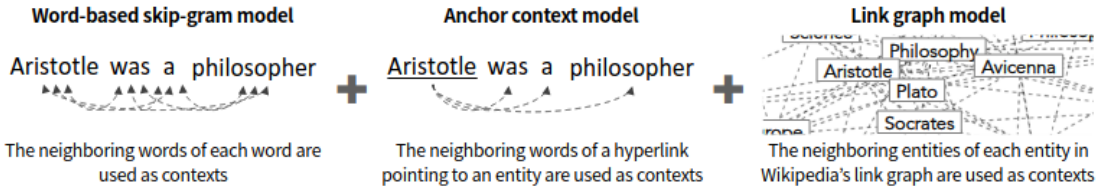

Figure 5.1 Wikipedia2Vec learns embeddings by jointly optimizing word-based skip-gram, anchor context, and link graph models (Yamada et al., 2020a).

- *Word-based skip-gram model.* By predicting the surrounding words of a given word, this model attempts to learn word embedding representations for each word in Wikipedia articles.
- *Anchor model.* Entities are referred to by the hyperlinks in a Wikipedia page. This model aims to predict the surrounding words of an entity to learn its embedding vector. In the vector space, the learned entity representation is placed closer to its surrounding words representations.
- *Link Graph Model.* This model uses Wikipedia link graph that is composed of Wikipedia pages as nodes and the hyperlinks between them as edges. The entity representation is learned by predicting the surrounding entities of a given entity in this graph.

– NASARI. [2] This resource proposed by (Camacho-Collados et al., 2015) combines complementary knowledge from different types of KB resources, including:

- lexico-semantic relations from the expert-based lexicographic WordNet (Miller et al., 1990) that covers highly accurate encoding of concepts and semantic relationships between them with a limited lexical coverage.
- texts of articles and the inter-article links of the collaboratively constructed encyclopedic Wikipedia that provides wide coverage of articles about named entities, concepts, and domain-specific lexicon that is frequently updated.[3]
- synset-to-article mappings provided by BabelNet (Navigli and Ponzetto, 2012) resource to link WordNet synsets and Wikipedia articles.

[2] http://lcl.uniroma1.it/nasari/

[3] English Wikipedia alone receives 566 new articles per day. Source: https://en.wikipedia.org/wiki/Wikipedia:Statistics

The concept (Wikipedia page p or WordNet synset s) modeling approach consists of three steps: first *information context collection* then *lexical vector computation*, and finally *embedding vector generation*, as follows:

- *Information context collection:* By leveraging the structural information in Wikipedia and WordNet, a set of relevant Wikipedia pages for a given concept is compiled. First, BabelNet is used to map a Wikipedia page to its Wordnet synset (and vice-versa). The first contextual information to consider for a Wikipedia page p is Wikipedia pages with an outgoing link to this page. For a synset s, all other synsets in its immediate surroundings are mapped using BabelNet to obtain their corresponding Wikipedia pages, which serve as the second contextual information to be used. Figure 5.2 summarizes this first step.

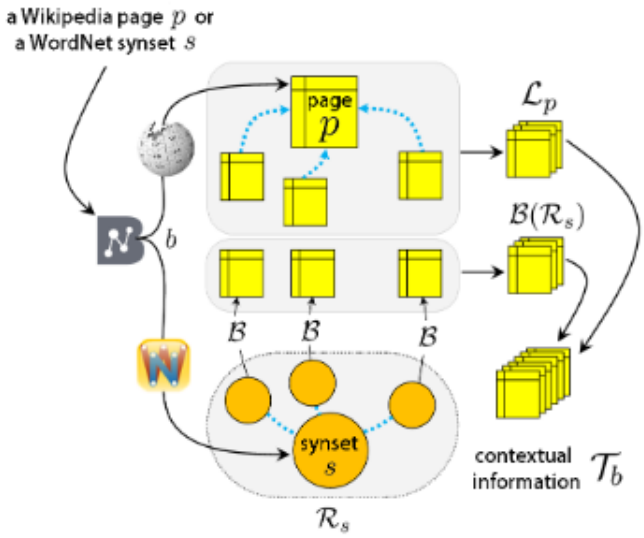

Figure 5.2 NASARI embeddings: The process of obtaining contextual information from a WordNet synset or a Wikipedia article (Camacho-Collados et al., 2016)

- *Lexical vector computation:* Lexical vectors have individual words as their dimensions. Instead of relying on tf/idf measure, lexical specificity (Lafon, 1980) is used to compute a weighted set of most representative words from the contextual bag-of-words with respect to the reference corpus, i.e., the whole Wikipedia.[4]

- *Embedding vector generation.* By exploiting the compositionality of word embeddings,[5] a trained word embeddings representations are used to generate the embedding

[4]More details about the calculation process are given in (Camacho-Collados et al., 2015, 2016).

[5]For example, the vector representation obtained by averaging the vectors of the words *Vietnam* and *capital* is very close to the vector representation of the word *Hanoi* in the semantic space of word embeddings.

vector corresponding to the lexical vector by averaging the embedding of its words while giving more importance to the higher weighted dimensions.

NASARI has already been employed in many NLP tasks such as word similarity (Camacho-Collados et al., 2016) and word sense disambiguation (Pasini and Navigli, 2020) but never for multi-class RE, as far as we know.

– Wikidata KG embeddings. These are pre-trained graph embeddings of entities provided by Lerer et al. (2019), who generated entity representations with an embedding dimension of 200 using a TransE algorithm (Bordes et al., 2013) on a Wikidata dump dated 2019-03-06. These vectors have been used in (Papaluca et al., 2022) as additional features to enhance a RE model.

– Wiki2vec[6] and Wikipedia Entity Vectors (WikiEntVec).[7] Entity embeddings are trained on Wikipedia as an input text using a standard word2vec model, where the entity annotations are replaced in the Wikipedia page with the unique identifier of their referent entities.

– RDF2vec. [8] Ristoski and Paulheim (2016) learned entity embeddings using the skip-gram model with inputs generated by random walks over a large RDF knowledge graphs such as Wikidata and DBpedia.

The main characteristics of the entity embeddings mentioned above are summarized in Table 5.2. The majority of resources used one KB to calculate the embedding vectors, except for the NASARI resource, which used WordNet and Wikipedia, and BabelNet as a bridge between them. Most of the released pre-trained entity embeddings are for English language, with the availability of the training code to calculate them for other languages if the KB used is available in these languages. Wikipedia2vec is the only multilingual resource available in 12 languages including: English, French, Spanish, German, Arabic, Chinese, Dutch, Italian, Russian, Polish, Portuguese, and Japanese. NASARI, on the other hand, is released for English and Spanish.

[6]https://github.com/idio/wiki2vec
[7]https://github.com/singletongue/WikiEntVec
[8]https://github.com/IBCNServices/pyRDF2Vec

Table 5.2 Entity embedding resources.

METHOD	SOURCE OF KNOWLEDGE	WAY OF LEARNING	LANG. OF RELEASED PRETRAINED VEC.
Wikipedia2vec	Wikipedia text Wikipedia hyper-links	word skip-gram+ link skip-gram+ anchor skip-gram	12 lang.
NASARI	Wikipedia Wordnet BabelNet	embedding of lexical vectors calculated using lexical specificity.	EN, ES
Wikidata KG	Wikidata	TansE algorithm on KG	EN entries
wiki2vec WikiEntVec	Wikipedia	word2vec	EN, DE, JA
rdf2vec	DbPedia or Wikidata	word2vec	EN

5.3 Multilevel Entity-Informed RE

While *Kin* models have shown to be quite effective for extracting generic relations, their use in business RE has not been investigated yet. In this section, we propose the first *Kin* model for business RE based on simple neural architectures that require neither additional training to learn factual knowledge about entities nor adaptation of knowledge representations. Hence, knowledge about entities is viewed as external features to be injected into the relation classifier along with the sentence representation (as given by BERT).

Compared to existing *Kin* models where different sources of knowledge about entities (entity type, pre-trained entity embeddings (P-EE), entity-aware attention mechanism) have been considered independently, as far as we know, no prior work attempted to measure the impact of combining multiple levels of knowledge on the performances of RE.

Our work is more comparable to E-BERT (Poerner et al., 2020), K-adapter (Wang et al., 2020b), and the model proposed by (Papaluca et al., 2022) in a way that we aim to inject factual knowledge about entities into PLM in a low computational way without extending the pre-training of PLM with knowledge-related objectives. In contrast to them, our work injects knowledge about entities at different level of the sentence representation and leverage pre-trained entity embeddings trained on raw and structured data from multiple KBs to extract the most informative elements in a sentence through an entity-knowledge attention

mechanism.

The architecture of our model (shown in Figure 5.3) is based on BERT PLM as a sentence encoder to encode the input sentence tokens and entity mentions into contextualized representations, as it has shown to be a quite effective language encoder for RE (see Section 1.4.5). The model is augmented at multiple levels with knowledge about entities. We newly consider three main levels of additional knowledge: entity generalization, multi-source entity embedding, and knowledge attention mechanism. We describe in the following the main components of this architecture.

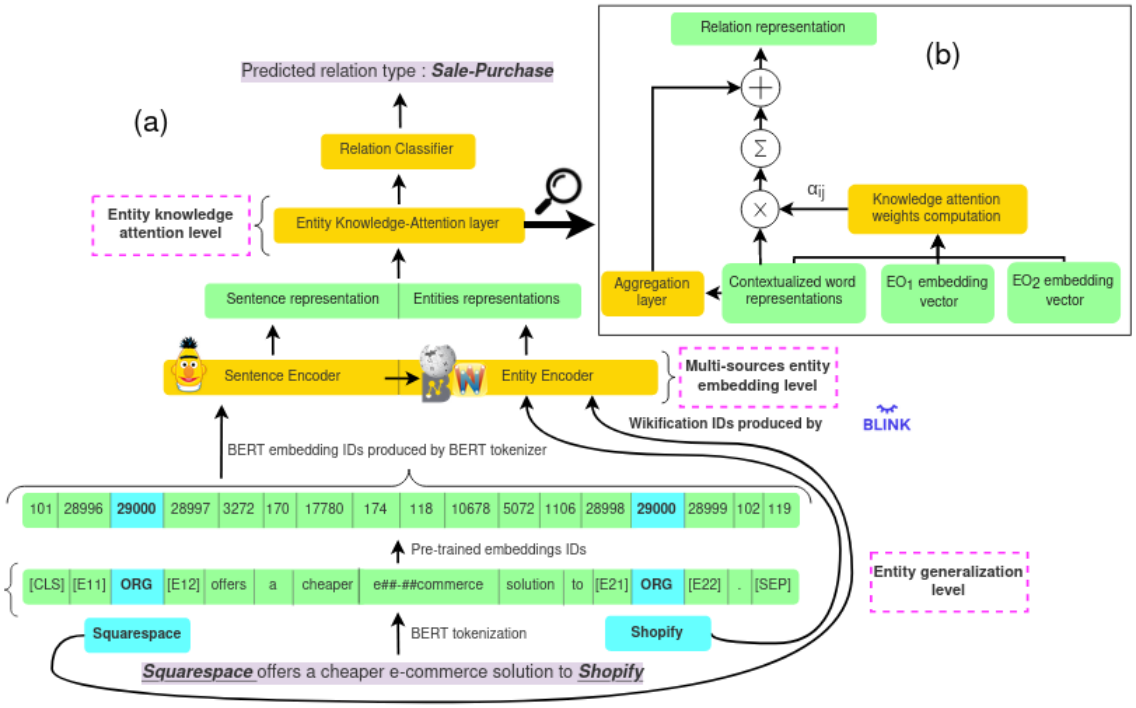

Figure 5.3 (a) Our multilevel entity-informed model for business relation extraction and (b) a detailed description of our knowledge-attention mechanism.

5.3.1 Input Representation

To capture the position of the two target entities EO_1 and EO_2 within the input sentence, both their beginning and ending are marked by specific tokens following (Baldini Soares et al., 2019): $[E_{11}]$, $[E_{12}]$ for EO_1 and $[E_{21}]$, $[E_{22}]$ for EO_2. We also add the classification token $[CLS]$ to the beginning of each sentence and $[SEP]$ to mark its end. The label ORG

referring to the entity type is considered as a generic term for replacing EO_1 and EO_2. Since these entities can belong to several subtypes (e.g., *startup*, *company*), this generalization strategy aims to prevent overfitting and helps the classifier to reason at the entity level rather than the entity mention itself which may be infrequent in the corpus or over-represented (cf. statistics about duplicated entities per relation type in Table 3.4 in Section 3.2.3, Chapter 3). For example, some entity pairs (e.g., *(Google, Microsoft)*) can be very frequent for a given relation (e.g., COMPETITOR) but rare for others (e.g., COOPERATION). This is *the first level that injects entity knowledge* to our model, and it extends argument entity type masking (Shi and Lin, 2019) and semantic type labeling (Wei et al., 2019) respectively proposed for generic and biomedical relations to business relations.

5.3.2 BERT Encoder

To encode input text, we use **BizBERT**, the BERT PLM fine-tuned on our business dataset, as it has shown to be an effective language encoder for many NLP tasks (Chiril et al., 2020; Li et al., 2020a) and particularly for RE task (Wu and He, 2019). The encoder takes a sequence of N tokens $(x_1, ..., x_N)$ as input, where an entity EO_i can be composed of several tokens. It then computes L layers of d-dimensional contextualized representations $H_i \in R^{N \times d}$, $1 \leq i \leq L$. Each layer of the encoder E_i is the combination of a multi-head self-attention and a multi-layer perceptron, and the encoder gets the representation of each layer by the following formula.

$$H_i = E_i(H_{i-1}) \tag{5.1}$$

Given a sentence with two entities EO_1 and EO_2, suppose the final hidden state output from BERT encoder is H. Let entity vector representations be H_{EO_1} and H_{EO_2}, obtained by applying an average operation on the final hidden state vectors (H_k^i to H_m^i) from BERT for each EO_i. Let the final hidden state vector of the classification token (i.e., '$[CLS]$') be H_{cls}. The vectors H_{EO_1}, H_{EO_2}, H_{cls} are the outputs of the encoder and will be fed into the next components.

5.3.3 Entity Encoder

The entity encoding consists of two steps: first, entity linking and disambiguation, then entity embedding lookup. The encoder first links every EO_i in the input sentence to its unique disambiguated textual identifier on Wikipedia using BLINK (Li et al., 2020a),[9] an open-

[9]https://github.com/facebookresearch/BLINK

source entity linker. For example, for an entity mention *Amazon* referring to the company *Amazon* in an input sentence, the entity linker returns the unique textual ID from Wikipedia *Amazon (company)* rather than *Amazon river* or *Amazon rainforest*. The entity ID is then used to get the entity dense representations from one of the two pre-trained entity embedding resources NASARI and Wikipedia2Vec that are described in Section 5.2.

We select these resources for the following reasons:

- NASARI resource is used to be able to inject knowledge about entities from different sources, as it was constructed based on Wikipedia and WordNet resources. Moreover, this resource has already been employed in many NLP tasks such as word similarity (Camacho-Collados et al., 2016) and word sense disambiguation (Pasini and Navigli, 2020) but never for multi-class RE, as far as we know.

- Wikipedia2vec is used to be able to evaluate our method on many languages, as it is available for 12 different languages.

Injecting P-EE into our model represents *the second level of entity knowledge*. In the course of the experiments, approximately 92% of entities in the training set can be found in Wikipedia2Vec English, and almost 83% of them in NASARI. When combining both resources, the coverage increases to 94%. P-EE can come from Wikipedia2Vec alone, NASARI alone, or both. In the last case, we first look up the P-EE in Wikipedia2Vec, as this resource has the best coverage rate with the training set. If it is not found, we use its NASARI representation instead. Otherwise, (i.e., the entity does not exist in both resources), its embedding vector is randomly initialized. The outputs of the entity encoder are 300-dimension vectors denoted K_{EO_i} (one vector K per entity EO_i).

5.3.4 Sentence-features Layer

Three type of aggregation layers are used to extract a single dense sentence representation vector (noted S) from the contextualized word representations (H_0, .., H_M) produced by the sentence encoder :

- **BizBERT**: we use the final hidden state generated by the sentence encoder for the $[CLS]$ classification token added at the beginning of each sentence as a sentence representation vector, as proposed in BERT's original paper (Devlin et al., 2019).

- **CNN-BizBERT**: a convolutional layer that applies 1-dimensional convolution followed by a max-pooling and an activation function is applied on (H_0, .., H_M).A dropout layer is added to reduce overfitting.

- **BiLSTM-BizBERT**: a BiLSTM layer is applied on (H_0, .., H_M) to extract a dense sentence vector using two LSTM layers taking into account left and right contexts of each word. We use the concatenation of the last hidden layer and the last cell state as a sentence representation vector, and add a dropout layer to prevent overfitting.

5.3.5 Knowledge-attention Layer

The third level of knowledge is a *knowledge-attention mechanism* (Figure 5.3 (b)) that exploits structural knowledge and statistical information about entities derived from semantic networks (e.g., WordNet, BabelNet, graph generated by links between Wikipedia pages) and text corpora (e.g., Wikipedia text) as given by NASARI and Wikipedia2Vec embeddings in order to focus on the most important words in a sentence that contribute significantly to the relation representation. Knowledge-attention has already been employed to select the most relevant entities from KBs to be integrated with sentence representation (Li et al., 2020c), or to incorporate information about how entities are linked in KBs (Li et al., 2019). Here, we adopt a different strategy by using pre-trained entity embeddings (i.e., K_{EO_1}, K_{EO_2}) to assign to each contextualized word representation of an input sentence (noted H_i, $i \in [0,N]$) an importance weight a_{ij}, calculated by Equation 5.3. The attention weight tells how much this word is correlated to the knowledge encapsulated by the entity vector. We assume that this information could be efficient for RE task.

$$Z = \sum_j a_{ij} H_i \tag{5.2}$$

$$a_{ij} = softmax(H_i^T W K_{EO_j}) \tag{5.3}$$

The sentence representation vector resulted after applying knowledge attention on contextualized word representations, noted Z (cf. Equation 5.2), is merged with the sentence representation vector S produced by the sentence-features layer through a weighted sum (cf. Equation 5.4).

$$S_f = \frac{1}{2}(a.S + b.Z) \tag{5.4}$$

5.3.6 Relation Classifier

A fully connected layer followed by a softmax classifier is applied on top of the final sentence representation vector S_f produced by the knowledge-attention layer concatenated with entity vectors (K_{EO_1}, K_{EO_2}, H_{EO_1}, H_{EO_2}) generated by the entity and sentence encoders, to produce a probability distribution p that predicts the relation type (cf. Equations 5.5 and 5.6) where

W and b are learnable parameters, with $W \in R^{|Y| \times d'}$, $b \in R^{|Y|}$, $|Y|$ is the number of relation types and d' is the classifier input dimension.

$$h = W[concat(S_f, H_{EO_1}, H_{EO_2}, K_{EO_1}, K_{EO_2})] + b \tag{5.5}$$

$$p = softmax(h). \tag{5.6}$$

We consider two configurations for the classifier: *monotask learning* and *multitask learning*. The first one is a multi-class learning problem, where the classifier has to predict the relation type that links a pair of entities (EO_i, EO_j) in a given sentence among the six relations that we consider (including OTHERS). The second configuration is designed to deal with data imbalance (cf. Section 4.1.2, Chapter 4), following recent studies that show that jointly learning common characteristics shared across multiple tasks can have a strong impact on RE performances (Yadav et al., 2020; Ye et al., 2019). To this end, we jointly train two classifiers using multitask objectives. The first one performs *relation identification* to detect whether a business relation holds between a given entity pair or not (i.e., business vs. non-business). It is trained on a more balanced dataset (business (37%) vs. non-business (63%)) to optimize a binary cross-entropy loss CL_{bin}. The second classifier performs *relation classification* and learns how to predict the relation type between two EO_i (this is a 6-class classification task) with a multi-class cross-entropy loss CL_{mlt}.

This configuration allows to learn common features about business relation types and to share knowledge among the two classifiers when trained jointly by multitask objectives, as shown in Equation 5.7 where α and β are weights associated to each classifier loss.

$$L = \alpha \cdot CL_{mlt} + \beta \cdot CL_{bin} \tag{5.7}$$

5.4 Experimental Settings and Baselines

We experiment with different models $\mathscr{M}_{\mathscr{E}}$ while varying the aggregation layer $\mathscr{M}$ (BizBERT, BizBERT+CNN, BizBERT+BILSTM) and the entity knowledge levels $\mathscr{E}$ (t, wiki, nas, att) among entity type generalization (t), multi-source entity embeddings from either Wikipedia (*wiki*) or NASARI (*nas*), and entity-attention (*att*).

In our experiments, the sentence encoder relies on the `bert-base-cased` model implemented in the HuggingFace library.[10] The sentence encoder always outputs a sentence representation of dimension 768, either using the BERT's [CLS] final embedding, a CNN

[10]https://huggingface.co/bert-base-cased

with a kernel size set to 5 applied to all the contextualized embeddings, or a BiLSTM with hidden units set to 768 applied to the same contextualized embeddings. All the models $\mathscr{M}_{\mathscr{E}}$ are trained either in a mono-task or a multitask configuration. BERT is fine-tuned on our business dataset for 5 *epochs* using the Adam optimizer with an initial learning rate of 2^{-5} and a batch size of 16.

Our multilevel entity-informed models have been evaluated on the BizRel English test set[11] and compared to the best performing *Kag* and *Kin* state of the art models for RE, as follows.

– CNN^{Kag} (Zeng et al., 2014). This model is based on a convolutional neural network that uses FastText (Mikolov et al., 2018) pre-trained word embedding vectors of 300-dimension, three 1D convolutional layers, each one using 100 filters and a stride of 1, and different window sizes (3, 4 and 5 respectively) with a ReLU activation function. Each layer is followed by a max-pooling layer. The output layer is composed of a fully connected layer followed by a softmax classifier. The results reported here were obtained using a dropout of 50% and optimized using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-3}.

– Attention-BiLSTM^{Kag} (Zhou et al., 2016). It adopts a BiLSTM model with an attention mechanism that attends over all hidden states and generates attention coefficients relying on FastText embeddings as input representation. During experiments, best results have been obtained using 100 hidden units, an embedding dropout rate of 70%, a final layer dropout rate of 70%, and an Adam optimizer learning rate of 1.

– R-BERT^{Kag} (Wu and He, 2019). This is an adaptation of BERT for RE that takes into account entities representation in the relation instance representation. The model relies on the `bert-base-cased` model for English that is fine-tuned on our dataset for 5 *epochs*. R-BERT^{Kag} has been trained with the same hyper-parameters used to train our models.

– KnowBert^{Kin} (Peters et al., 2019). We also compare with KnowBert, one of the best *Kin* systems for RE.[12] KnwoBert comes up with three models either pre-trained with Wikipedia (**KnowBert-Wiki**), WordNet (**KnowBert-WordNet**), or with both resources (**KnowBert-W+W**). KnowBert-Wiki entity embeddings are learned using a skip-gram model

[11] All the hyperparameters were tuned on a validation set (10% of the train set).

[12] Among existing entity-informed models (cf. Section 5.1), at the time of performing these experiments, and as far as we know, only KnowBert and ERNIE were actually available to the research community. In this section, we compare with Knowbert as it achieved the best results on the TACRED dataset (71.50% on F1-score) when compared to ERNIE (67.97%) (Wang et al., 2020b).

directly from Wikipedia descriptions without using any explicit graph structure between nodes. Entity embeddings are then incorporated into BERT using knowledge-attention and re-contextualization mechanism. Embeddings in KnowBert-WordNet are learned from both Wordnet synset glosses and a knowledge graph constructed from word-word and lemma-lemma relations. KnowBert models are fine-tuned on our dataset for 5 *epochs* using the same hyperparameters proposed in the original paper.

– LUKE^{Kin} (Yamada et al., 2020b). This PLM treats words and entities in a given text as independent tokens, and outputs contextualized representations of them. It is trained using a new pre-training MLM-based task that involves predicting randomly masked words and entities in a large entity-annotated corpus retrieved from Wikipedia. It also extends the self-attention mechanism of the transformer by an entity-aware self-attention mechanism while considering the types of tokens (words or entities) when computing attention scores. It has been trained for 4 *epochs* with a learning rate value of 1^{-5}. LUKE was added **at the time of writing this dissertation (on Oct. 2022)**, as it achieved the best F1 on TACRED (72.7%) outperforming available *Kin* models.

5.5 Results and Discussions

5.5.1 Baseline Results

Table 5.3 presents the results of state of the art *Kag* and *Kin* baselines in terms of macro-averaged F-score (F1), precision (P), and recall (R); the best scores are in bold. [13] Among the four *Kag* models, R-BERT achieves the best scores. The results are however lower when compared to KnowBERT, which confirms that injecting knowledge about entities is crucial for effective RE. KnowBERT-Wiki being the best baseline in terms of F1-score, we, therefore, consider this model as a strong baseline to compare with.

5.5.2 Results

Due to the high number of $\mathscr{M}_{\mathscr{E}}$ configurations (3 combinations for $\mathscr{M}$ and 16 for $\mathscr{E}$, leading to a total of 48 different models), we only present the best performing ones. Table 5.4 summarizes our results. Best scores in our table are underlined while bold ones are those that outperform the best baseline. Due to space limitation and to better compare the contributions

[13] We also experimented with Entity-Attention-BiLSTM following (Lee et al., 2019) but the results were not conclusive.

Table 5.3 Results of Knowledge-agnostic (*Kag*) and knowledge-informed (*Kin*) baselines.

MODEL^{Kag}	P	R	F1
CNN (Zeng et al., 2014)	63.5	58.7	59.7
Att.-BiLSTM (Zhou et al., 2016)	59.4	54.3	56.3
R-BERT (Wu and He, 2019)	63.6	67.4	65.2
MODEL^{Kin}	P	R	F1
KnowBERT-Wiki (Peters et al., 2019)	65.3	71.9	68.2
KnowBERT-Wordnet (Peters et al., 2019)	63.6	71.5	67.0
KnowBERT-W+W (Peters et al., 2019)	64.2	72.7	67.5
LUKE^{Kin} (Yamada et al., 2020b)	**66.6**	**76.1**	**70.7**

of level of knowledge, we present the entity type and P-EE sources (*t*, *wiki*, *nas*) horizontally, and the attention one (*att*) vertically along with the classifier setting (monotaks vs. multitask).

Monotask setting. In the monotask configuration, we can observe that BizBERT results are better than BizBERT+CNN and BizBERT+ BILSTM and that the sentence features obtained via BizBERT+BILSTM is the least productive. From the observed results, two other interesting findings can be drawn. First, models with only one level of entity knowledge do not outperform the KnowBERT baseline nor LUKE baseline (e.g., $F1 = 67.7\%$ for $\text{BizBERT}_{\text{t}}$, $F1 = 66.7\%$ for $\text{BizBERT+CNN}_{\text{wiki}}$ and $F1 = 66.1\%$ for $\text{BizBERT}_{\text{nas}}$). Second, P-EE from NASARI are more productive than those from Wikipedia2Vec. See for example $\text{BizBERT}_{\text{wiki}} = 65.7\%$ vs. $\text{BizBERT}_{\text{nas}} = 66.1\%$ and $\text{BizBERT+BILSTM}_{\text{wiki}} = 65.9\%$ vs. $\text{BizBERT+BILSTM}_{\text{nas}} = 66.5\%$. This shows that even with NASARI low coverage rate when performing entity linking (83% vs. 92% for Wikipedia2vec), the relation classifier could capture important knowledge about entities and that P-EE built from multiple sources (BabelNet, WordNet synsets, Wikipedia pages) are of better quality than those built from Wikipedia alone. When multiple levels of knowledge are injected into the model, most results increase outperforming the KnowBERT baseline, and are comparable to the best performing baseline LUKE. In particular, combining P-EE with generalization over entity type has been very productive, achieving 1.9% in terms of F1-score over KnowBERT when using wiki+t with BizBERT, and is only 0.6% lower than the best baseline LUKE. $\text{BizBERT}_{\text{wiki+t}}$ also outperforms its single level counterparts (i.e., $\text{BizBERT}_{\text{t}}$ and $\text{BizBERT}_{\text{wiki}}$) by 2.4% and 4.4% respectively. We observe the same tendency when training the models with nas+t vs. nas and t alone. When relying on wiki+nas+t, the results are better than those obtained

Table 5.4 Results[‡] of the MONOTASK and MULTITASK experiments on English BIZREL.

	MONOTASK			MONOTASK$_{att}$			MULTITASK			MULTITASK$_{att}$		
MODEL	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BizBERT$_{wiki}$	64.3	67.9	65.7	63.2	70.8	66.6	65.5	70.5	67.6	63.8	71.9	67.4
BizBERT$_{wiki+t}$	**68.5**	71.9	70.1	**67.8**	73.9	70.6	**67.2**	70.9	68.9	**67.2**	71.2	69.1
BizBERT$_{nas}$	64.7	68.6	66.1	62.7	71.2	66.4	65.6	70.8	67.8	64.3	71.1	67.3
BizBERT$_{nas+t}$	**66.8**	70.6	68.5	**68.1**	72.5	70.1	**69.8**	69.8	69.7	**68.0**	71.7	69.7
BizBERT$_{nas+wiki+t}$	**67.9**	71.4	69.5	**67.8**	73.4	70.4	**68.1**	70.3	69.1	**67.5**	71.6	69.4
BizBERT+CNN$_{wiki}$	63.6	70.6	66.7	65.5	71.6	68.0	64.4	71.3	67.5	65.4	71.4	68.0
BizBERT+CNN$_{wiki+t}$	64.7	70.6	67.2	66.2	70.8	68.1	66.3	72.6	69.1	66.1	72.9	69.0
BizBERT+CNN$_{nas}$	61.6	71.3	65.6	64.9	71.2	67.7	63.1	71.8	66.9	**65.5**	72.1	68.4
BizBERT+CNN$_{nas+t}$	**68.1**	72.5	69.9	65.3	71.0	67.7	**68.1**	71.3	69.3	65.0	72.2	68.0
BizBERT+BILSTM$_{wiki}$	62.5	70.8	65.9	64.3	71.7	67.4	63.2	69.6	65.9	64.3	71.6	67.4
BizBERT+BILSTM$_{wiki+t}$	64.9	72.0	67.9	64.4	70.2	67.1	**67.6**	73.5	70.1	64.3	71.0	67.1
BizBERT+BILSTM$_{nas}$	63.3	71.0	66.5	64.1	71.2	67.0	64.3	68.2	65.8	64.3	71.1	67.1
BizBERT+BILSTM$_{nas+t}$	64.0	72.0	67.3	63.7	71.1	67.0	65.5	72.5	68.4	**67.0**	72.5	69.3

[‡]Best scores in our table are underlined while bold ones are those that outperform the best baseline.

for wiki+nas, but still lower when compared to wiki+t. This can be explained by the weak converge of NASARI for the entities present in the test set. Finally, when the knowledge-attention layer is activated, almost all the models gained in terms of F1 score, yielding to the highest improvement (about 2.4%) over KnowBERT for BizBERT$_{wiki+t+att}$, our best model, and (-0.1%) lower than LUKE. This demonstrates that knowledge-attention is an important mechanism for RE when coupled with other levels of knowledge about entities regardless of the aggregation layer used. Overall, these results show that directly injecting knowledge about entities as external features to the relation classifier without neither PLM re-training nor architecture update is a simple and effective solution for RE. More importantly, multiple levels of knowledge are needed, the best level being Wikipedia P-EE when coupled with entity type and knowledge-attention. Furthermore, when comparing the models in terms of the number of parameters to fine-tune, LUKE has 483M parameters, whereas BizBERT has 110M parameters (23% of LUKE's) and achieves comparable results to it.

Multitask setting. The results of the multitask setting show the same general conclusions already drawn from the monotask experiments: multilevel knowledge about entities is better than injecting a single level alone. However, we notice that BizBERT scores are lower when compared to the monotask configurations, while those of the BizBERT+CNN and BizBERT+BILSTM increased. Indeed, the BizBERT+BILSTM model with nas+t beats the KnowBERT baseline with the highest difference in this multitask configuration (1.9% in terms of F1-score), which is still lower than the best performing model (i.e., BizBERT$_{wiki+t+att}$

Table 5.5 F1-score per relation type for our best performing model and the best *Kin* baselines.

MODEL ↓/ REL. →	Inv.	Com.	Coo.	Leg.	Sal.	Oth.
$\text{BizBERT}_{\text{wiki+t+att}}$	67.9	77.6	**67.8**	**82.4**	41.3	86.6
KnwoBERT-Wiki	64.3	78.0	62.3	77.8	40.5	86.6
LUKE	**70.7**	**79.5**	67.2	73.7	**44.9**	**87.9**

in monotask setting). This shows that learning to classify business relations (monotask setting) is more effective than learning simultaneously both relation identification and relation classification (multitask setting). This implies that discriminating business from non-business relations is a much more complex task than discriminating between business relations, making the relation identification task harder. Two reasons behind that could be: (a) the dataset imbalance between business relation types and OTHERS relation type, and (b) the variability of relation patterns that could be included in the relation type OTHERS which make learning features about this class difficult.

5.5.3 Analysis

Results per relation type. The F-scores per relation type achieved by $\text{BizBERT}_{\text{wiki+t+att}}$, our best performing model, and the two best performing *Kin* baselines, KnowBERT-Wiki and LUKE, are presented in Table 5.5.

$\text{BizBERT}_{\text{wiki+t+att}}$ scores are: INVESTMENT 67.9%, SALE-PURCHASE 41.3%, COMPETITION 77.6%, COOPERATION 67.8%, and LEGAL PROCEEDINGS 82.4%. When compared to KnowBERT-Wiki, our model gets better scores for COOPERATION (+5.5%), INVESTMENT (+3.6%), SALE-PURCHASE (+0.8%), and LEGAL PROCEEDINGS (+4.6%) whereas it fails to account for COMPETITION (-0.4%). It is interesting to note that our model is more effective than the baseline when it comes to classifying relations with few instances. This observation is more visible in complex sentences that contain more than 4 entities. Whereas for the best performing baseline LUKE, our model can only score higher F1 for LEGAL PROCEEDINGS (+8.7%) and COOPERATION relations (+0.6%).

Confusion matrix analysis. A closer look at the confusion matrices shows that both models do not perform well when differentiating between business relations and non-business relations (OTHERS). The multitask setting we developed did not help to mitigate this, since it

Table 5.6 Relation Distribution per relation type and dataset type (train/test) in French BIZREL.

DATASET ↓ / REL. →	Invest.	Compet.	Cooperat.	Legal.	Sale.	Others
Train.	268	1,492	726	50	228	5,764
Test.	47	263	129	9	40	1,018

gave less effective results than the monotask one.[14] This is more salient for SALE-PURCHASE and COOPERATION where 38% and 19% of instances respectively were predicted as OTHERS by our model. This is because OTHERS instances do not have common characteristics like the five business relations we consider, as it may represent any other relation that may exist between two ORG (e.g., attending the same event, etc.) (More details about this R^- are presented in Chapter 4).

5.6 Portability to French Business Relations

We investigate the portability of our proposed knowledge-informed approach for RE on business relationship extraction from French content. For these experiments, we use the French BIZREL dataset (cf. Section 3.2.3, Chapter 3). The distributions of instances per relation type and dataset type are summarized in Table 5.6. We describe below the modifications required to both baselines and the proposed architecture for these experiments, as well as the results obtained.

5.6.1 Experimental Settings

– Baselines. We modify the previously described baseline models (see Section 5.4) for evaluation on the French corpus. The adaptation for neural-based models (CNN and Att-BiLSTM) consists in replacing English pre-trained embeddings with FastText pre-trained embeddings for French. Adapting the R-BERT^{Kag} consists in replacing the `bert-base-cased` encoder with three different PLMs that handle French: FlauBERT (Le et al., 2020), CamemBERT (Martin et al., 2019), and finally m-BERT, which supports 104 languages, including French.

In terms of *Kin* models, mLUKE (Ri et al., 2022) which is an extension of LUKE (Yamada et al., 2020b) is the only multilingual *Kin* that was pre-trained on an entity-related task. It handles 24 languages, including French. Adapting the other *Kin* models to French

[14] This confirms the results obtained by MT-RE$_{bin+all}$ in Table 4.6, Section 4.4, Chapter 4.

needs more computational resources to either retrain from scratch the whole model on french data, or part of it.

– **Multilevel entity informed architecture.** We also adapt the proposed architecture to evaluate it on the French BIZREL dataset. The model adaptation consists in using a French PLM rather than an English BERT to initialize the sentence encoder, and to use Wikipedia2vec pre-trained embeddings as a source of entity knowledge in the entity encoder because it is available for French, as opposed to NASARI, which is only available for English and Spanish. As BLINK entity linker is only available for English, we use handcrafted rules to link organization mentions to their Wikipedia IDs.

5.6.2 Results and Discussions

Baseline Results

Table 5.7 presents the results of the evaluated baselines. We observe that the models based on transformers outperform those based on neural architectures such as *CNN* or *RNN*. We also observe that R-mBERT relying on the multilingual BERT achieve better results than the ones using French transformers FlauBERT and CamemBERT in terms of F1-score and recall. R-FlauBERT however scores a better precision. The *Kin* model mLUKE achieves the best results, outperforming therefore all *Kag* models.

Table 5.7 Results of Knowledge-agnostic (*Kag*) and knowledge-informed (*Kin*) baselines on French BIZREL.

MODELS^{Kag}	P	R	F1
CNN (Zeng et al., 2014)	66.8	51.3	57.0
Att-Bi-LSTM (Zhou et al., 2016)	56.6	55.0	53.3
R-mBERT (Wu and He, 2019)	71.2	64.1	67.1
R-CamemBERT	74.6	53.8	59.5
R-FlauBERT	**77.2**	59.7	66.3
MODELS^{Kin}	P	R	F1
mLUKE (Ri et al., 2022)	70.8	**72.8**	**71.6**

Results of the Knowledge Enhanced models

Results obtained by the adaptation of the multilevel entity informed architecture to French are reported in Table 5.8. We report results obtained using mBERT, as it obtained better F1 than the two other French PLMs in baseline models (cf. Section 5.6.2).

Table 5.8 Results[‡] of the MONOTASK and MULTITASK experiments on French BIZREL.

	MONOTASK			MONOTASK$_{att}$			MULTITASK			MULTITASK$_{att}$		
MODEL	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Biz-mBERT$_{wiki}$	69.6	65.0	66.5	70.6	66.6	68.1	69.6	64.9	66.6	72.6	66.1	68.6
Biz-mBERT$_{wiki+t}$	68.4	67.5	67.5	70.1	68.4	69.1	69.9	65.3	66.4	68.1	66.2	66.7
Biz-mBERT+CNN$_{wiki}$	69.9	63.4	65.7	71.4	64.9	67.4	70.7	64.1	66.3	69.7	66.0	67.5
Biz-mBERT+CNN$_{wiki+t}$	68.1	64.0	65.6	66.7	66.7	66.5	70.0	64.5	66.1	69.3	66.5	67.4
Biz-mBERT+BILSTM$_{wiki}$	71.0	67.0	69.7	72.6	66.5	68.8	69.0	62.4	65.1	69.5	64.7	66.3
Biz-mBERT+BILSTM$_{wiki+t}$	71.9	70.3	70.5	73.2	68.1	69.9	67.7	64.4	65.4	70.2	68.2	68.9

[‡]Best scores in our table are underlined while bold ones are those that outperform the best baseline.

In contrast to experiments on English data, Biz-mBERT+BILSTM results outperform Biz-mBERT+CNN and Biz-mBERT, while the sentence features obtained through Biz-mBERT+CNN are the least productive.

Some of the findings from the English data could be generalized to the French data in terms of combining multiple levels of knowledge. In most cases, combining multiple levels of knowledge yields better results than one level alone. For example, Biz-mBERT$_{wiki+t}$, which combines P-EE with generalization over entity type, outperforms its variant that uses only P-EE (Biz-mBERT$_{wiki+t}$) (+1% F1). Same is the case for Biz-mBERT+BILSTM$_{wiki+t}$ that outperforms Biz-mBERT+BILSTM$_{wiki}$ (+0.1% F1). Moreover, knowledge-attention layer is as effective as for English data. When being activated, almost all the models gained in terms of F1 score. It is however less effective when the sentence encoder Biz-mBERT+BILSTM is used in a monotask setting. Overall, these results confirms that directly injecting multiple levels of knowledge about entities as external features to the relation classifier without neither PLM re-training nor architecture update is a simple and effective solution for RE.

Finally, we confirm that the combination of Wikipedia pre-trained embeddings with entity type and knowledge-attention forms the most productive sources of knowledge for RE when using the sentence encoder Biz-mBERT+BILSTM, scoring 1% shorter than the best performing baseline while being 30% of its size,[15] while no additional pre-training or modification pf the PLM architecture was required.

[15]mLuke model has 585M parameters, while our model is based on mBERT which has 177M parameters.

In contrast to English, combining pre-trained embeddings and entity type generalization in a multitask setup is less productive than P-EE alone, in both Biz-mBERT and Biz-mBERT+CNN. However, regardless of the sentence encoder used, knowledge-attention remains effective. When comparing the performances of monotask models and their multitask variants, we can notice that multitask setup has achieved better results for both Biz-mBERT and Biz-mBERT+CNN, while the monotask setup performed better for Biz-mBERT+BiLSTM.

Analysis

F1-score per relation type. The F-scores per relation type achieved by Biz-mBERT+ $\text{BILSTM}_{\text{wiki+t}}$, our best performing model, and the best performing *Kin* baseline mLUKE, are presented in Table 5.9. We can notice than mLUKE outperforms our best performing model for all-relation types except for the relation type LEGAL-PROCEEDINGS.

Table 5.9 F1-score per relation type for our best performing model and the best *Kin* baselines.

MODEL ↓/ REL. →	Inv.	Com.	Coo.	Leg.	Sal.	Oth.
Biz-mBERT+$\text{BILSTM}_{\text{wiki+t}}$	57.5	70.1	68.0	85.7	55.8	86.1
mLUKE	**62.6**	**72.8**	**68.2**	73.7	**80.0**	**87.2**

5.7 Conclusion

This chapter presented a simple but effective multilevel entity informed neural architectures to extract business relations from web documents. We conducted for the first time a systemic evaluation of the contribution of different levels of knowledge, experimenting with entity type generalization, pre-trained entity embeddings from Wikipedia2vec and NASARI, and entity-knowledge-attention both in a monotask and multitask settings.

Evaluating our model on both English and French BIZREL, our results show that multiple levels of knowledge are needed for effective RE, beating very competitive knowledge-agnostic, and achieving comparable results to strong knowledge-informed state of the art models. Our approach only requires entity knowledge as input alongside with the sentence representation provided by BERT pre-trained language model without any additional trained layer or parameters re-training. It is therefore generic and can be easily applied to extract other types of relations between named entities thanks to different sources of knowledge.

One way to improve our approach would be to create a single multilingual entity-informed model that uses the Wikipedia2vec resource, which is available in 14 languages, as a source of entity embeddings and leverages the cross-linguality power of multilingual pre-trained language models. This work has been published as a long paper in NLDB 2021 conference (Khaldi et al., 2021)

We end this dissertation by presenting how the models we proposed during this thesis have been integrated within the Geotrend platform, showing the effectiveness of our approach in real-time market intelligence application.

Chapter 6

Business Relation Extraction In the Geotrend Pipeline

In this chapter, we present the NLP components of Geotrend, a market intelligence platform that collects and processes open-web textual data (news articles, Wikipedia pages, etc.) to automatically extract market actors, technologies and business relations, and then makes this information accessible through a browsable interface to assist stakeholders in making strategic decisions.

This chapter is organized as follows: First, Section 6.1 presents the Geotrend platform, and highlighting how the RE component is integrated into it. Then, to assess its performances, a system demonstration is run on real-world Web pages with the goal of analyzing a specific market. The experimental results demonstrate that the platform can extract major business entities and relations from the Web for that given market. The extracted data is represented graphically in a graph of business interactions. This demonstrator is described in Section 6.2).[1] Finally, Section 6.3 presents some potential use cases of this platform, illustrating the advantages of such an NLP-based solution in real-world business applications.

6.1 The Geotrend Platform

The Geotrend platform is a complex platform comprised of many components ranging from authentication, information extraction to visualization. This chapter focuses on the IE pipeline, with RE being the most important, business-centric component.

The platform is made up of five key cloud-hosted components, which are depicted in Figure 6.1 and will be discussed in the following subsections. RE is the only component that

[1] A demo is available at: https://youtu.be/guoOCsvXRew.

uses a completely in-house built dataset (see Sections 3.1 and 3.2, Chapter 3) and models (cf. Chapters 4 and 5). The platform relies in addition on off-the-shelf components or slightly modified industrial ready models that have proven to be effective in practice.

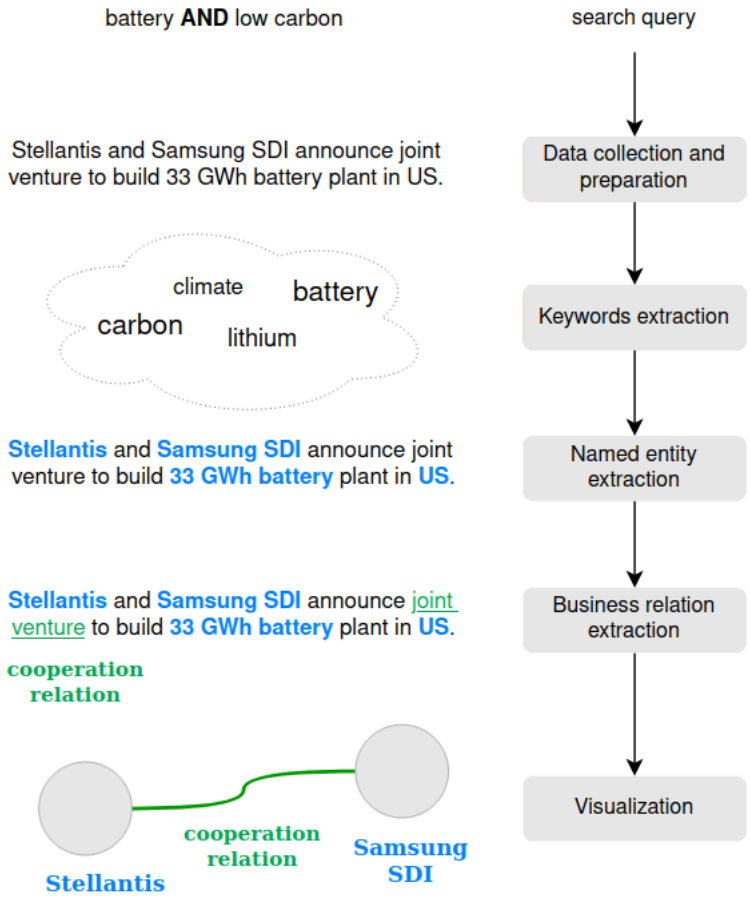

Figure 6.1 Geotrend Platform key components with illustrating examples of outputs per component.

6.1.1 Data Collection and Preparation

To retrieve the most accurate documents related to a specific company's market from the web, a combination of keywords and boolean operators is used. Numerous data sources, such as the open web indices (e.g., Google), user specified documents (upload of internal documents in various formats, or list of URLs), social media (e.g., Twitter), patent databases (e.g., Espacenet), and scientific publications databases (e.g., Scopus), can be queried for this purpose. The user can specify the language of the analyzed documents (English, French,

Spanish, Italian, German, Chinese, Russian, Arabic, Dutch, and Portuguese) and the data source to tailor their search.

The retrieved documents are cleaned and converted into a standardized format that the following IE components can analyze. Common text pre-processing operations are performed, including the removal of HTML tags, special characters, and extra white spaces, as well as the handling of variable encoding formats.

6.1.2 Keywords Extraction

Given the large number of documents to be analyzed, the ability to automatically identify the most relevant keywords representing these documents is critical. The TextRank algorithm (Mihalcea and Tarau, 2004) is used to extract the most important words or phrases in the collected textual documents. The algorithm is inspired from PageRank (Page et al., 1999), which is commonly used to calculate the importance of web pages. The textrank algorithm begins by constructing a word graph by observing which words are connected to one another. The graph's nodes are words or word sequences, and the edges represent the co-occurrence relationship between them. When two words follow one another, a link is formed between them; the link is given more weight if these two words appear together more frequently in the text. The Pagerank is a recursive algorithm applied to the resulting network to determine the importance of each word that is calculated by taking into account the incoming edges and the importance of the words from which these edges originate. The top one-third of these words are retained and deemed relevant.

In addition to PageRank, we also experimented with KeyBERT, an embedding-based solution that has been evaluated for keyword and key-phrase extraction (Sharma and Li, 2019). The method involves calculating cosine similarity between the embedding vectors of each N-gram word/phrase in a document and the embedding vector of the document itself. This is an expensive step given the length of the document to be analyzed. Despite the positive results, this method is not used in production due to efficiency concerns.

6.1.3 Named Entity Extraction

We use the publicly available models, spaCy[2] and Stanza,[3] to extract named entities from retrieved sentences. These models, based on Convolutional Neural Networks for spaCy and Recurrent Neural Networks for Stanza, have been trained on a variety of written materials,

[2]https://spacy.io/
[3]https://stanfordnlp.github.io/stanza/

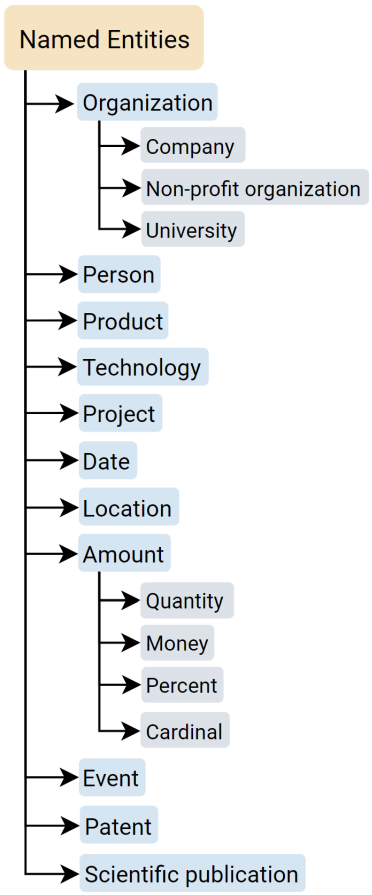

Figure 6.2 General taxonomy of named entities extracted by Geotrend platform.

including blogs, media, news, and comments (Honnibal et al., 2022; Qi et al., 2020). Depending on the language of the analyzed sentences and available models, and the purpose of the market study, various types of entities can be extracted. Figure 6.2 depicts the general taxonomy of these types. Moreover, a variety of techniques, including handcrafted regular expressions and the fine-tuning of existing models using additional internal annotations, are used to improve the extraction of certain entity types in some languages (e.g., product, University).

The named-entity extraction models mentioned above can detect locations in text. The Geonames database is used to convert them to a standard geographic format (city, country, etc.).[4] When a textual location can be mapped to multiple geographic locations, the one with the most people is chosen.

Date extraction is performed using the Heideltime temporal tagger (Strötgen and Gertz, 2013) to identify temporal expressions and convert them to standard TIMEX format. This

[4]https://www.geonames.org/

tagger can recognize both explicit (for example, 05/07/2020) and implicit (for example, next month) temporal expressions. Post-processing is used to improve the quality of the detected dates and remove some false positive extractions such as distance measurements (e.g., 2020 km detected as the year 2020), weight measurements (e.g., 2000 kg detected as the year 2000), or money measurements (e.g., 1999$ detected as the year 1999).

6.1.4 Business Relation Extraction

This module takes as input sentences containing at least two tagged target organizations as input and attempts to predict the type of business relationship that exists between them (it considers two entities at a time, per sentence).

It is based on a supervised model that was trained on an annotated dataset created in-house. The multilingual dataset presented in Chapter 3 accounting for four languages (English, French, Spanish, and Chinese) and 25k instances has been expanded by 11k examples in three new languages: German, Portuguese, and Arabic. The corpora generated with 36k manually annotated relation instances in seven languages is used to train a multilingual Transformer-based model based on Wu and He (2019) architecture in a joint learning setup.[5]

Target entities in the input sentence are identified using entity markers. Their generated representations are concatenated to form the relation instance, which is then passed to the classification layer for prediction. The model's performance in English, French, Spanish, and Chinese is reported in Section 3.3, Chapter 3. The contributions in Chapter 4 and 5 were partially integrated, with small and easy steps:

(1) Entities at the input level are replaced with their type in order to prevent over-fitting and for more generalization over the entity type.

(2) To tackle data imbalance, we use the Weighted Cross-Entropy to train the model;

To fully integrate the proposed models into the Geotrend IE pipeline, a thorough evaluation of model efficiency in terms of memory usage, latency, and inference time will be carried out as a future work of this thesis, for the purpose of resource consumption optimization.

6.1.5 Visualization and Collaboration

The above-mentioned components enable the extraction of valuable entities from a business standpoint (organizations, locations, products, persons, amounts, and so on), as well as the classification of numerous relationships between business organizations (investment,

[5]Which means training it on all instances from different languages at the same time.

cooperation, competition, and so on). The extracted relations and entities are stored in a database without duplication and visualized in faceted views by this component, allowing market stakeholders to analyze and explore them.

A *graph view* is created to map entities and relations from different documents into the same space, providing a snapshot of business interactions in a specific market. Each node in the graph represents an entity, and an edge connecting two entities represents their commercial interaction. A *sources view* enables the user to examine the data sources per link and determine the most frequently requested websites for this search. In addition, we create a *key figures view* to visualize the extracted amounts, as well as a *keywords view* to display a cloud of the most representative keywords in the analyzed documents. A *map view* is created to make it easier to explore the extracted locations. It combines the locations with the documents from which they were extracted.

Finally, users can refine their searches, browse data using a variety of filters that update displayed information in real time, and share their findings by granting access to their search to other users or exporting the results in a variety of formats (e.g., PNG, CSV). Examples of the obtained visualizations will be presented in the following section.

6.2 System Demonstration

To assess the Geotrend platform, we conduct a market analysis for "*zero-carbon and low-carbon batteries*". This is a new emerging technology that aids in the reduction of greenhouse gas emissions and plays an important role in climate change mitigation. The pipeline is tested on real web content, and the results are reviewed by domain experts.

To this end, the two sources Google News and Google are queried through their APIs to retrieve from the web 2,742 documents, between June 2001 and June 2022, using the following search query : *battery AND ("low carbon" OR "zero carbon" OR sustainable OR green)*.[6]

Quantitative results. Our platform extracted the following from these documents: 500 keywords whose top 10 are presented in Table 6.1, a total of 71,738 named entities, and approximately 41,245 business relations between entities of type *Organization*, which are presented per type in Table 6.2.

The most dominant keywords are the one used in the search query, which are *battery* and *carbon*. New keywords are extracted which are related to the search topic like *electric*,

[6]Using quotation marks for "low carbon" keyword, for example, will search for its exact occurrence in documents.

Table 6.1 Top 10 extracted keywords.

Keywords	**#Count**
battery	7,075
carbon	3,124
electric	2,274
lithium	2,237
storage	1,691
emission	1,325
solar	1,207
hydrogen	1,147
climate	875
production	728

Table 6.2 Extracted named entities and relations per type.

Entities	**#Count**	**Relations**	**#Count**
Organization	55,603	Cooperation	1,974
Person	14,065	Competition	822
Product	1,944	Investment	774
Technology	126	Sale-Purch.	594
Date	17,696	Legal-proc.	44
Location	19,903	Others	27,491
Amount	6,691		

lithium, *solar*, or *hydrogen*. Others reveal the actions taken in this domain such as *storage*, *emission* or *production*.

From Table 6.2, we can notice that the most extracted entity types are: *Organization*, *Location*, *Date*, *person*, followed by the rest of the types. The extracted relations are dominated by the negative relation *Others*, followed by *Cooperation*, *Competition*, *Investment*, *Sale-Purchase*, and *Legal-proceeding*. This is consistent with the relation types distribution in our training dataset (cf. Section 3.2.3, Chapter 3), and reflects the reality that there is not always a business interaction between two organizations mentioned in the same sentence.

Business Interactions Graph Insights. The extracted data is displayed in a graph, then analyzed by experts to generate meaningful insights about the search topic. Overall, the ecosystem surrounding zero-carbon and low-carbon batteries is dominated by car manufacturers, automotive suppliers, and battery manufacturers (cf. Figure 6.3).

Figure 6.4 shows that Tesla, the market leader in electric vehicles, dominates the market. It is followed by long-established automakers such as Volkswagen, Ford, BMW, Renault, and Toyota, which are gradually joining the race.

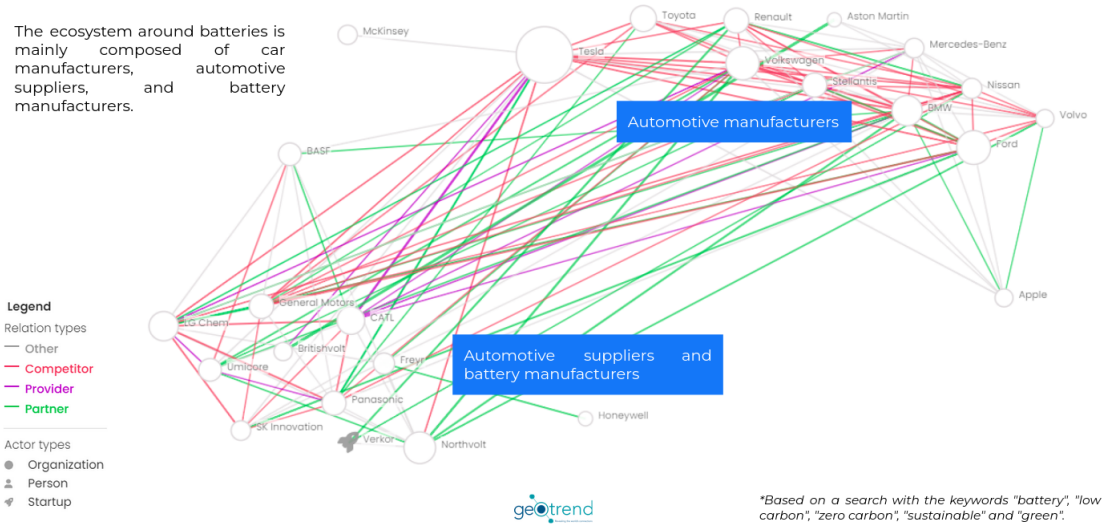

Figure 6.3 Graph view analysis for *zero-carbon and low-carbon batteries* ecosystem.

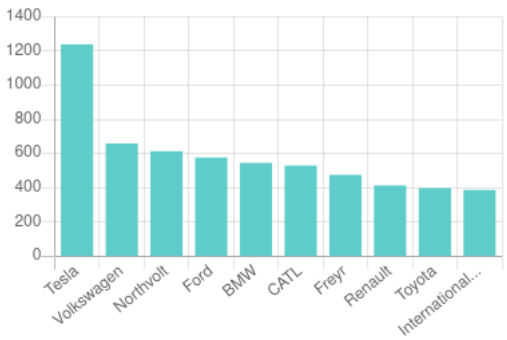

Figure 6.4 Top actors in terms of the number of business relations they participate in.

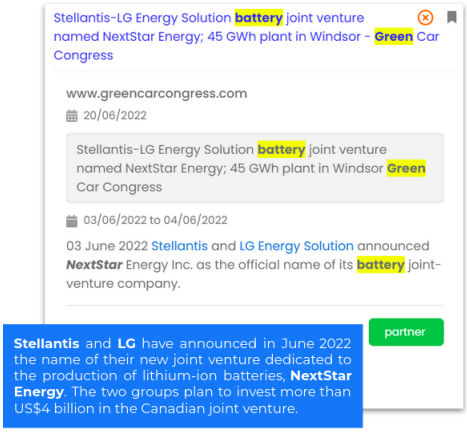

Figure 6.5 Relation analysis on a sentence level between the two actors *Stellantis* and *LG*.

A closer look at the business relation between the two actors *Stellantis* and *LG* reveals that the battery market is expanding, and significant investment is being made in the production of lithium-ion batteries, which bodes well for battery manufacturers (cf. Figure 6.5).

6.3 Possible Use Cases

Geotrend examines millions of data points from various sources and aggregates them into a single, multidimensional visualization to highlight economic interactions between market participants. This information extraction pipeline can benefit a variety of potential use cases and assist various business professionals in making pertinent and global decisions.

Among them users can **explore new ecosystems**. Indeed, to enter a new market or develop a novel technology, decision makers must have a thorough understanding of the various aspects of the new targeted ecosystem. The Geotrend platform can be used to analyze a specific market segment. The different views can be used to visualize key actors and their business interactions, identify strategic geographic areas, identify key projects and contracts, find new clients, partners, and investors, and finally detect market trends and technologies.

A second interesting use case is **competitive analysis**. In a rapidly changing business environment, companies must keep track of their most important competitors to anticipate their strategic moves. A user can conduct a search on our platform for a specific economic actor (competitor) and identify its strategic investments, partnerships, and service providers. Patent and scientific paper databases can also be requested via our platform to analyze competitors' R&D efforts and forecast their new product or project orientations.

Other use cases are also possible such as: technology analysis, risk analysis, customer knowledge, supply chain analysis, trends and innovation detection, and mergers and acquisitions opportunities assessment.

6.4 Conclusion

In this chapter, we introduced Geotrend, an NLP-based market intelligence platform that analyzes textual web content to extract and structure useful business information from it. The overall architecture is described, with emphasis on the business relation extraction component, which was the aim of this thesis.

Among the contributions presented in the previous chapters, only some of them have been partially integrated into the platform, namely *using a cost sensitive loss function* to handle data imbalance, and *performing a generalization over the entity type* in the extracted sentences before passing them to the relation extraction component. Indeed, their full integration into the pipeline in an industrial context requires extensive evaluations of their resource consumption in terms of memory and bandwidth, as well as their time of inference. This is envisioned as a future work to this thesis.

Conclusion

Main Contributions

This dissertation had five main objectives: (1) *Propose a new characterization of business relations* and *a new multilingual business relation dataset* manually annotated following this characterization, (2) *Develop models able to detect business relations between organizations in multilingual textual content*, (3) Propose solutions *to tackle data imbalance in our dataset*, (4) Investigate the impact of *injecting different sources of knowledge about target entities into a RE model*, and finally (5) Integrate these models in the Geotrend market intelligence platform.

To achieve these objectives, we first started by providing an overview of the state-of-the-art on generic and domain specific RE, with a particular focus on business relation extraction (cf. Chapters 1 and 2). This study has revealed three main shortcomings:

1. Despite previous efforts in the context of structuring business textual content, no prior work has addressed the extraction of business relations from multilingual texts. Previous work has all focused on a single language at a time. Generally, supervised models were used to perform this task casting it into a multi-class classification problem. However, the datasets used to train these models have three major flaws: **(a)** They are too small to train neural models that have proved to be very efficient for generic and other domain specific RE, **(b)** They are not always available to the research community, and **(c)** They are annotated using different annotation schemes, making comparison between different works difficult in this context.

2. When extracting domain-specific relations from the open web, a data imbalance problem arises between the limited set of targeted relations (positive) and the other relations not included in this set (negative relations).

3. Although several studies showed a significant improvement when injecting knowledge about target entities into transformer-based RE models, no one explored this approach

for business relation extraction. In addition, most of these studies considered one source of knowledge while requiring the model to be re-trained from scratch or the model's architecture to be modified. As far as we know, no one has empirically measured the impact of injecting multiple sources of knowledge about entities at different levels.

In this work, we bridge the gap by proposing solutions to each of the aforementioned shortcomings, as follows.

Firstly, we proposed the first multilingual dataset for extracting business relations. This dataset is unique in that it accounts for four languages: English, French, Spanish, and Chinese, and it is manually annotated using a unified characterization composed of five business relations between organizations: INVESTMENT, COOPERATION, COMPETITION, LEGAL-PROCEEDING, and SALE-PURCHASE. To account for other possible types of relations between organizations or the complete absence of any semantic link between them, the negative relation OTHERS is added. A pilot study was carried out to investigate the ability of multilingual pre-trained language models to extract these business relationships from multilingual contents. The results of various cross-lingual transfer configurations affirms that combining all languages during training was the best, beating all monolingual baselines (cf. Chapter 3).

Second, we proposed three approaches for learning RE from imbalanced data. The first is unsupervised and is based on sentence similarity to augment under-represented relation types with unlabeled data. The other two solutions aim at modifying the model architecture to learn more specific characteristics about positive and negative relationships: (a) a multitask learning architecture which allowed for the extraction of features of positive relations as well as the distinction between positive and negative relations, (b) a knowledge distillation architecture that enabled transferring knowledge of positive and negative relations from a binary classification model to a multi-class relation classification model. Overall, the three proposed models outperformed strong state-of-the-art models based on comparable approaches, with knowledge distillation being the most productive solution (cf. Chapter 4).

Finally, for the first time, we conducted an empirical study to investigate the impact of injecting multiple sources of knowledge about target entities into an RE model. The entity types, which are used to perform entity generalization at the input sentence, are the first source of knowledge. The second source of knowledge is entity embeddings trained on textual and structural data and used as extra features in the relation instance representation. The final source is obtained by employing a knowledge attention mechanism to identify the most important part of the sentence regarding the target entity's external knowledge. When evaluated on our dataset, the results show that incorporating these sources of knowledge achieve very good results on the English and French portion of BIZREL (cf. Chapter 5).

The findings of this thesis are intended to aid in improving the performances of Geotrend, a market intelligence platform performing business relation extraction. The RE component was initially based on manually constructed regular expressions created by domain experts to identify these business relations between organizations. Given the cost of writing these rules and adapting them to new languages, Geotrend decided to use supervised methods instead that handles content in different languages. Thus, the creation of a multilingual annotated corpus was a necessary step in this process.

Furthermore, the set of deep learning experiments conducted during this thesis will be extremely useful to Geotrend. Indeed, it can be used in the automatic construction and update of company knowledge graphs from multilingual content. The generated knowledge graph can serve in different direct business applications such as stock market prediction (Chen et al., 2019; Usmani and Shamsi, 2021), perceiving market trends and industries structures (Berns et al., 2021; Braun et al., 2018; Han et al., 2018a; Yamamoto et al., 2017), assisting investors decisions, risk analysis (Hogenboom et al., 2015; Liang et al., 2020; Yan et al., 2019), and competitive intelligence (Xu et al., 2011; Zhao et al., 2010). It can also be used in other NLP applications, such as the enrichment of a cross-domains KB such as Wikidata with up-to-date domain-specific knowledge (Waagmeester et al., 2021).

Future Directions

This work has several interesting future directions. We detail below some of them.

Towards hybrid RE models. Recently, the combination of neural networks and hand-crafted features has achieved impressive results on the ACE dataset (Chen et al., 2021) (more than 8% over state-of-the-art results), confirming that these methods are completely complementary, where the effort in neural network models is concentrated on designing network architecture that can learn productive features, whereas feature-engineering models aim to design these features manually while using prior knowledge and experience derived from previous work. As a result, investigating the integration of manually selected features with transformer-based models could be a promising research direction to explore.

Towards smaller knowledge enhanced RE models. Pretraining ever-larger language models on massive corpora and entity-knowledge related objectives has resulted in significant advances in NLP tasks, particularly for RE (Yamada et al., 2020b; Zhang et al., 2019). On the other hand, training and applying these large models requires massive amounts of computation, resulting in a significant carbon footprint and making them difficult to use for

both researchers and professionals (Schick and Schütze, 2021). To reduce the size of large models while maintaining the same level of performance, various approaches have been proposed, such as knowledge distillation (Sanh et al., 2019), quantization of models weights (Shen et al., 2020), or reducing the embedding layers size (Abdaoui et al., 2020; Mehta et al., 2019). However, the application of these methods to large knowledge enhanced models had not yet been investigated.

Another path to explore is **multilingual data augmentation for RE**. The scarcity of positive relations of interest is especially important when extracting relations from poorly annotated data, in a multilingual setting. Mixup (Guo et al., 2019a) is a data augmentation technique that interpolates input examples and labels linearly. When combined with transformers, it has demonstrated its effectiveness in many NLP classification tasks in either a monolingual (Chen et al., 2022a; Sun et al., 2020a) or multilingual setting (Yang et al., 2021). As far as we know, no prior work attempted to adapt it to the RE task.

Towards more domain-specific knowledge integration. Our multi-label relation classification models performed better when we fed them knowledge from structured encyclopedic resources in the form of pre-trained entity embeddings. It would be interesting to exploit the available knowledge in domain-specific resources such as stock market indices that report companies' stock price increases or decreases,[7][8] existing financial lexicon,[9] financial KB such as CrunchBase,[10] which covers over 100,000 companies, investors, acquisitions, and funding rounds, and finally the Financial Industry Business Ontology (FIBO),[11] which models financial concepts.

Last but not least, since we showed that the models proposed in this thesis are portable across domains (cf. Section 4.5, Chapter 4) and languages (cf. Section 5.6, Chapter 5), a new area of study to investigate would be **cross-domain relation extraction**. Recent efforts in this direction have recently been carried out by Bassignana and Plank (2022a) who proposed the first cross-domain dataset named CROSSRE, covering six diverse domains (news, politics, natural science, music, literature, AI) with annotations spanning 17 relation types. It would be therefore interesting to measure the possibility for transferring specific patterns for relations between entities across-domains.

[7]https://www.nasdaq.com/market-activity/stocks/screener
[8]https://markets.businessinsider.com/index/components/dow_jones
[9]https://markets.ft.com/glossary/searchLetter.asp?letter=A
[10]https://www.crunchbase.com/
[11]https://github.com/edmcouncil/fibo

Appendix A

Translated Introduction

Contexte et Motivations

Sur l'importance de structurer les relations d'affaires

L'économie du XXIe siècle a modifié la façon dont les acteurs du marché interagissent les uns avec les autres dans un marché mondial où les frontières nationales se sont dissoutes et où le commerce est devenu plus ouvert et plus libre (Hameed et al., 2021). La rivalité est passée du niveau du marché local au niveau multinational (Gorodnichenko et al., 2008). Cela incite les entreprises et les industries à renforcer leur capacité d'innovation pour fournir des produits et des services compétitifs, et accroître leur croissance et leurs performances économiques (Hameed et al., 2021; Passaris, 2006).

Dans un environnement commercial complexe et en évolution rapide, l'intelligence compétitive (IC) désigne le processus de collecte, d'analyse et de partage d'informations sur l'environnement économique telles que les capacités et les intentions des concurrents, puis de les transformer en connaissances qui peuvent être utilisées par les entreprises pour la prise de décision (Gilad and Gilad, 1986; Kahaner, 1997; Montgomery and Weinberg, 1979; Oberlechner and Hocking, 2004). En raison de l'énorme quantité d'informations publiques partagées et diffusées chaque jour sur Internet, les contenus web non structurés sont devenus une source cruciale d'IC, rendant leur exploitation manuelle peu pratique (Boncella, 2003). L'extraction automatique d'informations commerciales est donc un outil précieux pour identifier les liens entre des acteurs spécifiques du marché et construire des réseaux commerciaux.

Une façon possible de structurer les relations commerciales et de faciliter la génération de réseaux d'entreprises est d'organiser le contenu textuel en graphes de connaissances financières, où les nœuds sont des entités financières et commerciales et les arêtes reliant

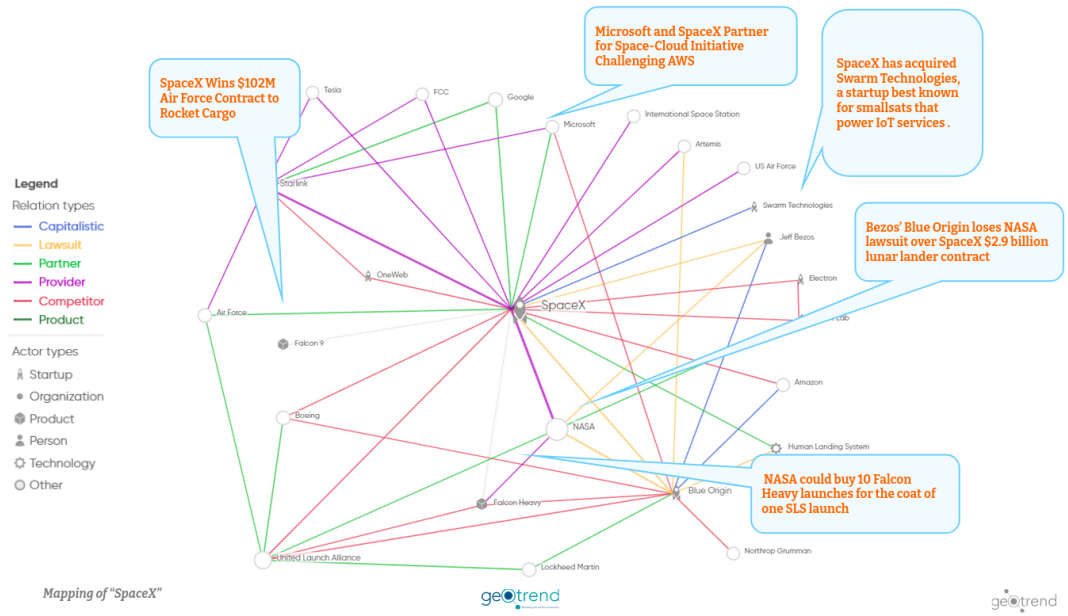

Figure A.1 Un exemple de graphe de connaissances sur SpaceX, tel que fourni par la plateforme Geotrend.

ces entités représentent les interactions commerciales entre elles. La figure A.1 illustre un tel graphe de connaissances tel que fourni par la plateforme Geotrend[1]. Geotrend est une PME française, qui a développé une plateforme de "Market Intelligence" visant à soutenir la découverte, l'analyse et le suivi de tout marché en temps réel. Cette plateforme est basée sur des composants d'extraction d'informations qui extraient du web les principaux acteurs du marché, les relations qui existent entre eux (par exemple, partenariat, concurrence, filiale, etc.), les valeurs monétaires qui caractérisent le marché, ainsi que les dates et lieux importants qui y sont liés. Dans cette figure, le graphe a été créé en analysant des centaines de documents récupérés sur le Web à propos de l'entreprise *SpaceX*. Ensuite, les relations commerciales sont extraites à partir de phrases à l'aide d'un composant d'extraction de relations (ER). Par exemple, à partir de la phrase dans (1), la plateforme peut identifier les entités *SpaceX*, et *Swarm Technologies*, et la relation commerciale entre elles (`acheté_par`).

(1) SpaceX a acquis Swarm Technologies, une startup surtout connue pour ses petits satellites qui alimentent des services IoT.

Les concurrents de SpaceX peuvent utiliser le graphique généré pour détecter les menaces ou opportunités potentielles basées sur les activités de l'entreprise, ce qui leur permet d'ajuster leurs stratégies pour prospérer et rester compétitifs sur le marché (Sewlal, 2004). Ce

[1]https://www.Geotrend.fr/fr/

réseau d'affaires peut également être utilisé par les banques et les investisseurs pour analyser les relations d'affaires de leurs clients et des entreprises dans lesquelles ils investissent, afin d'évaluer les risques d'un prêt ou d'un investissement, et donc de maximiser les gains éventuels tout en minimisant les pertes potentielles (Yan et al., 2019; Zuo et al., 2017).

En 2019, le composant Geotrend ER était initialement basé sur des expressions régulières conçues manuellement et créées par des experts du domaine. En revanche, l'écriture de ces règles et leur adaptation à de nouveaux langages sont coûteuses. En outre, ces relations peuvent être exprimées de manière indirecte ou métaphorique dans le texte, ce qui rend l'extraction basée sur des règles encore plus difficile. C'est ce qu'illustre la phrase dans (2) qui exprime une relation `en_compétition_avec` entre les entités *Delphi Automotive* et *Volkswagen* en utilisant l'expression *a délivré des permis d'essai de véhicules autonomes,*, ce qui implique qu'ils sont tous des fabricants de véhicules autonomes.

(2) Wheego et Valeo rejoignent désormais Google, Tesla, GM Cruise et Ford sur la liste des entreprises auxquelles le DMV californien a délivré des permis d'essai de véhicules autonomes, ainsi que Volkswagen, Mercedes Benz, Delphi Automotive et Bosch.

Malgré leur importance stratégique, l'extraction de relations commerciales a reçu moins d'attention dans la littérature par rapport à d'autres domaines spécifiques, tels que les domaines biomédical (Bunescu et al., 2005; Krallinger et al., 2017; Lee et al., 2013; Segura-Bedmar et al., 2013; Van Mulligen et al., 2012; Wu et al., 2019) et scientifique (Bruches et al., 2020; Buscaldi et al., 2018; Luan et al., 2018a; Ma et al., 2022). Zhao et al. (2010) a présenté le premier travail dans cette direction visant à identifier la taxonomie des relations commerciales à extraire entre les entreprises, les personnes, les dates, les lieux, etc. et son application dans le contexte de l'IC. Peu d'articles ont été publiés dans les années qui ont suivi, avec une orientation industrielle typique, où l'ensemble des relations ciblées est principalement limité aux interactions de concurrence et de coopération entre les organisations (Lau and Zhang, 2011; Yamamoto et al., 2017). Récemment, divers ateliers de recherche ont été consacrés à l'analyse des informations financières sur le web, encourageant ainsi les avancées dans ce domaine (par exemple, Financial Technology and Natural Language Processing Workshop, [2]. Atelier sur la découverte de connaissances à partir de données non structurées dans les services financiers, et atelier sur le traitement des récits financiers [3]).

[2]https://aclanthology.org/venues/finnlp/
[3]https://aclanthology.org/venues/fnp/

L'objectif de cette thèse, réalisée dans le cadre d'un contrat CIFRE entre le laboratoire IRIT de Toulouse et Geotrend, est de proposer de nouvelles approches d'apprentissage supervisé basées sur des architectures modernes d'apprentissage profond pour détecter des relations commerciales dans un contexte multilingue, améliorant ainsi le composant initial Geotrend ER à base de règles.

L'extraction de relations économiques en tant que tâche du TAL

Jusqu'à présent, diverses approches ont été proposées dans la littérature pour extraire les relations entre les entités. Les premières, *basés sur des règles*, reposaient sur la définition manuelle de patrons qui identifient le type de relations sémantiques entre les entités dans le texte sur la base de divers patrons linguistiques lexico-syntaxiques utilisés pour exprimer un type de relation donné (Akbik and Broß, 2009; Aussenac-Gilles and Jacques, 2008; Batista et al., 2015; Hearst, 1992; Snow et al., 2004; Suchanek et al., 2006). Cependant, cette approche a un rappel plus faible et nécessite une expertise humaine pour créer ces patrons ainsi que pour les adapter à de nouveaux domaines.

Pour surmonter ces limites, des *approches supervisées* basées sur des algorithmes d'apprentissage automatique ont été proposées, principalement en raison du volume croissant de corpus textuels (notamment sur le web) qui peuvent être utilisés comme données d'entraînement après avoir été annotés par des experts du domaine. Les méthodes *basés sur les traits* (Kambhatla, 2004; Nguyen et al., 2007b; Zhou et al., 2005) et *à base de noyaux*. (Collins and Duffy, 2001; Culotta and Sorensen, 2004; Mooney and Bunescu, 2006) sont parmi les premières approches où des caractéristiques lexicales, sémantiques et syntaxiques dédiées représentant une phrase du corpus d'apprentissage sont conçues manuellement puis introduites dans un algorithme de classification qui apprend à prédire le type de relation reliant deux entités préalablement identifiées. Bien que ces approches soient nettement plus efficaces, le choix de l'ensemble sous-optimal de caractéristiques représentatives n'est pas une tâche facile. De plus, l'annotation des données d'apprentissage représente un coût élevé chaque fois qu'un nouveau domaine est abordé.

Pour réduire la phase fastidieuse de l'identification des caractéristiques les plus pertinentes, des modèles neuronaux (en particulier les réseaux neuronaux convolutionnels (CNN) et les réseaux neuronaux récurrents (RNN)) ont été proposés pour automatiser l'extraction des caractéristiques. Dans ces modèles, les phrases sont représentées par des vecteurs sémantiques statiques appelés *word embeddings*, qui sont calculés sur de grands corpus pour apprendre les représentations des mots à partir de leurs différents contextes (Mikolov et al., 2013b; Pennington et al., 2014). Les architectures neuronales sont entraînées sur ces vecteurs pour extraire automatiquement des caractéristiques et prédire le type de relation exprimé dans

la phrase d'entrée. Récemment, les architectures *transformer* basées sur des mécanismes d'auto-attention à têtes multiples ont permis d'obtenir des représentations de mots contextualisées générées par des modèles de langage pré-entraînés sur des corpus textuels à grande échelle (Devlin et al., 2019), atteignant des scores maximaux sur des jeux de données d'ER très connus (Tao et al., 2019; Wu and He, 2019). Ces architectures (et leurs performances associées) ont été améliorées en injectant des connaissances sur les entités cibles fournies par des ressources linguistiques externes et des pré-traitements supplémentaires (Wang et al., 2019; Yamada et al., 2020b; Zhang et al., 2019).

Dans cette étude, nous avons expérimenté plusieurs nouvelles architectures basées sur des transformateurs tout en évaluant leurs performances sur un nouveau jeu de données multilingues pour l'extraction de relations commerciales.

Pourquoi l'extraction de relations d'affaires est-elle une tâche difficile?

Dans le cas des relations d'affaires, la plupart des méthodes existantes reposent sur des modèles générés manuellement ou automatiquement qui sont difficiles à maintenir (Braun et al., 2018; Burdick et al., 2015; Lau and Zhang, 2011). Des approches supervisées ont été récemment proposées (Collovini et al., 2020; De Los Reyes et al., 2021; Yamamoto et al., 2017; Yan et al., 2019) mais les ensembles de données utilisés dans ces études présentent les lacunes suivantes :

- Ils se concentrent tous sur une seule langue. Or, en raison du manque de modèles multilingues, une grande quantité d'informations textuelles commerciales et financières générées en ligne dans diverses langues est difficile à exploiter automatiquement par les professionnels;
- Ils sont peu nombreux ou pas toujours disponibles pour la communauté des chercheurs;
- Elles sont annotées à l'aide de divers schémas d'annotation, ce qui rend difficile la comparaison de divers travaux réalisés par différents chercheurs dans ce contexte.

En outre, comme les approches supervisées disposent d'un ensemble limité de relations ciblées, les modèles qui extraient des relations du web ouvert souffrent d'une pénurie de relations positives d'intérêt. Dans le contexte des relations commerciales, par exemple, deux entreprises mentionnées dans une même phrase ne sont pas nécessairement liées par une relation commerciale sémantique, comme le montre l'exemple suivant : (3). Cette phrase exprime une relation négative (c'est-à-dire, `None`) entre les deux entreprises, *Intel* et *Tesla*.

En même temps, la relation `acquired_by` existe entre *Intel* et *Mobileye*, et `partner_with` entre *Tesla* et *Mobileye*. Cela pose un problème de déséquilibre des données, qui entrave l'entrainement des modèles d'apprentissage automatique.

(3) Mobileye a été rachetée par Intel en 2017 pour 15,3 milliards de dollars américains. Cette société israélienne de vision était également un partenaire de Tesla, pour disposer de la première génération d'Autopilot.

Dans cette thèse, nous visons à combler le fossé en proposant des solutions à chacune des lacunes susmentionnées.

Questions de recherche et contributions

Pour explorer l'extraction des relations d'affaires, notre recherche peut être formulée sous la forme des questions de recherche (QR) suivantes.

(*RQ*1) Comment les relations commerciales sont-elles caractérisées et annotées dans un contenu textuel multilingue ?

(*RQ*2) L'entraînement d'un modèle ER unique sur des données multilingues peut-il être plus performant que l'entraînement de plusieurs modèles uniques sur des données monolingues ?

(*RQ*3) Comment un modèle ER peut-il gérer le déséquilibre des données entre les relations commerciales et non commerciales ?

(*RQ*4) L'injection de connaissances factuelles sur les entités dans un modèle ER à différents niveaux de granularité peut-elle améliorer ses performances ?

(*RQ*5) Comment les résultats de notre recherche peuvent-ils être utilisés dans une application d'intelligence économique en temps réel ?

Sur la base des questions de recherche ci-dessus, les principales contributions (C) de cette thèse sont résumées comme suit :

(C1) Une caractérisation unifiée des relations commerciales entre *Organisations*, basée sur une taxonomie composée de cinq relations, à savoir : INVESTISSEMENT, COOPÉRATION, VENTE-ACHAT, CONCURRENCE, PROCÉDURE JUDICIAIRE, et une relation négative AUTRES qui regroupe les autres types de relations non ciblées ;

(C2) Un jeu de données d'ER d'entreprise multilingue annoté manuellement à l'aide de cette caractérisation en quatre langues : *Français*, *Espagnol*, *Anglais*, et *chinois*. Une partie de l'ensemble de données est disponible pour la communauté des chercheurs.[4] Pour autant que nous le sachions, il s'agit du premier ensemble de données multilingues dans ce domaine, car tous les ensembles de données proposés précédemment se concentraient sur une seule langue à la fois (Khaldi et al., 2022c).

(C3) Un ensemble de modèles de type *tranformer* pour réaliser la tache de ER à partir de textes multilingues, reposant sur des modèles de langage pré-entraînés monolingues et multilingues (Khaldi et al., 2022c).

(C4) Évaluation empirique de diverses approches pour résoudre le problème du déséquilibre des données entre les relations commerciales et non commerciales (c'est-à-dire négatives), en apportant des améliorations aux données d'apprentissage ou aux modèles d'ER. (Khaldi et al., 2022a,b). Nous étudions en particulier trois nouvelles solutions : l'augmentation des données à l'aide de la similarité des phrases, l'extraction multitâche des relations et les étiquettes binaires générées par la distillation des connaissances pour superviser l'ER.

(C5) Une évaluation empirique de l'impact de l'intégration de différentes sources de connaissances sur les entités dans le modèle d'ER, à différents niveaux de granularité. (Khaldi et al., 2020, 2021), allant au-delà des études récentes qui se sont concentrées sur une seule source de connaissances (Papaluca et al., 2022; Poerner et al., 2020; Zhang et al., 2019).

(C6) L'intégration de nos modèles dans la plateforme Geotrend Market Intelligence montrant qu'une extraction multilingue de relations commerciales informée par les entités qui gère le déséquilibre des données est cruciale dans un contexte industriel.

Plan de dissertation

La thèse est organisée en six chapitres. Les deux premiers présentent une vue d'ensemble de l'état de l'art en matière d'extraction de relations binaires, tandis que les quatre autres se concentrent sur l'une des contributions susmentionnées.

Dans le chapitre 1, nous présentons la tâche de *extraction de relations génériques* au niveau de la phrase. Nous commençons par définir les principaux concepts liés et le pipeline global de l'ER. Nous présentons ensuite les principaux jeux de données existants annotés

[4] https://github.com/Geotrend-research/business-relation-dataset

manuellement pour l'ER, ainsi qu'une description de leur construction et une caractérisation quantitative de leurs instances de relations annotées. Enfin, nous nous concentrons sur les approches supervisées pour l'ER, avec un accent particulier sur les approches neuronales et celles basées sur les transformateurs, qui servent de base à notre travail.

Le chapitre 2 poursuit l'état de l'art, en se concentrant cette fois sur trois *relations spécifiques à un domaine*, à savoir les domaines biomédical, scientifique et commercial, qui ont suscité un grand intérêt dans la communauté des chercheurs. Pour chaque domaine, nous passons en revue les principaux jeux de données proposés et les approches entraînées sur ces jeux de données, ainsi que certaines applications qui exploitent les relations extraites dans différents systèmes déployés. Nous terminons ce chapitre en soulignant les principales contributions de ce travail.

Le chapitre 3 détaille le processus de collecte de données que nous avons suivi et caractérise la typologie des relations commerciales qui a été utilisée pour annoter notre jeu de données multilingue (c'est-à-dire *(RQ1)* et *(RQ2)*). Nous présentons ensuite les expériences réalisées pour détecter les relations d'affaires à partir de contenus multilingues avec différents paramètres de transfert interlinguistique, allant d'un transfert à zéro-données à un transfert conjoint.

Le chapitre 4 aborde la question du déséquilibre des données dans les modèles ER (c'est-à-dire *(RQ3)*), en particulier le déséquilibre entre la relation négative et les relations positives d'intérêt. Nous commençons par donner une vue d'ensemble des approches existantes au niveau des données et des modèles pour le déséquilibre des données dans le langage naturel. Nous présentons ensuite les trois approches que nous venons de proposer, à savoir : l'augmentation des données basée sur la similarité des phrases, l'identification et la classification multitâches des relations pour améliorer leur extraction, et enfin l'utilisation d'étiquettes binaires générées par la distillation des connaissances pour superviser l'extraction des relations.

Nous tentons de répondre à notre quatrième question de recherche *(RQ4)* dans le chapitre 5. Nous commençons par donner une vue d'ensemble des modèles de langage pré-entraînés améliorés par de la connaissance, en soulignant leurs principales caractéristiques. Nous présentons ensuite l'architecture que nous proposons et qui injecte plusieurs sources de connaissances sur les entités cibles dans un modèle ER à plusieurs niveaux. Les expériences menées pour étudier l'importance de chaque niveau de connaissance sont ensuite présentées.

Dans le chapitre 6, nous décrivons la plateforme Geotrend, un pipeline industriel d'intelligence économique pour l'extraction de relations commerciales. Nous présentons l'architecture globale de ce pipeline et détaillons comment les modèles proposés dans cette

thèse y ont été intégrés (c'est-à-dire *(RQ5)*).

Enfin, nous concluons en donnant une vue d'ensemble de ce travail, en soulignant ses contributions et ses limites. Nous soulignons également nos perspectives pour les travaux futurs.

Appendix B

Translated Conclusion

Contributions principales

Cette thèse avait cinq objectifs principaux : (1) *Proposer une nouvelle caractérisation des relations d'affaires* et *un nouveau jeu de données multilingues de relations d'affaires* annoté manuellement selon cette caractérisation, (2) *Développer des modèles capables de détecter les relations d'affaires entre organisations dans un contenu textuel multilingue*, (3) Proposer des solutions *pour remédier au déséquilibre des données dans notre jeu de données*, (4) Etudier l'impact de *l'injection de différentes sources de connaissances sur les entités cibles dans un modèle RE*, et enfin (5) Intégrer ces modèles dans la plateforme d'intelligence économique Geotrend.

Pour atteindre ces objectifs, nous avons commencé par fournir un aperçu de l'état de l'art sur l'ER générique et spécifique à un domaine, avec un accent particulier sur l'extraction de relations commerciales (cf. chapitres 1 et 2). Cette étude a révélé trois principales lacunes :

1. Malgré les efforts précédents dans le contexte de la structuration du contenu textuel commercial, aucun travail antérieur n'a abordé l'extraction des relations commerciales à partir de textes multilingues. Les travaux précédents se sont tous concentrés sur une seule langue à la fois. Généralement, des modèles supervisés ont été utilisés pour effectuer cette tâche, la transformant en un problème de classification multi-classes. Cependant, les jeux de données utilisés pour entraîner ces modèles présentent trois défauts majeurs : **(a)** Ils sont trop petits pour entraîner des modèles neuronaux qui se sont avérés très efficaces pour l'ER génériques et d'autres domaines spécifiques, **(b)** Ils ne sont pas toujours disponibles pour la communauté des chercheurs, et **(c)**

Ils sont annotés à l'aide de différents schémas d'annotation, ce qui rend difficile la comparaison entre les différents travaux dans ce contexte.

2. Lors de l'extraction de relations de domaines spécifiques à partir du web ouvert, un problème de déséquilibre des données se pose entre l'ensemble limité de relations ciblées (positives) et les autres relations non incluses dans cet ensemble (relations négatives).

3. Bien que plusieurs études aient montré une amélioration significative lors de l'injection de connaissances sur les entités cibles dans les modèles ER basés sur les transformateurs, aucune étude n'a exploré cette approche pour l'extraction de relations commerciales. De plus, la plupart de ces études n'ont considéré qu'une seule source de connaissances tout en exigeant que le modèle soit ré-entraîné à partir de zéro ou que l'architecture du modèle soit modifiée. À notre connaissance, personne n'a mesuré empiriquement l'impact de l'injection de sources multiples de connaissances sur les entités à différents niveaux.

Dans ce travail, nous comblons cette lacune en proposant des solutions à chacune des lacunes susmentionnées, comme suit.

Tout d'abord, nous avons proposé le premier ensemble de données multilingues pour l'extraction de relations commerciales. Ce jeu de données est unique dans la mesure où il prend en compte quatre langues : L'anglais, le français, l'espagnol et le chinois, et il est annoté manuellement en utilisant une caractérisation unifiée composée de cinq relations d'affaires entre organisations : INVESTISSEMENT, COOPÉRATION, CONCURRENCE, PROCÉDURE LÉGALE, et VENTE-ACHAT. Pour tenir compte d'autres types de relations possibles entre les organisations ou de l'absence totale de tout lien sémantique entre elles, la relation négative AUTRES est ajoutée. Une étude pilote a été réalisée afin d'examiner la capacité des modèles linguistiques multilingues pré-entraînés à extraire ces relations commerciales à partir de contenus multilingues. Les résultats de diverses configurations de transfert interlinguistique affirment que la combinaison de toutes les langues pendant l'entraînement était la meilleure, battant toutes les modèles de base monolingues (cf. chapitre 3).

Deuxièmement, nous avons proposé trois approches pour l'apprentissage des ER à partir de données déséquilibrées. La première est non supervisée et se base sur la similarité des phrases pour augmenter les types de relations sous-représentés avec des données non étiquetées. Les deux autres solutions visent à modifier l'architecture du modèle pour apprendre des caractéristiques plus spécifiques des relations positives et négatives : (a) une architecture d'apprentissage multitâche qui a permis l'extraction des caractéristiques des relations positives ainsi que la distinction entre les relations positives et négatives, (b) une architecture

de distillation des connaissances qui a permis de transférer les connaissances des relations positives et négatives d'un modèle de classification binaire à un modèle de classification des relations multi-classes. Dans l'ensemble, les trois modèles proposés ont obtenu de meilleures performances que les modèles de pointe basés sur des approches comparables, la distillation des connaissances étant la solution la plus productive (cf. chapitre 4).

Enfin, pour la première fois, nous avons mené une étude empirique pour examiner l'impact de l'injection de plusieurs sources de connaissances sur les entités cibles dans un modèle d'ER. Les types d'entités, qui sont utilisés pour effectuer la généralisation des entités à la phrase d'entrée, constituent la première source de connaissances. La deuxième source de connaissances est constituée par les incorporations des representations des entités entrainées sur des données textuelles et structurelles et utilisées comme caractéristiques supplémentaires dans la représentation de l'instance de relation. La dernière source est obtenue en employant un mécanisme d'attention aux connaissances pour identifier la partie la plus importante de la phrase concernant les connaissances externes de l'entité cible. Une fois évalués sur notre jeu de données, les résultats montrent que l'incorporation de ces sources de connaissances permet d'obtenir de très bons résultats sur les partie anglaise et française de BIZREL. (cf. chapitre 5).

Les résultats de cette thèse ont pour but d'aider à améliorer les performances de Geotrend, une plateforme d'intelligence économique réalisant l'extraction de relations commerciales. Le composant ER était initialement basé sur des expressions régulières construites manuellement et créées par des experts du domaine pour identifier ces relations d'affaires entre organisations. Compte tenu du coût de l'écriture de ces règles et de leur adaptation à de nouvelles langues, Geotrend a décidé d'utiliser plutôt des méthodes supervisées qui traitent le contenu dans différentes langues. Ainsi, la création d'un corpus annoté multilingue était une étape nécessaire dans ce processus.

En outre, l'ensemble des expériences d'apprentissage profond menées au cours de cette thèse sera extrêmement utile à Geotrend. En effet, il peut être utilisé dans la construction et la mise à jour automatique de graphes de connaissances d'entreprise à partir de contenus multilingues. Le graphe de connaissances généré peut servir dans différentes applications commerciales directes telles que la prédiction du marché boursier (Chen et al., 2019; Usmani and Shamsi, 2021), la perception des tendances du marché et des structures des industries (Berns et al., 2021; Braun et al., 2018; Han et al., 2018a; Yamamoto et al., 2017), l'aide aux décisions des investisseurs, l'analyse des risques (Hogenboom et al., 2015; Liang et al., 2020; Yan et al., 2019), et la veille concurrentielle (Xu et al., 2011; Zhao et al., 2010). Elle peut également être utilisée dans d'autres applications de traitement automatique des langues,

comme l'enrichissement d'une base de données interdomaines telle que Wikidata avec des connaissances actualisées spécifiques à un domaine (Waagmeester et al., 2021).

Future Directions

Ce travail présente plusieurs directions futures intéressantes. Nous en détaillons quelques-unes ci-dessous.

Vers des modèles hybrides de RE. Récemment, la combinaison de réseaux neuronaux et les modèles à base de trats créées à la main a permis d'obtenir des résultats impressionnants sur le jeu de données ACE (Chen et al., 2021) (plus de 8 % par rapport aux résultats de l'état de l'art), ce qui confirme que ces méthodes sont totalement complémentaires, l'effort des modèles de réseaux neuronaux étant concentré sur la conception d'une architecture de réseau capable d'apprendre des caractéristiques productives, tandis que les modèles d'ingénierie des caractéristiques visent à concevoir ces caractéristiques manuellement tout en utilisant les connaissances et l'expérience antérieures tirées de travaux antérieurs. Par conséquent, l'étude de l'intégration de caractéristiques sélectionnées manuellement dans les modèles basés sur les transformateurs pourrait être une direction de recherche prometteuse à explorer.

Vers des modèles plus petits améliorés par des connaissances. D'une part, le pré-entraînement de modèles de langage de plus en plus grands sur des corpus massifs et des objectifs liés à la connaissance des entités a permis des avancées significatives dans les tâches de TAL, en particulier pour l'ER (Yamada et al., 2020b; Zhang et al., 2019). D'autre part, l'entraînement et l'application de ces grands modèles nécessitent des quantités massives de ressources de calcul, ce qui entraîne une empreinte carbone, et des couts financiers importants et rend leur utilisation difficile pour les chercheurs et les professionnels (Schick and Schütze, 2021). Pour réduire la taille des grands modèles tout en conservant le même niveau de performance, diverses approches ont été proposées, comme la distillation des connaissances (Sanh et al., 2019), la quantification des poids des modèles (Shen et al., 2020), ou la réduction de la taille des couches des embeddings (Abdaoui et al., 2020; Mehta et al., 2019). Cependant, l'application de ces méthodes à de grands modèles enrichis de connaissances n'avait pas encore été étudiée.

Une autre voie à explorer est celle du **augmentation des données multilingues pour l'ER**. La rareté des relations positives d'intérêt est particulièrement importante lors de l'extraction de relations à partir de données peu annotées, dans un contexte multilingue.

Mixup (Guo et al., 2019a) est une technique d'augmentation des données qui interpole linéairement les exemples et les étiquettes en entrée. Lorsqu'elle est combinée à des transformateurs, elle a démontré son efficacité dans de nombreuses tâches de classification NLP dans un cadre monolingue (Chen et al., 2022a; Sun et al., 2020a) ou multilingue (Yang et al., 2021). À notre connaissance, aucun travail antérieur n'a tenté de l'adapter à la tâche RE.

Vers une intégration des connaissances plus spécifique au domaine. Nos modèles de classification de relations multi-labels ont obtenu de meilleurs résultats lorsque nous leur avons fourni des connaissances provenant de ressources encyclopédiques structurées sous la forme d'embedding d'entités pré-entraînés. Il serait intéressant d'exploiter les connaissances disponibles dans les ressources spécifiques à un domaine, telles que les indices boursiers qui signalent les hausses ou les baisses du cours des actions des entreprises, les lexiques financiers existants,[1] des bases de données financières telles que CrunchBase,[2] qui couvre plus de 100 000 entreprises, investisseurs, acquisitions et cycles de financement, et enfin la Financial Industry Business Ontology (FIBO),[3] qui modélise les concepts financiers.

Enfin, puisque nous avons montré que les modèles proposés dans cette thèse sont portables à travers les domaines (cf. Section **??**, Chapitre 4) et les langages (cf. Section **??**, Chapitre 5), un nouveau domaine d'étude à explorer serait **l'extraction de relations inter-domaines**. Des efforts récents dans cette direction ont été menés par Bassignana and Plank (2022a) qui a a proposé le premier ensemble de données inter-domaines appelé CROSSRE, couvrant six domaines divers (actualités, politique, sciences naturelles, musique, littérature, IA) avec des annotations couvrant 17 types de relations. Il serait donc intéressant de mesurer la possibilité de transférer des modèles d'extraction de relations spécifiques entre entités à travers les domaines.

[1] https://markets.ft.com/glossary/searchLetter.asp?letter=A
[2] https://www.crunchbase.com/
[3] https://github.com/edmcouncil/fibo

Bibliography

Third Message Uunderstanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991, 1991. URL https://aclanthology.org/M91-1000.

Proceedings of the 4th Conference on Message Understanding, MUC 1992, McLean, Virginia, USA, June 16-18, 1992, 1992. ACL. ISBN 1-55860-273-9. doi: 10.3115/1072064. URL https://doi.org/10.3115/1072064.

Proceedings of the 5th Conference on Message Understanding, MUC 1993, Baltimore, Maryland, USA, August 25-27, 1993, 1993. ACL. ISBN 1-55860-336-0. doi: 10.3115/1072017. URL https://doi.org/10.3115/1072017.

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. Load what you need: Smaller versions of mutililingual bert. In Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pages 119–123, 2020.

Mostafa Abdou, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein, and Anders Søgaard. X-WikiRE: A large, multilingual resource for relation extraction as machine comprehension. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pages 265–274, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6130. URL https://aclanthology.org/D19-6130.

Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In Proceedings of the fifth ACM conference on Digital libraries, pages 85–94. ACM, 2000.

Mahtab Ahmed, Jumayel Islam, Muhammad Rifayat Samee, and Robert E Mercer. Identifying protein-protein interaction using tree lstm and structured attention. In 2019 IEEE 13th international conference on semantic computing (ICSC), pages 224–231. IEEE, 2019.

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC bioinformatics, 9(11):1–12, 2008.

Alan Akbik and Jügen Broß. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In www workshop, volume 48, 2009.

Abbas Akkasi and Mari-Francine Moens. Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey. Journal of Biomedical Informatics, 119:103820, 2021.

Nazanin Alipourfard, Beatrix Arendt, Daniel M Benjamin, Noam Benkler, Michael Bishop, Mark Burstein, Martin Bush, James Caverlee, Yiling Chen, Chae Clark, et al. Systematizing confidence in open research and evidence (score). 2021.

Abduladem Aljamel, Taha Osman, and Giovanni Acampora. Domain-specific relation extraction: Using distant supervision machine learning. In 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), volume 1, pages 92–103. IEEE, 2015.

Christoph Alt, Marc Hübner, and Leonhard Hennig. Improving relation extraction by pre-trained language representations. In Automated Knowledge Base Construction (AKBC), 2018.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. Tacred revisited: A thorough evaluation of the tacred relation extraction task. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1558–1569, 2020.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7383–7390, 2020.

Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research, 38:135–187, 2010.

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. Combining distant and partial supervision for relation extraction. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1556–1567, 2014.

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019.

Cecilia N Arighi, Zhiyong Lu, Martin Krallinger, Kevin B Cohen, W John Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy H Wu. Overview of the biocreative iii workshop. BMC bioinformatics, 12(8):1–9, 2011.

Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Extracting drug-drug interactions with attention cnns. In BioNLP 2017, pages 9–18, 2017.

Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Using drug descriptions and molecular structures for drug–drug interaction extraction from literature. Bioinformatics, 37(12): 1739–1746, 2021.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In The semantic web, pages 722–735. Springer, 2007.

I. Augenstein, D. Maynard, and F. Ciravegna. Distantly supervised web relation extraction for knowledge base population. Semantic Web Journal, 2016a.

Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. Distantly supervised web relation extraction for knowledge base population. Semantic Web, 7(4):335–349, 2016b.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 546–555, 2017.

Nathalie Aussenac-Gilles and Marie-Paule Jacques. Designing and evaluating patterns for relation acquisition from texts with caméléon. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 14(1):45–73, 2008.

Nathalie Aussenac-Gilles and Patrick Séguéla. Les relations sémantiques: du linguistique au formel. Cahiers de grammaire, (25):175–198, 2000.

Mehmet Aydar, Ozge Bozal, and Furkan Ozbay. Neural relation extraction: a survey. arXiv e-prints, pages arXiv–2007, 2020.

Nguyen Bach and Sameer Badaskar. A review of relation extraction. Literature review for Language and Statistics II, 2, 2007.

Hyeong-Ryeol Baek and Yong-Suk Choi. Enhancing targeted minority class prediction in sentence-level relation extraction. Sensors, 22(13):4911, 2022.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1279. URL https://www.aclweb.org/anthology/P19-1279.

Elisa Bassignana and Barbara Plank. Crossre: A cross-domain dataset for relation extraction. arXiv preprint arXiv:2210.09345, 2022a.

Elisa Bassignana and Barbara Plank. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 67–83, 2022b.

Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. Recon: relation extraction using knowledge graph context in a graph neural network. In Proceedings of the Web Conference 2021, pages 1673–1685, 2021.

David S Batista, Bruno Martins, and Mário J Silva. Semi-supervised bootstrapping of relationship extractors with distributional semantics. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 499–504, 2015.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, 2019.

John Berns, Patty Bick, Ryan Flugum, and Reza Houston. Do changes in md&a section tone predict investment behavior? Financial Review, 2021.

Yanelys Betancourt and Sergio Ilarri. Use of text mining techniques for recommender systems. In ICEIS (1), pages 780–787, 2020.

Abhyuday Bhartiya, Kartikeya Badola, et al. Dis-rex: A multilingual dataset for distantly supervised relation extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 849–863, 2022.

Jari Björne and Tapio Salakoski. Biomedical event extraction using convolutional neural networks and dependency parsing. In Proceedings of the BioNLP 2018 workshop, pages 98–108, 2018.

Christian Blaschke, Miguel A Andrade, Christos A Ouzounis, and Alfonso Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. In Ismb, volume 7, pages 60–67, 1999.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl_1):D267–D270, 2004.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, page 1247–1250, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581026. doi: 10.1145/1376616.1376746. URL https://doi.org/10.1145/1376616.1376746.

Robert J Boncella. Competitive intelligence and the web. Communications of the Association for Information Systems, 12(1):21, 2003.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26, 2013.

Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 615–620, 2014.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modelling under imbalanced distributions. CoRR, abs/1505.01658, 2015. URL http://arxiv.org/abs/1505.01658.

Daniel Braun, Anne Faber, Adrian Hernandez-Mendez, and Florian Matthes. Automatic relation extraction for building smart city ecosystems using dependency parsing. In Pierpaolo Basile, Valerio Basile, Danilo Croce, Felice Dell'Orletta, and Marco Guerini, editors, Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI 2018) co-located with 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018), Trento, Italy, November 22nd to 23rd, 2018, volume 2244 of CEUR Workshop Proceedings, pages 29–39. CEUR-WS.org, 2018. URL http://ceur-ws.org/Vol-2244/paper_03.pdf.

Sergey Brin. Extracting patterns and relations from the world wide web. In International workshop on the world wide web and databases, pages 172–183. Springer, 1998.

E Bruches, T Batura, A Pauls, and V Isachenko. Entity recognition and relation extraction from scientific and technical texts in russian. In Proceedings-2020 Science and Artificial Intelligence Conference, SAI ence 2020, pages 41–45. IEEE, 2020.

Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. BMC bioinformatics, 9(1):1–14, 2008.

Razvan Bunescu and Raymond Mooney. A shortest path dependency kernel for relation extraction. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 724–731, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/H05-1091.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong. Comparative experiments on learning information extractors for proteins and their interactions. Artificial Intelligence in Medicine, 33(2): 139–155, 2005.

Douglas Burdick, Mauricio Hernández, Howard Ho, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Constantin Popa, Ioana Stanoi, Shivakumar Vaithyanathan, and Sanjiv Ranjan Das. Extracting, linking and integrating data from public sources: A financial case study. Available at SSRN 2666384, 2015.

Davide Buscaldi, Anne-Kathrin Schumann, Behrang Qasemizadeh, Haïfa Zargayouna, and Thierry Charnois. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In Proceedings of the 12th international workshop on semantic evaluation, pages 679–688, 2018.

Pere-Lluís Huguet Cabot and Roberto Navigli. Rebel: Relation extraction by end-to-end language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2370–2381, 2021.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. Bidirectional recurrent convolutional neural network for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 756–765, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1072. URL https://aclanthology.org/P16-1072.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: a novel approach to a semantically-aware representation of items. In Proceedings of the 2015 NAACL: Human Language Technologies, pages 567–577, 2015.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. Artificial Intelligence, 240:36–64, 2016.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In PML4DC at ICLR 2020, 2020.

Yee Seng Chan and Dan Roth. Exploiting background knowledge for relation extraction. In Proceedings of the 23rd COLING, pages 152–160. ACL, 2010.

Deli Chen, Yanyan Zou, Keiko Harimoto, Ruihan Bao, Xuancheng Ren, and Xu Sun. Incorporating fine-grained events in stock movement prediction. In Proceedings of the Second Workshop on Economics and Natural Language Processing, pages 31–40, 2019.

Elizabeth S. Chen, George Hripcsak, Hua Xu, Marianthi Markatou, and Carol Friedman. Automated Acquisition of Disease–Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. Journal of the American Medical Informatics Association, 15(1):87–98, 01 2008. ISSN 1067-5027. doi: 10.1197/jamia.M2401. URL https://doi.org/10.1197/jamia.M2401.

Hui Chen, Wei Han, Diyi Yang, and Soujanya Poria. Doublemix: Simple interpolation-based data augmentation for text classification. In Proceedings of the 29th International Conference on Computational Linguistics, pages 4622–4632, 2022a.

Liwei Chen, Yansong Feng, Songfang Huang, Yong Qin, and Dongyan Zhao. Encoding relation requirements for relation extraction via joint inference. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 818–827, 2014.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In Proceedings of the ACM Web Conference 2022, pages 2778–2788, 2022b.

Yanping Chen, Weizhe Yang, Kai Wang, Yongbin Qin, Ruizhang Huang, and Qinghua Zheng. A neuralized feature engineering method for entity relation extraction. Neural Networks, 141:249–260, 2021.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. Hacred: A large-scale relation extraction dataset toward hard cases in practical applications. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2819–2831, 2021.

Nancy A Chinchor. Overview of muc-7/met-2. Technical report, SCIENCE APPLICATIONS INTERNATIONAL CORP SAN DIEGO CA, 1998.

Patricia Chiril. Automatic Hate Speech Detection on Social Media. (Détection automatique des messages haineux sur les réseaux sociaux). PhD thesis, Paul Sabatier University, Toulouse, France, 2021. URL https://tel.archives-ouvertes.fr/tel-03599458.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. He said "who's gonna take care of your children when you are at ACL?": Reported Sexist Acts are Not Sexist. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4055–4066, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.373. URL https://www.aclweb.org/anthology/2020.acl-main.373.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2833–2844, 2021.

Rajesh Chowdhary, Jinfeng Zhang, and Jun S Liu. Bayesian inference of protein–protein interactions from biological literature. Bioinformatics, 25(12):1536–1542, 2009.

Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. Automatic recognition of topic-classified relations between prostate cancer and genes using medline abstracts. In BMC bioinformatics, volume 7, pages 1–8. Springer, 2006.

Philipp Cimiano. Ontology learning and population from text - algorithms, evaluation and applications. 2006.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. Relation classification as two-way span-prediction. arXiv preprint arXiv:2010.04829, 2020.

Jacob Cohen. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1):37–46, 1960.

Michael Collins and Nigel Duffy. Convolution kernels for natural language. In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01, page 625–632, Cambridge, MA, USA, 2001. MIT Press.

Sandra Collovini, Patricia Nunes Gonçalves, Guilherme Cavalheiro, Joaquim Santos, and Renata Vieira. Relation extraction for competitive intelligence. In International Conference on Computational Processing of the Portuguese Language, pages 249–258. Springer, 2020.

Marco Costantino, Russell J Collingham, and Richard G Morgan. Qualitative information in finance: Natural language processing and information extraction. Neuro Ve t Journal, 4, 1996a.

Marco Costantino, Richard G Morgan, and Russell J Collingham. Financial information extraction using pre-defined and user-definable templates in the lolita system. Journal of computing and information technology, 4(4):241–255, 1996b.

Claude Coulombe. Text data augmentation made simple by leveraging NLP cloud apis. CoRR, abs/1812.04718, 2018. URL http://arxiv.org/abs/1812.04718.

Mark Craven, Johan Kumlien, et al. Constructing biological knowledge bases by extracting information from text sources. In ISMB, volume 1999, pages 77–86, 1999.

Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 423–429, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1219009. URL https://www.aclweb.org/anthology/P04-1054.

Aron Culotta, Andrew McCallum, and Jonathan Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In Proceedings of the Main Conference on Human Language Technology NAACL, pages 296–303. ACL, 2006.

Zihang Dai, Lei Li, and Wei Xu. Cfo: Conditional focused neural question answering with large-scale knowledge bases. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 800–810. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1076. URL http://aclweb.org/anthology/P16-1076.

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. Comparative toxicogenomics database (ctd): update 2021. Nucleic acids research, 49(D1):D1138–D1143, 2021.

Sandra Collovini de Abreu and Renata Vieira. Relp: Portuguese open relation extraction. KO KNOWLEDGE ORGANIZATION, 44(3):163–177, 2017.

Daniel De Los Reyes, Allan Barcelos, Renata Vieira, and Isabel Manssour. Related named entities classification in the economic-financial context. In Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 8–15, 2021.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessieres, and Claire Nédellec. Overview of the bacteria biotope task at bionlp shared task 2016. In Proceedings of the 4th BioNLP shared task workshop, pages 12–22, 2016.

Vinicio DeSola, Kevin Hanna, and Pri Nonis. Finbert: pre-trained model on sec filings for financial natural language tasks. 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Lee R Dice. Measures of the amount of ecologic association between species. Ecology, 26 (3):297–302, 1945.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In LREC confrence, Lisbon Portugal, volume 2, pages 837–840, 2004.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. Question answering over freebase with multi-column convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 260–269, 2015.

Cícero Dos Santos, Bing Xiang, and Bowen Zhou. Classifying relations by ranking with convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 626–634, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1061. URL https://www.aclweb.org/anthology/P15-1061.

Jeffrey L Elman. Finding structure in time. Cognitive science, 14(2):179–211, 1990.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-rex: A large scale alignment of natural language with knowledge base triples. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

Safaa Eltyeb and Naomie Salim. Chemical named entities recognition: a review on approaches and applications. Journal of cheminformatics, 6(1):1–12, 2014.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. Open information extraction from the web. Dec 2008. URL https://allenai.org/content/team/orene/etzioni-cacm08.pdf.

Matan Eyal, Asaf Amrami, Hillel Taub-Tabib, and Yoav Goldberg. Bootstrapping relation extractors using syntactic search by examples. arXiv preprint arXiv:2102.05007, 2021.

Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In Proceedings of the 2011 conference on EMNLP, pages 1535–1545, 2011.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 889–898, 2018.

Manaal Faruqui and Shankar Kumar. Multilingual open relation extraction using cross-lingual projection. arXiv preprint arXiv:1503.06450, 2015.

Jean-Philippe Fauconnier and Mouna Kamel. Discovering hypernymy relations using text layout. In 4th Joint Conference on Lexical and Computational Semantics (SEM 2015), pages pp–249, 2015.

Jean-Philippe Fauconnier, Mouna Kamel, and Bernard Rothenburger. A supervised machine learning approach for taxonomic relation recognition through non-linear enumerative structures. In Proceedings of the 30th Annual ACM Symposium on Applied Computing, pages 423–425, 2015.

Jenny Rose Finkel, Christopher D. Manning, and Andrew Y. Ng. Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, pages 618–626, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610162.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378, 1971.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relex—relation extraction using dependency parse trees. Bioinformatics, 23(3):365–371, 2007.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fewrel 2.0: Towards more challenging few-shot relation classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6250–6255, 2019.

Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6174–6181, 2020.

Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Goran Glavas, and Mathias Niepert. Benchie: Open information extraction evaluation based on facts, not tokens. arXiv preprint arXiv:2109.06850, 2021.

Adel Ghamnia, Mouna Kamel, Cassia Trojahn dos Santos, Cécile Fabre, and Nathalie Aussenac-Gilles. Extraction de relations: combiner les techniques pour s' adapter à la diversité du texte. In 28es Journées francophones d'Ingénierie des Connaissances IC 2017, pages 86–97, 2017.

Tamar Gilad and Benjamin Gilad. Smr forum: business intelligence-the quiet revolution. Sloan Management Review (1986-1998), 27(4):53, 1986.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. Semeval-2007 task 04: Classification of semantic relations between nominals. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 13–18, 2007.

Yuriy Gorodnichenko, Jan Svejnar, and Katherine Terrell. Globalization and innovation in emerging markets. Technical report, National Bureau of Economic Research, 2008.

Ralph Grishman and Beth M Sundheim. Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996.

Jonatas S Grosman, Pedro HT Furtado, Ariane MB Rodrigues, Guilherme G Schardong, Simone DJ Barbosa, and Hélio CV Lopes. Eras: Improving the quality control in the annotation process for natural language processing tasks. Information Systems, 93:101553, 2020.

Paul Groth, Mike Lauruhn, Antony Scerri, and Ron Daniel Jr. Open information extraction on scientific text: An evaluation. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3414–3423, 2018.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. A knowledge-enhanced pretraining model for commonsense story generation. Transactions of the Association for Computational Linguistics, 8:93–108, 2020.

Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. arXiv preprint arXiv:1905.08941, 2019a.

Qiushi Guo, Xin Wang, and Dehong Gao. Dependency position encoding for relation extraction. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1601–1606, 2022.

Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 241–251, 2019b.

Benjamin Hachey. Towards generic relation extraction. Institute for Communicating and Collaborative Systems School of Informatics, 2009.

Martin Had, Felix Jungermann, and Katharina Morik. Relation extraction for monitoring economic networks. In International Conference on Application of Natural Language to Information Systems, pages 103–114. Springer, 2009.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1735–1742. IEEE, 2006.

Kamran Hameed, Noman Arshed, Naveed Yazdani, and Mubbasher Munir. On globalization and business competitiveness: A panel data country classification. Estudios De Economia Aplicada, 39(2):1–27, 2021.

DB Han. Klue annotation guidelines-version 2.0. Technical report, Technical Report RC25042, IBM Research, August, 2010.

Songqiao Han, Xiaoling Hao, and Hailiang Huang. An event-extraction approach for business analysis from online chinese news. Electronic Commerce Research and Applications, 28: 244–260, 2018a.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4803–4809, 2018b.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. More data, more relations, more context and more openness: A review and outlook for relation extraction. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 745–758, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.75.

Zellig S Harris. Distributional structure. Word, 10(2-3):146–162, 1954.

Peter Hart. The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory, 14(3):515–516, 1968.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics, pages 539–545. Association for Computational Linguistics, 1992. URL https://www.aclweb.org/anthology/C92-2082.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, page 33–38, USA, 2010. Association for Computational Linguistics.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of biomedical informatics, 46(5):914–920, 2013.

Lena Hettinger, Alexander Dallmann, Albin Zehe, Thomas Niebler, and Andreas Hotho. Claire at semeval-2018 task 7: classification of relations using embeddings. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 836–841, 2018.

Lars Hillebrand, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. Kpi-bert: A joint named entity recognition and relation extraction model for financial reports. arXiv preprint arXiv:2208.02140, 2022.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In In Proc. of NeurIPS, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9 (8):1735–1780, 1997.

Frederik Hogenboom, Michael de Winter, Flavius Frasincar, and Uzay Kaymak. A news event-driven approach for the historical value at risk method. Expert Systems with Applications, 42(10):4667–4675, 2015.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and al et. spaCy: Industrial-strength Natural Language Processing in Python. Version 3.4 [Computer Software] https://doi.org/10.5281/zenodo.1212303, 2022. URL https://doi.org/10.5281/zenodo.1212303.

Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel S Weld, Roy Schwartz, and Hannaneh Hajishirzi. Extracting a knowledge base of mechanisms from covid-19 papers. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4489–4503, 2021.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5203–5213, 2019.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge-enhanced pre-trained language models. arXiv preprint arXiv:2211.05994, 2022.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. Balancing methods for multi-label text classification with long-tailed class distribution. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8153–8161, 2021a.

Yuan Huang, Zhixing Li, Wei Deng, Guoyin Wang, and Zhimin Lin. D-bert: Incorporating dependency-based attention into bert for relation extraction. CAAI Transactions on Intelligence Technology, 2021b.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. A french corpus and annotation schema for named entity recognition and relation extraction of financial news. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 2293–2299, 2020.

Gilles Jacobs and Veronique Hoste. Sentivent: enabling supervised information extraction of company-specific events in economic and financial news. LANGUAGE RESOURCES AND EVALUATION, 2021.

Gilles Jacobs, Els Lefever, and Véronique Hoste. Economic event detection in company-specific news text. In Proceedings of the First Workshop on Economics and Natural Language Processing, pages 1–10, 2018.

Marie-Paule Jacques and Nathalie Aussenac-Gilles. Variabilité des performances des outils de tal et genre textuel cas des patrons lexico-syntaxiques. Revue TAL, 47(1):11–32, 2006.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. Scirex: A challenge dataset for document-level information extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7506–7516, 2020.

Amarin Jettakul, Duangdao Wichadakul, and Peerapon Vateekul. Relation extraction between bacteria and biotopes from biomedical texts with attention mechanisms and domain-specific contextual representations. 2019.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.

Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1148–1158. Association for Computational Linguistics, 2011.

Jing Jiang and ChengXiang Zhai. A systematic exploration of the feature space for relation extraction. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 113–120, 2007.

Di Jin, Franck Dernoncourt, Elena Sergeeva, Matthew McDermott, and Geeticka Chauhan. Mit-medg at semeval-2018 task 7: Semantic relation classification via convolution neural network. In Proceedings of the 12th international workshop on semantic evaluation, pages 798–804, 2018.

Daniel Jurafsky and James H. Martin. Speech and Language Processing An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition. San Val, 2018. URL https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Salomon Kabongo, Jennifer D'Souza, and Sören Auer. Automated mining of leaderboards for empirical ai research. In International Conference on Asian Digital Libraries, pages 453–470. Springer, 2021.

Larry Kahaner. Competitive Intelligence: how to gather analyze and use information to move your business to the top. Simon and Schuster, 1997.

Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, ACLdemo '04, page 22–es, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1219044.1219066. URL https://doi.org/10.3115/1219044.1219066.

Mouna Kamel, Cassia Trojahn dos Santos, Adel Ghamnia, Nathalie Aussenac-Gilles, and Cécile Fabre. Extracting hypernym relations from wikipedia disambiguation pages: comparing symbolic and machine learning approaches. In International Conference on Computational Semantics (IWCS 2017), pages 1–12, 2017a.

Mouna Kamel, Cassia Trojahn, Adel Ghamnia, Nathalie Aussenac-Gilles, and Cécile Fabre. A distant learning approach for extracting hypernym relations from wikipedia disambiguation pages. Procedia computer science, 112:1764–1773, 2017b.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In International Conference on Learning Representations, 2019.

Hadjer Khaldi, Amine Abdaoui, Farah Benamara, Grégoire Sigel, and Nathalie Aussenac-Gilles. Classification de relations pour l'intelligence économique et concurrentielle. In 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles, number 2, pages 27–39. ATALA: Association pour le traitement automatique des langues; AFCP . . . , 2020.

Hadjer Khaldi, Farah Benamara, Amine Abdaoui, Nathalie Aussenac-Gilles, and EunBee Kang. Multilevel entity-informed business relation extraction. In International Conference on Applications of Natural Language to Information Systems, pages 105–118. Springer, 2021.

Hadjer Khaldi, Farah Benamara, Camille Pradel, and Nathalie Aussenac-Gilles. How can a teacher make learning from sparse data softer? application to business relation extraction. In 4th Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2022), pages 22–28, 2022a.

Hadjer Khaldi, Farah Benamara, Camille Pradel, and Nathalie Aussenac Gilles. A closer look to your business network: Multitask relation extraction from economic and financial french content. In The AAAI-22 Workshop on Knowledge Discovery from Unstructured Data in Financial Services, 2022b.

Hadjer Khaldi, Farah Benamara, Camille Pradel, Grégoire Sigel, and Nathalie Aussenac-Gilles. How's business going worldwide? a multilingual annotated corpus for business relation extraction. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), page 3696–3705, 2022c.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. Corpus annotation for mining biomedical events from literature. BMC bioinformatics, 9(1):10, 2008.

Seongsik Park Harksoo Kim. Improving sentence-level relation extraction through curriculum learning. 2021.

Sun Kim, Haibin Liu, Lana Yeganova, and W John Wilbur. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. Journal of biomedical informatics, 55:23–30, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. arXiv preprint arXiv: 1412.6980.

Mitchell Koch, John Gilmer, Stephen Soderland, and Daniel S Weld. Type-aware distantly supervised relation extraction with linked arguments. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1891–1901, 2014.

Abdullatif Köksal and Arzucan Özgür. The relx dataset and matching the multilingual blanks for cross-lingual relation classification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 340–350, 2020.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, et al. Alignment-augmented consistent translation for multilingual open information extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2502–2517, 2022.

Natalia Konstantinova. Review of relation extraction methods: What is new out there? In International Conference on Analysis of Images, Social Networks and Texts, pages 15–28. Springer, 2014.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. Overview of the protein-protein interaction annotation extraction task of biocreative ii. Genome biology, 9(2):1–19, 2008.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. Overview of the biocreative vi chemical-protein interaction track. In Proceedings of the sixth BioCreative challenge evaluation workshop, volume 1, pages 141–146, 2017.

Ruben Kruiper, Julian Vincent, Jessica Chen-Burger, Marc Desmulliez, and Ioannis Konstas. In layman's terms: Semi-open relation extraction from scientific texts. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1489–1500, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.137. URL https://aclanthology.org/2020.acl-main.137.

Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In Icml, volume 97, pages 179–186. Citeseer, 1997.

Shantanu Kumar. A survey of deep learning methods for relation extraction. CoRR, 2017.

Pierre Lafon. Sur la variabilité de la fréquence des formes dans un corpus. Mots. Les langages du politique, 1(1):127–165, 1980.

Dan Lahav, Jon Saad-Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S. Weld, and Tom Hope. A search engine for discovery of scientific challenges and directions. In AAAI, 2022.

Tuan Lai, Heng Ji, Cheng Xiang Zhai, and Quan Hung Tran. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, pages 6248–6260. Association for Computational Linguistics (ACL), 2021.

RY Lau and Wenping Zhang. Semi-supervised statistical inference for business entities extraction and business relations discovery. In SIGIR 2011 workshop, Beijing, China, pages 41–46, 2011.

Jorma Laurikkala. Improving identification of difficult small classes by balancing class distribution. In Conference on Artificial Intelligence in Medicine in Europe, pages 63–66. Springer, 2001.

Anne Lauscher, Yide Song, and Kiril Gashteovski. Minscie: Citation-centered open information extraction. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pages 386–387, 2019. doi: 10.1109/JCDL.2019.00083.

LDC. Annotation guidelines for relation detection and characterization (rdc). Technical report, Linguistic Data Consortium, 2004.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.302.

Hoang Quynh Le, Duy Cat Can, Quang Thuy Ha, and Nigel Collier. A richer-but-smarter shortest dependency path with attentive augmentation for relation extraction. In 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, volume 1, pages 2902–2912. Association for Computational Linguistics, 2019.

William Léchelle, Fabrizio Gotti, and Philippe Langlais. Wire57: A fine-grained benchmark for open information extraction. In Proceedings of the 13th Linguistic Annotation Workshop, pages 6–15, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

Hee-Jin Lee, Sang-Hyung Shim, Mi-Ryoung Song, Hyunju Lee, and Jong C Park. Comagc: a corpus with multi-faceted annotations of gene-cancer relations. BMC bioinformatics, 14 (1):1–17, 2013.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. MIT at SemEval-2017 task 10: Relation extraction with convolutional neural networks. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 978–984, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2171. URL https://www.aclweb.org/anthology/S17-2171.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240, 2020.

Joohong Lee, Sangwoo Seo, and Yong Suk Choi. Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. Symmetry, 11 (6), june 2019. ISSN 2073-8994. doi: 10.3390/sym11060785.

Els Lefever and Véronique Hoste. A classification-based approach to economic event detection in Dutch news text. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 330–335, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1051.

Joël Legrand, Yannick Toussaint, Chedy Raïssi, and Adrien Coulet. Syntax-based transfer learning for the task of biomedical relation extraction. Journal of Biomedical Semantics, 12(1):1–11, 2021.

Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. Pytorch-biggraph: A large scale graph embedding system. Proceedings of Machine Learning and Systems, 1:120–131, 2019.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. Sensebert: Driving some sense into bert. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4656–4667, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Belinda Z Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. Efficient one-pass end-to-end entity linking for questions. In Proceedings of EMNLP, pages 6433–6441, 2020a.

J Li, Y Sun, RJ Johnson, D Sciaky, CH Wei, R Leaman, AP Davis, CJ Mattingly, TC Wiegers, and Z Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. Database: the Journal of Biological Databases and Curation, 2016, 2016.

Jun Li, Guimin Huang, Jianheng Chen, and Yabing Wang. Dual cnn for relation extraction with knowledge-based attention and word embeddings. Computational Intelligence and Neuroscience, 2019, 2019.

Qian Li, Hao Peng, Jianxin Li, Yiming Hei, Rui Sun, Jiawei Sheng, Shu Guo, Lihong Wang, Jia Wu, Amin Beheshti, et al. A comprehensive survey on schema-based event extraction with deep learning. arXiv preprint arXiv:2107.02126, 2021.

R. Li, C. Yang, T. Li, and S. Su. Midtd: A simple and effective distillation framework for distantly supervised relation extraction. ACM Transactions on Information Systems (TOIS), (4):1–32, 2022.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 465–476, 2020b.

Zhijing Li, Yuchen Lian, Xiaoyong Ma, Xiangrong Zhang, and Chen Li. Bio-semantic relation extraction with attention-based external knowledge reinforcement. BMC Bioinformatics, 21:1–18, 2020c.

Percy Liang, Michael I. Jordan, and Dan Klein. Learning dependency-based compositional semantics. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, pages 590–599, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL http://dl.acm.org/citation.cfm?id=2002472.2002547.

Xin Liang, Dawei Cheng, Fangzhou Yang, Yifeng Luo, Weining Qian, and Aoying Zhou. F-hmtc: Detecting financial events for investment decisions based on neural hierarchical multi-label text classification. In IJCAI, pages 4490–4496, 2020.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. Modeling multi-granularity hierarchical features for relation extraction. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5088–5098, 2022.

Sangrak Lim and Jaewoo Kang. Chemical–gene relation extraction using recursive neural network. Database, 2018.

Sangrak Lim, Kyubum Lee, and Jaewoo Kang. Drug drug interaction extraction from the literature using a recursive neural network. PloS one, 13(1):e0190926, 2018.

H. Lin, Y. Lu, X. Han, and L. Sun. Adaptive scaling for sparse detection in information extraction. arXiv preprint arXiv:1805.00250, 2018.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017.

Xiangyu Lin, Tianyi Liu, Weijia Jia, and Zhiguo Gong. Distantly supervised relation extraction using multi-layer revision network and confidence-based multi-instance learning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 165–174, 2021.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2124–2133, 2016.

ChunYang Liu, WenBo Sun, WenHan Chao, and Wanxiang Che. Convolution neural network for relation extraction. In International Conference on Advanced Data Mining and Applications, pages 231–242. Springer, 2013.

Jianyi Liu, Xi Duan, Ru Zhang, Youqiang Sun, Lei Guan, and Bingjie Lin. Relation classification via bert with piecewise convolution and focal loss. Plos one, 16(9):e0257092, 2021a.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv e-prints, pages arXiv–2107, 2021b.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 2901–2908, 2020.

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. A dependency-based neural network for relation classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 285–290, 2015.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv: 1907.11692, 2019.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pages 4513–4519, 2021c.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. Journal of Machine Learning Research, 2(Feb): 419–444, 2002.

Keming Lu, I Hsu, Wenxuan Zhou, Mingyu Derek Ma, Muhao Chen, et al. Summarization as indirect supervision for relation extraction. arXiv preprint arXiv:2205.09837, 2022.

Wei Lu, Hwee Tou Ng, Wee Sun Lee, and Luke S. Zettlemoyer. A generative model for parsing natural language to meaning representations. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, pages 783–792, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613815.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. Scientific information extraction with semi-supervised neural tagging. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2641–2651, 2017.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3219–3232, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1360. URL https://aclanthology.org/D18-1360.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. The uwnlp system at semeval-2018 task 7: Neural relation extraction model with selectively incorporated concept embeddings. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 788–792, 2018b.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. Biored: A comprehensive biomedical relation extraction dataset. arXiv preprint arXiv:2204.04263, 2022.

Shengfei Lyu and Huanhuan Chen. Relation classification with entity type restriction. arXiv preprint arXiv:2105.08393, 2021.

Shengfei Lyu, Jin Cheng, Xingyu Wu, Lizhen Cui, Huanhuan Chen, and Chunyan Miao. Auxiliary learning for relation extraction. IEEE Transactions on Emerging Topics in Computational Intelligence, 2020.

Yongqiang Ma, Jiawei Liu, Wei Lu, and Qikai Cheng. Beyond tasks, methods, and metrics: extracting metrics-driven mechanism from the abstracts of ai articles. 2022.

Ian Magnusson and Scott Friedman. Extracting fine-grained knowledge graphs of scientific claims: Dataset and transformer-based results. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4651–4658, 2021.

Angrosh Mandya, Danushka Bollegala, Frans Coenen, and Katie Atkinson. A dataset for inter-sentence relation extraction using distant supervision. In LREC 2018-11th International Conference on Language Resources and Evaluation, pages 1559–1565, 2019.

Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets, volume 126. ICML United States, 2003.

Stefano Marchesin and Gianmaria Silvello. Tbga: A large-scale gene-disease association dataset for biomedical relation extraction. BMC Bioinformatics, 23, 2022.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model, 2019. arXiv preprint arXiv: 1911.03894.

Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. Information extraction meets the semantic web: a survey. Semantic Web, pages 1–81, 2020.

Mausam Mausam. Open information extraction systems and downstream applications. In Proceedings of the twenty-fifth international joint conference on artificial intelligence, pages 4074–4077, 2016.

Andrew McCallum and David Jensen. A note on the unification of information extraction and data mining using conditional-probability, relational models. Computer Science Department Faculty Publication Series, page 42, 2003.

Sachin Mehta, Rik Koncel-Kedziorski, Mohammad Rastegari, and Hannaneh Hajishirzi. Define: Deep factorized input token embeddings for neural sequence modeling. In International Conference on Learning Representations, 2019.

Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing, pages 404–411, 2004.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013b.

Tomáš Mikolov, Édouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In Proceedings of LREC, 2018.

George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38 (11):39–41, 1995.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. International journal of lexicography, 3(4):235–244, 1990.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV), pages 565–571. IEEE, 2016.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011, 2009.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. Ace 2004 multilingual training corpus. Linguistic Data Consortium, Philadelphia, pages 1–1, 2005.

Sayantan Mitra, Sriparna Saha, and Mohammed Hasanuzzaman. A multi-view deep neural network model for chemical-disease relation extraction from imbalanced datasets. IEEE Journal of Biomedical and Health Informatics, 24(11):3315–3325, 2020. doi: 10.1109/JBHI.2020.2983365.

Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. arXiv preprint arXiv:1601.00770, 2016.

S. S. Moein, E. Romina, and S. Mehrnoush. Improving persian relation extraction models bu data augmentation. In Proceedings of The Second International Workshop NSURL 2021 co-located with ICNLSP 2021, pages 32–37, 2021.

Reham Mohamed, Nagwa M El-Makky, and Khaled Nagi. Hybqa: Hybrid deep relation extraction for question answering on freebase. In KEOD, pages 128–136, 2017.

David B Montgomery and Charles B Weinberg. Toward strategic intelligence systems. Journal of Marketing, 43(4):41–52, 1979.

Raymond J Mooney and Razvan C Bunescu. Subsequence kernels for relation extraction. In Advances in neural information processing systems, pages 171–178, 2006.

Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In Proceedings of the AAAI conference on artificial intelligence, volume 30, 2016.

Ajay Nagesh. Exploring relational features and learning under distant supervision for information extraction tasks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 40–47, 2015.

Sangha Nam, Kijong Han, Eun-kyung Kim, and Key-Sun Choi. Distant supervision for relation extraction with multi-sense word embedding. In Proceedings of the 9th Global Wordnet Conference, pages 239–244, 2018.

Hidetsugu Nanba, Yoko Doi, Miho Tsujita, Toshiyuki Takezawa, and Kazutoshi Sumiya. Construction of a cooking ontology from cooking recipes and patents. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pages 507–516, 2014.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193:217–250, 2012.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. Overview of bionlp shared task 2013. In Proceedings of the BioNLP shared task 2013 workshop, pages 1–7, 2013.

Dat P. T Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. Exploiting syntactic and semantic information for relation extraction from wikipedia. In In IJCAI07-TextLinkWS, 2007a.

Dat P. T. Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. Relation extraction from wikipedia using subtree mining. In Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2, AAAI'07, page 1414–1420. AAAI Press, 2007b. ISBN 9781577353232.

Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 39–48, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1506. URL https://www.aclweb.org/anthology/W15-1506.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. A survey on open information extraction. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3866–3878. Association for Computational Linguistics, 2018.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. Sirius-ltg-uio at semeval-2018 task 7: Convolutional neural networks with shortest dependency paths for semantic relation extraction and classification in scientific papers. arXiv preprint arXiv:1804.08887, 2018.

Christopher Norman, Mariska Leeflang, René Spijker, Evangelos Kanoulas, and Aurélie Névéol. A distantly supervised dataset for automated data extraction from diagnostic studies. In Proceedings of the 18th BioNLP Workshop and Shared Task, pages 105–114, 2019.

Esmaeil Nourani and Vahideh Reshadat. Association extraction from biomedical literature based on representation and transfer learning. Journal of theoretical biology, 488:110112, 2020.

Diarmuid Ó Séaghdha. Annotating and learning compound noun semantics. In Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop, ACL '07, pages 73–78, 2007.

Thomas Oberlechner and Sam Hocking. Information sources, news, and rumors in financial markets: Insights into the foreign exchange market. Journal of economic psychology, 25 (3):407–424, 2004.

Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458, 2015.

Arzucan Özgür, Thuy Vu, Güneş Erkan, and Dragomir R Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics, 24(13):i277–i285, 2008.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Andrea Papaluca, Daniel Krefl, Hanna Suominen, and Artem Lenskiy. Pretrained knowledge base embeddings for improved sentential relation extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 373–382, 2022.

Yannis Papanikolaou and Andrea Pierleoni. Dare: Data augmented relation extraction with gpt-2. arXiv preprint arXiv:2004.13845, 2020.

Chanhee Park, Jinuk Park, and Sanghyun Park. Agcn: Attention-based graph convolutional networks for drug-drug interaction extraction. Expert Systems with Applications, 159: 113538, 2020.

Tommaso Pasini and Roberto Navigli. Train-o-matic: Supervised word sense disambiguation with no (manual) effort. Artificial Intelligence, 279:103215, 2020.

Constantine Passaris. The business of globalization and the globalization of business. Journal of Comparative International Management, 9(1):3–18, 2006.

Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. Relation extraction: A survey. arXiv preprint arXiv:1712.05191, 2017.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Learning from context or names? an empirical study on neural relation extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3661–3672, 2020a.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. An empirical study of multi-task learning on bert for biomedical text mining. arXiv preprint arXiv:2005.02799, 2020b.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. In Proceedings of EMNLP-IJCNLP, pages 43–54, 2019.

Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In International conference on machine learning, pages 4095–4104. PMLR, 2018.

Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research, page gkw943, 2016.

Vassilis Plachouras and Jochen L Leidner. Information extraction of regulatory enforcement actions: From anti-money laundering compliance to countering terrorism finance. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pages 950–953, 2015.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. E-BERT: Efficient-yet-effective entity embeddings for BERT. In EMNLP, pages 803–818. ACL, 2020.

Bhanu Pratap, Daniel Shank, Oladipo Ositelu, and Byron Galbraith. Talla at semeval-2018 task 7: Hybrid loss optimization for relation classification using convolutional neural networks. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 863–867, 2018.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 101–108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14. URL https://aclanthology.org/2020.acl-demos.14.

Yu Qian, Xiongwen Deng, Qiongwei Ye, Baojun Ma, and Hua Yuan. On detecting business event from the headlines and leads of massive online news articles. Information Processing & Management, 56(6):102086, 2019.

Han Qin, Yuanhe Tian, and Yan Song. Relation extraction with word graphs from n-grams. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2860–2868, 2021.

Pengda Qin, Weiran Xu, and William Yang Wang. Dsgan: Generative adversarial training for distant supervision relation extraction. arXiv preprint arXiv:1805.09929, 2018a.

Pengda Qin, Weiran Xu, and William Yang Wang. Robust distant supervision relation extraction via deep reinforcement learning. arXiv preprint arXiv:1805.09927, 2018b.

Chanqin Quan, Lei Hua, Xiao Sun, and Wenjun Bai. Multichannel convolutional neural network for biological relation extraction. BioMed research international, 2016, 2016.

Chris Quirk and Hoifung Poon. Distant supervision for relation extraction beyond the sentence boundary. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1171–1182, 2017.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, 2016.

B. Raphael. SIR: A Computer Program for Semantic Information Retrieval. Ad 608. Massachusetts Institute of Technology, 1964. URL https://books.google.fr/books?id=7osaSwAACAAJ.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

Feiliang Ren, Di Zhou, Zhihui Liu, Yongcheng Li, Rongsheng Zhao, Yongkang Liu, and Xiaobo Liang. Neural relation classification with text descriptions. In Proceedings of the 27th international conference on computational linguistics, pages 1167–1177, 2018.

Tim Repke and Ralf Krestel. Extraction and representation of financial entities from text. In Data Science for Economics and Finance, pages 241–263. Springer, Cham, 2021.

D.D.L. Reyes, D. Trajano, I. Manssour, R. Vieira, and R. Bordini. Entity relation extraction from news articles in portuguese for competitive intelligence based on bert. In Brazilian Conference on Intelligent Systems, pages 449–464. Springer, 2021.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. mluke: The power of entity representations in multilingual pretrained language models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7316–7330, 2022.

Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 148–163. Springer, 2010.

Petar Ristoski and Heiko Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In International Semantic Web Conference, pages 498–514. Springer, 2016.

Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004, pages 1–8, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-2401.

Jonathan Rotsztejn, Nora Hollenstein, and Ce Zhang. Eth-ds3lab at semeval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 689–696, 2018.

Jacobo Rouces, Gerard de Melo, and Katja Hose. Framebase: Enabling integration of heterogeneous knowledge. Semantic Web, 8(6):817–850, 2017.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. Label verbalization and entailment for effective zero and few-shot relation extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1199–1212, 2021.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

Jasmin Šarić, Lars Juhl Jensen, Rossitza Ouzounova, Isabel Rojas, and Peer Bork. Extraction of regulatory gene/protein networks from medline, 2006.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2339–2352, 2021.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In European semantic web conference, pages 593–607. Springer, 2018.

Gustavo Schwenkler and Hannan Zheng. The network of firms implied by the news. Boston University Questrom School of Business Research Paper, (3320859), 2019.

Alessandro Seganti, Klaudia Firlag, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. Multilingual entity and relation extraction dataset and model. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1946–1955, 2021.

Patrick Séguéla. Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. Terminologies nouvelles, 19:52–60, 1999.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 341–350, 2013.

Satoshi Sekine. Named entity: History and future. Project notes, New York University, page 4, 2004.

Ravina Sewlal. Effectiveness of the web as a competitive intelligence tool. South African Journal of Information Management, 6(1), 2004.

Anuj Sharma, Vassilis Virvilis, Tina Lekka, and Christos Andronis. Binary relation extraction from biomedical literature using dependency trees and svms. bioRxiv, 2016.

Prafull Sharma and Yingbo Li. Self-supervised contextual keyword and keyphrase retrieval with self-labelling. Preprints, 2019.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. Finred: A dataset for relation extraction in financial domain. In Companion Proceedings of the Web Conference 2022, pages 595–597, 2022.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Q-bert: Hessian based ultra low precision quantization of bert. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8815–8821, 2020.

Yatian Shen and Xuanjing Huang. Attention-based convolutional neural network for semantic relation extraction. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2526–2536, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1238.

Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. arXiv preprint arXiv:1904.05255, 2019.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. Syntactic search by example. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 17–23, 2020.

Alisa Smirnova and Philippe Cudré-Mauroux. Relation extraction using distant supervision: A survey. ACM Computing Surveys (CSUR), 51(5):1–35, 2018.

Alisa Smirnova, Julien Audiffren, and Philippe Cudré-Mauroux. Apcnn: tackling class imbalance in relation extraction through aggregated piecewise convolutional neural networks. In 2019 6th Swiss Conference on Data Science (SDS), pages 63–68. IEEE, 2019.

Rion Snow, Daniel Jurafsky, and Andrew Ng. Learning syntactic patterns for automatic hypernym discovery. In L. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper/2004/file/358aee4cc897452c00244351e4d91f69-Paper.pdf.

D. Song, J. Xu, J. Pang, and H. Huang. Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data. Information Sciences, 573:222–238, 2021.

J. F. Sowa. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984. ISBN 0-201-14472-7.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.

Gabriel Stanovsky and Ido Dagan. Creating a large benchmark for open information extraction. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2300–2305, 2016.

George J Stigler. The development of utility theory. i. Journal of political economy, 58(4): 307–327, 1950.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. Re-tacred: Addressing shortcomings of the tacred dataset. In Proceedings of the Thirty-fifth AAAI Conference on Aritificial Intelligence, 2021.

Jannik Strötgen and Michael Gertz. Multilingual and cross-domain temporal tagging. Language Resources and Evaluation, 47(2):269–298, 2013.

Peng Su, Yifan Peng, and K Vijay-Shanker. Improving bert model using contrastive learning for biomedical relation extraction. In Proceedings of the 20th Workshop on Biomedical Language Processing, pages 1–10, 2021a.

Yusheng Su, Xu Han, Zhengyan Zhang, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. Cokebert: Contextual knowledge selection and embedding towards enhanced pre-trained language models. AI Open, 2:127–134, 2021b.

Fabian M Suchanek, Georgiana Ifrim, and Gerhard Weikum. Combining linguistic and statistical analysis to extract relations from web documents. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 712–717, 2006.

Dianbo Sui, Yubo Chen, Kang Liu, and Jun Zhao. Distantly supervised relation extraction in federated settings. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 569–583, 2021.

Le Sun and Xianpei Han. A feature-enriched tree kernel for relation extraction. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 61–67, 2014.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, S Yu Philip, and Lifang He. Mixup-transformer: Dynamic data augmentation for nlp tasks. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3436–3440, 2020a.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. CoLAKE: Contextualized language and knowledge embedding. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3660–3670, Barcelona, Spain (Online), December 2020b. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.327. URL https://aclanthology.org/2020.coling-main.327.

Xia Sun, Ke Dong, Long Ma, Richard Sutcliffe, Feijuan He, Sushing Chen, and Jun Feng. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. Entropy, 21(1):37, 2019.

Beth M Sundheim. Overview of the fourth message understanding evaluation and conference. Technical report, NAVAL COMMAND CONTROL AND OCEAN SURVEILLANCE CENTER RDT AND E DIV SAN DIEGO CA, 1992.

Mihai Surdeanu, David McClosky, Mason Smith, Andrey Gusev, and Christopher D Manning. Customizing an information extraction system to a new domain. In Proceedings of the ACL 2011 Workshop on Relational Models of Semantics, pages 2–10, 2011.

Thorvald A Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biol. Skar., 5:1–34, 1948.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. arXiv preprint arXiv:2203.10900, 2022a.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. Revisiting docred–addressing the overlooked false negative problem in relation extraction. arXiv preprint arXiv:2205.12696, 2022b.

Q. Tao, X. Luo, H. Wang, and R. Xu. Enhancing relation extraction using syntactic indicators and sentential contexts. In ICTAI, pages 1574–1580. IEEE, 2019.

Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. In The 2010 International Joint Conference on Neural Networks (IJCNN), pages 1–8, 2010. doi: 10.1109/IJCNN.2010.5596486.

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. Dependency-driven relation extraction with attentive graph convolutional networks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4458–4471, 2021.

Yuanhe Tian, Yan Song, and Fei Xia. Improving relation extraction through syntax-induced pre-training with dependency masking. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1875–1886, 2022.

Aryeh Tiktinsky, Vijay Viswanathan, Danna Niezni, Dana Meron Azagury, Yosi Shamay, Hillel Taub-Tabib, Tom Hope, and Yoav Goldberg. A dataset for n-ary relation extraction of drug combinations. arXiv preprint arXiv:2205.02289, 2022.

Ivan Tomek. Two modifications of cnn. IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(11):769–772, 1976. doi: 10.1109/TSMC.1976.4309452.

Van-Hien Tran, Van-Thuy Phi, Hiroyuki Shindo, and Yuji Matsumoto. Relation classification using segment-level attention-based cnn and dependency-based rnn. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2793–2798, 2019.

Vu Tran, Van-Hien Tran, Phuong Nguyen, Chau Nguyen, Ken Satoh, Yuji Matsumoto, and Minh Nguyen. Covrelex: A covid-19 retrieval system with relation extraction. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 24–31, 2021.

Shazia Usmani and Jawwad A Shamsi. News sensitive stock market prediction: literature review and suggestions. PeerJ Computer Science, 7:e490, 2021.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5):552–556, 2011.

Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. Journal of biomedical informatics, 45(5): 879–884, 2012.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. RESIDE: Improving distantly-supervised neural relation extraction using side information. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1257–1266, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1157. URL https://aclanthology.org/D18-1157.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

Mihaela Vela and Thierry Declerck. Concept and relation extraction in the finance domain. In Proceedings of the Eight International Conference on Computational Semantics, pages 346–350, 2009.

Vijay Viswanathan, Graham Neubig, and Pengfei Liu. Citationie: Leveraging the citation graph for scientific information extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 719–731, 2021.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. Commun. ACM, 57(10):78–85, September 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

Ngoc Thang Vu, Heike Adel, Pankaj Gupta, et al. Combining recurrent and convolutional neural networks for relation classification. In Proceedings of NAACL-HLT, pages 534–539, 2016.

Andra Waagmeester, Egon L Willighagen, Andrew I Su, Martina Kutmon, Jose Emilio Labra Gayo, Daniel Fernández-Álvarez, Quentin Groom, Peter J Schaap, Lisa M Verhagen, and Jasper J Koehorst. A protocol for adding knowledge to wikidata: aligning resources on human coronaviruses. BMC biology, 19(1):1–14, 2021.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. Zero-shot information extraction as a unified text-to-triple translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1225–1238, 2021a.

Hailin Wang, Ke Qin, Rufai Yusuf Zakari, Guoming Lu, and Jin Yin. Deep neural network-based relation extraction: an overview. Neural Computing and Applications, pages 1–21, 2022.

Haitao Wang, Tong Zhu, Mingtao Wang, Guoliang Zhang, and Wenliang Chen. A prior information enhanced extraction framework for document-level financial event extraction. Data Intelligence, pages 1–12, 2021b.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. Relation classification via multi-level attention CNNs. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1298–1307, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1123. URL https://www.aclweb.org/anthology/P16-1123.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, et al. Cord-19: The covid-19 open research dataset. In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, 2020a.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-adapter: Infusing knowledge into pre-trained models with adapters. CoRR, abs/2002.01808, 2020b. URL https://arxiv.org/abs/2002.01808.

Weijie Wang and Wenxin Hu. Improving relation extraction by multi-task learning. In Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence, pages 152–157, 2020.

William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2557–2563, 2015.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. arXiv preprint arXiv:1911.06136, 2019.

Neha Warikoo, Yung-Chun Chang, and Wen-Lian Hsu. Lptk: a linguistic pattern-aware dependency tree kernel approach for the biocreative vi chemprot task. Database, 2018, 2018.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382–6388, 2019.

Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. Relation extraction from clinical narratives using pre-trained language models. In AMIA Annual Symposium Proceedings, volume 2019, page 1236. American Medical Informatics Association, 2019.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew Arnold. Knowledge enhanced pretrained language models: A compreshensive survey. arXiv preprint arXiv:2110.08455, 2021.

Michael Wiegand, Benjamin Roth, Eva Lasarcyk, Stephanie Köser, and Dietrich Klakow. A gold standard for relation extraction in the food domain. In Proceedings of LREC, 2012.

Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, (3):408–421, 1972.

Limsoon Wong. Pies, a protein interaction extraction system. In Biocomputing 2001, pages 520–531. World Scientific, 2000.

Fei Wu and Daniel S Weld. Open information extraction using wikipedia. In Proceedings of the 48th ACL, pages 118–127, 2010.

Haoyu Wu, Qing Lei, Xinyue Zhang, and Zhengqian Luo. Creating a large-scale financial news corpus for relation extraction. In 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), pages 259–263. IEEE, 2020.

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. Graph neural networks for natural language processing: A survey. arXiv preprint arXiv:2106.06090, 2021.

Shanchan Wu and Yifan He. Enriching pre-trained language model with entity information for relation classification. In Proceedings of ACM CIKM'19, page 2361–2364, 2019.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 833–844, 2019.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 120–130, 2020.

Ye Wu, Ruibang Luo, Henry Leung, Hing-Fung Ting, and Tak-Wah Lam. Renet: A deep learning approach for extracting gene-disease associations from literature. In International Conference on Research in Computational Molecular Biology, pages 272–284. Springer, 2019.

Bolun Xia, Vipula D Rawte, Mohammed J Zaki, Aparna Gupta, et al. Fetilda: An effective framework for fin-tuned embeddings for long financial text documents. arXiv preprint arXiv:2206.06952, 2022.

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. Revisiting the negative data of distantly supervised relation extraction. arXiv preprint arXiv:2105.10158, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems, 33, 2020.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. Data noising as smoothing in neural network language models. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. URL https://openreview.net/forum?id=H1VyHY9gg.

Rui Xing, Jie Luo, and Tengwei Song. Biorel: towards large-scale biomedical relation extraction. BMC bioinformatics, 21(16):1–13, 2020.

Chen Xingyue, Ni Liping, and Ni Zhiwei. Financial event extraction based on electra and part-of-speech. Data Analysis and Knowledge Discovery, page 1, 2021.

Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. In Proceedings of the 26th international conference on world wide web, pages 1271–1279, 2017.

Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, and Yuxia Song. Mining comparative opinions from customer reviews for competitive intelligence. Decision Support Systems, 50(4): 743–754, 2011. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2010.08.021. URL https://www.sciencedirect.com/science/article/pii/S0167923610001454. Enterprise Risk and Security Management: Data, Text and Web Mining.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. Semantic relation classification via convolutional neural networks with simple negative sampling. arXiv preprint arXiv:1506.07650, 2015a.

Rong Xu, Li Li, and QuanQiu Wang. Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. Bioinformatics, 29 (17):2186–2194, 2013.

Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14167–14175, 2021.

Wang Xu, Kehai Chen, Lili Mou, and Tiejun Zhao. Document-level relation extraction with sentences importance estimation and focusing. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2920–2929, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.212. URL https://aclanthology.org/2022.naacl-main.212.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency paths. In proceedings of the 2015 conference on empirical methods in natural language processing, pages 1785–1794, 2015b.

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. Improved relation classification by deep recurrent neural networks with data augmentation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1461–1470, 2016.

Shweta Yadav, Srivastsa Ramesh, Sriparna Saha, and Asif Ekbal. Relation extraction from biomedical and clinical text: Unified multitask learning framework. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020.

Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In 27th International Conference on Computational Linguistics, COLING 2018, pages 2145–2158. Association for Computational Linguistics (ACL), 2018.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 23–30, Online, October 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.4. URL https://aclanthology.org/2020.emnlp-demos.4.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: deep contextualized entity representations with entity-aware self-attention. arXiv preprint arXiv:2010.01057, 2020b.

Ayana Yamamoto, Yuichi Miyamura, Kouta Nakata, and Masayuki Okamoto. Company relation extraction from web news articles for analyzing industry structure. In 2017 IEEE 11th International Conference on Semantic Computing (ICSC), pages 89–92. IEEE, 2017.

Chenwei Yan, Xiangling Fu, Weiqiang Wu, Shilun Lu, and Ji Wu. Neural network based relation extraction of enterprises in credit risk management. In 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), pages 1–6. IEEE, 2019.

Hu Yanan. Cross lingual relation extraction via machine translation. Journal of Chinese Information Processing, 27(5):191–197, 2013.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. Enhancing cross-lingual transfer by manifold mixup. In International Conference on Learning Representations, 2021.

Nian Yang, Dongxin Shi, and Yan Hua. Bidirectional gated recurrent unit neural networks for relation extraction of chinese enterprises. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), volume 1, pages 1539–1543, 2020a. doi: 10.1109/ITNEC48623.2020.9084718.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097, 2020b.

Xuchen Yao and Benjamin Van Durme. Information extraction over structured data: Question answering with freebase. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 956–966. Association for Computational Linguistics, 2014. doi: 10.3115/v1/P14-1090. URL http://aclweb.org/anthology/P14-1090.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In Proceedings of ACL, pages 764–777, 2019.

Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. Codred: A cross-document relation extraction dataset for acquiring knowledge in the wild. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4452–4472, 2021.

Wei Ye, Bo Li, Rui Xie, Zhonghao Sheng, Long Chen, and Shikun Zhang. Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data. In Proceedings of ACL, pages 1351–1360, 2019.

Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. ACL - Association for Computational Linguistics, July 2015. URL https://www.microsoft.com/en-us/research/publication/semantic-parsing-via-staged-query-graph-generation-question-answering-with-knowledge-base/. Outstanding Paper Award.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541, 2018.

Bei Yu, Yingya Li, and Jun Wang. Detecting causal language use in science findings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4664–4674, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1473. URL https://aclanthology.org/D19-1473.

Junjie Yu, Tong Zhu, Wenliang Chen, Wei Zhang, and Min Zhang. Improving relation extraction with relational paraphrase sentences. In Proceedings of the 28th International Conference on Computational Linguistics, pages 1687–1698, 2020.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. Dwie: An entity-centric dataset for multi-task document-level information extraction. Information Processing & Management, 58(4):102563, 2021.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2335–2344, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-1220.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 1753–1762, 2015.

Ce Zhang. DeepDive: a data management system for automatic knowledge base construction. PhD thesis, The University of Wisconsin-Madison, 2015.

Dongxu Zhang and Dong Wang. Relation classification via recurrent neural network. arXiv preprint arXiv:1508.01006, 2015.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017a.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning, pages 11328–11339. PMLR, 2020a.

Min Zhang, Jie Zhang, and Jian Su. Exploring syntactic features for relation extraction using a convolution tree kernel. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, pages 288–295, 2006.

Qingchuan Zhang, Menghan Li, Wei Dong, Min Zuo, Siwei Wei, Shaoyi Song, and Dongmei Ai. An entity relationship extraction model based on bert-blstm-crf for food safety domain. Computational Intelligence and Neuroscience, 2022, 2022.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia conference on language, information and computation, pages 73–78, 2015a.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. Advances in neural information processing systems, 28:649–657, 2015b.

Xingxing Zhang, Jianwen Zhang, Junyu Zeng, Jun Yan, Zheng Chen, and Zhifang Sui. Towards accurate distant supervision for relational facts extraction. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 810–815, 2013.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. A hybrid model based on neural networks for biomedical relation extraction. Journal of biomedical informatics, 81:83–92, 2018a.

Yijia Zhang, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. Drug–drug interaction extraction via hierarchical rnns on sequence and shortest dependency paths. Bioinformatics, 34(5):828–835, 2018b.

Yue Zhang, Hongliang Fei, and Ping Li. Readsre: Retrieval-augmented distantly supervised relation extraction. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2257–2262, 2021.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 35–45, Copenhagen, Denmark, September 2017b. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL https://www.aclweb.org/anthology/D17-1004.

Yuhao Zhang, Peng Qi, and Christopher D Manning. Graph convolution over pruned dependency trees improves relation extraction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2205–2215, 2018c.

Z. Zhang, X. Shu, B. Yu, T. Liu, J. Zhao, Q. Li, and L. Guo. Distilling knowledge from well-informed soft labels for neural relation extraction. In Proceedings of the AAAI Conference, pages 9620–9627, 2020b.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In Proceedings of ACL, pages 1441–1451, 2019.

Di Zhao, Jian Wang, Yijia Zhang, Xin Wang, Hongfei Lin, and Zhihao Yang. Incorporating representation learning and multihead attention to improve biomedical cross-sentence n-ary relation extraction. BMC bioinformatics, 21(1):1–17, 2020.

Jie Zhao, Peiquan Jin, and Yanhong Liu. Business relations in the web: Semantics and a case study. Journal of Software, 5(8):826–833, 2010.

Jun Zhao, Frank van Harmelen, Jie Tang, Xianpei Han, Quan Wang, and Xianyong Li, editors. Knowledge Graph and Semantic Computing. Knowledge Computing and Language Understanding - Third China Conference, CCKS 2018, Tianjin, China, August 14-17, 2018, Revised Selected Papers, volume 957 of Communications in Computer and Information Science, 2019. Springer.

Sendong Zhao, Chang Su, Zhiyong Lu, and Fei Wang. Recent advances in biomedical literature mining. Briefings in Bioinformatics, 22(3):bbaa057, 2021.

Wei Zheng, Hongfei Lin, Ling Luo, Zhehuan Zhao, Zhengguang Li, Yijia Zhang, Zhihao Yang, and Jian Wang. An attention-based effective neural model for drug-drug interactions extraction. BMC bioinformatics, 18(1):1–11, 2017.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 427–434, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219893. URL https://www.aclweb.org/anthology/P05-1053.

Guodong Zhou, Min Zhang, DongHong Ji, and Qiaoming Zhu. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 728–736, 2007.

Kai Zhou, Xiangfeng Luo, Hao Wang, and Richard Xu. Multi-task learning for relation extraction. In 31st International Conference on Tools with Artificial Intelligence (ICTAI), pages 1480–1487. IEEE, 2019.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 207–212. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-2034. URL http://aclweb.org/anthology/P16-2034.

Wenxuan Zhou and Muhao Chen. An improved baseline for sentence-level relation extraction. arXiv preprint arXiv:2102.01373, 2021.

Wei Zhu. Autorc: Improving bert based relation classification models via architecture search. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 33–43, 2021.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27, 2015.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. Adversarial feature adaptation for cross-lingual relation classification. In Proceedings of the 27th International Conference on Computational Linguistics, pages 437–448, 2018.

Zhe Zuo, Michael Loster, Ralf Krestel, and Felix Naumann. Uncovering business relationships: Context-sensitive relationship extraction for difficult relationship types. In Michael Leyer, editor, Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings, Rostock, Germany, September 11-13, 2017, volume 1917 of CEUR Workshop Proceedings, page 271. CEUR-WS.org, 2017. URL http://ceur-ws.org/Vol-1917/paper36.pdf.