Extracting Clinical Event Timelines : Temporal Information Extraction and Coreference Resolution in Electronic Health Records

Julien Tourille

université PARIS-SACLAY

Limsi

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

NNT : 2018SACLS603

Extracting Clinical Event Timelines: Temporal Information Extraction and Coreference Resolution in Electronic Health Records

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Orsay, le 18 décembre 2018, par

JULIEN TOURILLE

Composition du Jury :

Claire Nédellec Directrice de Recherche, INRA (MaIAGE)	Présidente
Philippe Muller Maître de Conférences HDR, Université Paul Sabatier (IRIT)	Rapporteur
Mathieu Roche Chercheur HDR, CIRAD (TETIS)	Rapporteur
Sandra Bringay Professeure, Université de Montpellier 3 (LIRMM)	Examinatrice
Guergana Savova Associate Professor, Harvard Medical School et Boston Children's Hospital Informatics Program	Examinatrice
Xavier Tannier Professeur, Sorbonne Université (LIMICS)	Directeur de thèse
Olivier Ferret Chercheur HDR, CEA (LIST)	Encadrant de thèse
Aurélie Névéol Chargée de Recherche, CNRS (LIMSI)	Encadrante de thèse

Thèse de doctorat

Abstract

Important information for public health is contained within electronic health records. Being able to extract this information automatically would allow to improve the daily medical care and to foster advances in clinical research. However, the large majority of it remains locked in documents written in natural language.

Among all the clinical aspects that are of interest in these records, the patient timeline is one of the most important. Being able to retrieve clinical timelines would allow for a better understanding of some clinical phenomena such as disease progression and longitudinal effects of medications. It would also allow to improve medical question answering and clinical outcome prediction systems. Accessing the clinical timeline is needed to evaluate the quality of the healthcare pathway by comparing it to clinical guidelines, and to highlight the steps of the pathway where specific care should be provided.

In this thesis, we focus on building such timelines by addressing two related Natural Language Processing (NLP) topics which are temporal information extraction and clinical event coreference resolution.

Beyond the obvious necessity of tackling both issues at the same time, we could argue that both topics are interdependent. Intuitively, two event mentions that are coreferent would share the same temporal information. This has a number of implications when considering that tasks are done in a specific order (coreference resolution *then* temporal information extraction). Information brought by the coreference links could be used to increase the performance of a temporal information extraction approach.

Conversely, two event mentions that share the same temporal location and the same meaning have high chances to be coreferent. The additional temporal information that would be available to a coreference resolution system could be valuable and improve performance.

Based on these observations, it becomes clear that temporal information extraction and clinical event coreference resolution need to be handled jointly not only because they are needed for clinical timeline extraction but also because there are complementary.

Temporal information extraction is a complex subject that requires carefully designed annotated corpora. These resources are mostly in English and the approaches that are developed in the community are biased toward that language. This motivates our first research question: is it possible to devise a generic approach for temporal information extraction that could be used in different languages?

These corpora are often packing a rich set of categorical features describing the annotated entities. Moreover, there is a large number of clinical text preprocessing tools that allow to add more information about these entities. Given this context, how could this diversity of categorical features be used in neural approaches? This brings a follow-up question: how does the use of these categorical features impact the performance of such approaches?

Coreference resolution is an active topic in the clinical domain. As we mentioned above, both temporal information extraction and coreference resolution are interlinked. This leads

to our fourth research question: what sort of temporal information could be useful for coreference resolution? Coreference resolution approaches are complex and it is very hard to improve the performance of simple vanilla models. In this context, how could this temporal information be integrated in a neural approach for coreference resolution?

To answer these questions, we present several contributions. We devise a feature-based approach for temporal relation extraction. We evaluate our system on an English corpus. Then, we perform an empirical evaluation on two corpora of documents written in English and in French and showed that a similar approach can be used for both languages.

Our next contribution consists in a hybrid approach for end-to-end temporal information extraction that incorporates categorical features. We perform an empirical study on how categorical features impact the performance of our neural-based approach. Then, we evaluate our approach on an English dataset.

Concerning coreference resolution, we devise a neural-based approach for coreference resolution in the clinical domain inspired by recent approaches in the general domain. We devise a temporal feature and evaluate its contribution in the context of an empirical study which aimed at measuring how categorical features and neural network components such as attention mechanisms and token character-level representations influence the performance.

In addition to the above-mentioned contributions, we contribute to the NLP community by devising two resources. First, we convert the i2b2 task1c corpus to the CoNLL format. This transformation could foster NLP research on coreference resolution in the clinical domain by allowing a better reproduction of results and by facilitating corpus processing. Second, we pack our neural sequence labeling module into an open source tool called Yet Another SEquence Tagger (YASET) that can be used for any NLP sequence labeling tasks.

Résumé

Les dossiers patients électroniques contiennent des informations importantes pour la santé publique. Être en mesure d'extraire automatiquement ces informations permettrait d'améliorer les soins médicaux et de soutenir la recherche clinique. Cependant, la majeure partie de ces informations est contenue dans des documents rédigés en langue naturelle.

Parmi toutes les informations cliniques intéressantes présentes dans ces dossiers, la chronologie médicale du patient est l'une des plus importantes. Être capable d'extraire automatiquement cette chronologie permettrait d'acquérir une meilleure connaissance de certains phénomènes cliniques tels que la progression des maladies et les effets à long-terme des médicaments. De plus, cela permettrait d'améliorer la qualité des systèmes de question–réponse et de prédiction de résultats cliniques. Par ailleurs, accéder aux chronologies médicales est nécessaire pour évaluer la qualité du parcours de soins en le comparant aux recommandations officielles et pour mettre en lumière les étapes de ce parcours auxquelles une attention particulière doit être portée.

Dans notre thèse, nous nous concentrons sur la création de ces chronologies médicales en abordant deux questions connexes en Traitement Automatique des Langues (TAL) : l'extraction d'informations temporelles et la résolution de la coréférence dans des documents cliniques.

Au-delà de la nécessité évidente d'aborder les deux tâches en même temps, nous soutenons que les deux sujets sont interdépendants. Intuitivement, deux événements médicaux qui sont coréférents partagent les mêmes informations temporelles. Cette observation a des implications si l'on considère l'ordre dans lequel les tâches sont effectuées (le résolution de la coréférence puis l'extraction d'informations temporelles). L'information apportée par la présence de liens de coréférence entre les événements médicaux pourrait être utilisée pour améliorer la performance d'une approche d'extraction d'informations temporelles.

Inversement, deux mentions d'événements qui partagent les mêmes informations temporelles et la même signification ont une grande chance d'être coréférentes. Cette information temporelle pourrait permettre d'améliorer la qualité d'un système de résolution de la coréférence.

Sur la base de ces observations, il apparaît clairement que l'extraction d'informations temporelles et la résolution de la coréférence sont des tâches qui doivent être étudiées conjointement, non seulement car elles sont nécessaires à l'extraction de chronologies médicales, mais aussi car elles sont complémentaires.

L'extraction d'informations temporelles est un sujet complexe qui nécessite des corpus soigneusement annotés. Ces ressources sont principalement rédigées en anglais et les approches élaborées dans la communauté sont biaisées en faveur de cette langue. De cette observation découle notre première question de recherche : est-il possible d'élaborer une approche générique pour l'extraction d'informations temporelles qui pourrait être utilisée pour différentes langues ?

Ces corpus contiennent souvent un ensemble de traits catégoriels décrivant les entités annotées. De plus, il existe un grand nombre d'outils de prétraitement des textes cliniques qui permettent d'ajouter des informations sur ces entités. Dans ce contexte, comment cette diversité de traits catégoriels pourrait-elle être utilisée dans le cadre des approches neuronales ? Par ailleurs, comment l'utilisation de ces traits catégoriels influe-t-elle sur la performance de ces approches ?

La résolution de la coréférence est un sujet de recherche actif en TAL clinique. Comme nous l'avons mentionné précédemment, l'extraction d'informations temporelles et la résolution de la coréférence sont des sujets interdépendants. Cela nous amène à notre quatrième question de recherche : quelle sorte d'informations temporelles pourrait être utile pour la résolution de la coréférence ? Les approches développées pour la résolution de la coréférence sont souvent complexes et il est très difficile d'améliorer significativement la performance de modèles simples. Dans ce contexte, comment cette information temporelle pourrait-elle être intégrée dans une approche neuronale pour la résolution de la coréférence ?

Pour répondre à ces questions, nous présentons plusieurs contributions. Nous concevons une approche à base de traits pour l'extraction des relations temporelles. Nous évaluons notre approche sur un corpus de documents écrits en anglais. Ensuite, nous effectuons une évaluation empirique sur deux corpus de documents rédigés en anglais et en français et nous montrons qu'une approche similaire peut être utilisée pour les deux langues.

Notre seconde contribution consiste en une approche hybride pour l'extraction d'informations temporelles qui incorpore des traits catégoriels. Nous effectuons une étude empirique sur la façon dont ces traits affectent la performance de notre approche. Ensuite, nous évaluons notre approche sur un corpus de documents écrits en anglais.

En ce qui concerne la résolution de la coréférence, nous concevons une approche neuronale inspirée par les travaux récents dans le domaine général. Nous concevons un trait temporel et évaluons sa contribution dans le cadre d'une étude empirique visant à mesurer l'impact des traits catégoriels et des différents composants habituellement utilisés dans les approches neuronales tels que les mécanismes d'attention et les représentations au niveau des caractères.

En plus des contributions susmentionnées, nous convertissons le corpus i2b2 task1c au format CoNLL. Cette transformation pourrait favoriser la recherche sur la résolution de la coréférence dans le domaine clinique en permettant une meilleure reproduction des résultats et en facilitant l'utilisation du corpus. Deuxièmement, nous empaquetons notre module d'étiquetage de séquences dans un outil open-source appelé YASET qui peut être utilisé dans des tâche d'étiquetage de séquence en TAL.

Remerciements

Je souhaite commencer par remercier mes encadrants de thèse, Aurélie, Olivier et Xavier pour leur soutien sans faille et leurs conseils avisés. Merci d'avoir accepté d'encadrer mon travail, pour toutes ces discussions que nous avons eu semaine après semaine et de m'avoir donné goût à la recherche.

Je remercie les membres de mon jury de thèse. Merci à Philippe Muller et Mathieu Roche d'avoir accepté d'être rapporteurs de ma thèse et à Sandra Bringay, Claire Nédellec et Guergana Savova d'avoir accepté de l'évaluer.

Je remercie Guergana, Tim, Chen et toute l'équipe du Boston's Children Hospital de m'avoir accueilli pendant deux mois dans leur laboratoire. Cette escapade américaine a été enrichissante, pleine de bons souvenirs et m'a permis d'initier des liens amicaux et scientifiques durables.

La vie d'un laboratoire c'est un va-et-vient continu de gens tous plus intéressants les uns que les autres. Un grand merci à tous les collègues de passage au LIMSI, pour toutes ces discussions passionnantes sur la science et le reste qui ont animé nos pauses cafés et nos pauses déjeuners. Merci aux anciens (José, Vincent, Romain, Marine, Arthur, Pierre, Charlotte, Lucie et Eva) et aux nouveaux (Christopher, Anna, Antoine, Léon, Nicolas, Arnaud, Yuming, Rachel, Rashedur, Swen, Zheng, Sanjay et Leonardo). Mention spéciale à mes camarades de promotion, Arnaud et Rachel; on l'a fait, on y est! Un grand merci aussi à Billy, pour ses avis nuancés et ses propos raffinés sur l'actualité et la vie en général.

Merci aux permanents du groupe ILES pour maintenir la cohésion de tout ce joyeux groupe. Merci à Anne pour toutes ses casquettes et à Pierre pour l'animation du groupe. Mention spéciale à Patrick pour les discussions incroyables que l'on a eu et pour le partage de son savoir sans fond. Merci aussi à Michael, Brigitte, Cyril, Thomas, Anne-Laure, Aurélie, Sophie et les autres pour leurs conseils, leurs oreilles attentives et nos discussions autour d'un café.

Ce travail n'aurait pas été possible sans un soutien administratif et logistique. Merci à Olivier, Jean-Claude et les autres pour leur aide et leur gestion des machines qui ont fait tourner mes expériences! Merci à Bénédicte, Pascal, Carole, Isabelle, Sophie, Laurence et les autres pour faire tourner la machine LIMSI sans accrocs.

Merci aussi aux amis de ch'nord que j'ai quitté pour la recherche. Merci aux runners du club de Massy de m'avoir aidé à me changer les idées. Merci à ma famille de m'avoir accompagné durant toutes ces années d'études. Merci à ma belle famille pour leur soutien.

Enfin, merci à Léa, sans qui je ne rédigerais pas ces remerciements. Merci d'avoir supporté mes humeurs dans les derniers mois et de m'avoir accompagné dans ce voyage difficile.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Important information for public health is contained within Electronic Health Records (EHRs). The vast majority of clinical data available in these records takes the form of narratives written in natural language. Although free text is convenient to describe complex medical concepts, it is difficult to use for medical decision support, clinical research or statistical analysis. The need to access information combined to the rapid growth of EHRs led to the development of NLP approaches tailored for the clinical domain.

Information Extraction (IE) has been applied successfully to a variety of tasks in the clinical domain over the last decades. Various examples of such applications can be found in the literature. One example is the automatic extraction of codes from clinical text. It consists usually in assigning codes dealing with diagnoses, such as International Classification of Diseases (ICD) codes. Several datasets were released to the community to foster NLP research on this topic. For instance, Pestian et al. (2007) offer to work on radiology reports while Névéol et al. (2016) release a large dataset for ICD-10 coding of death certificates.

Surveillance is also an important area of research. One valuable application is adverse event detection (Velupillai et al. 2015a). These events may be related to medical procedures (Penz et al. 2007) or drugs (Wang et al. 2009). Another use-case scenario is syndromic surveillance which consists in monitoring patient records to spot disease outbreaks (Meystre et al. 2008) or hospital-acquired infections (Velupillai et al. 2015a).

Another active area of research is EHR enrichment for computerized decision support. Meystre et al. (2008) identify several sub-areas: automatic structuring of documents which consists in converting free text by segmenting and re-arranging them according to a template, automatic summarization which allows to have a concise view of clinical narratives, and case finding which allows to search for a specific patient according to some criteria. Also, the i2b2

initiative[1] led several corpus annotation efforts to foster NLP research. It includes the annotation of patient smoking status (Uzuner et al. 2008) obesity and comorbidities (Uzuner 2009), medication information (Uzuner et al. 2010) and concepts, assertions and relations (Uzuner et al. 2011).

All these research efforts would not have been possible without access to data. As clinical narratives contain highly sensitive personal information about the patients and their conditions, one requirement is that clinical narratives must be properly de-identified. Automatic de-identification of clinical narratives has been addressed in many research efforts and is still an active area of research (Grouin and Névéol 2014; Neamatullah et al. 2008; Stubbs and Uzuner 2015; Uzuner et al. 2007).

Devising a comprehensive list of NLP applications to the clinical domain is a difficult task. For a detailed overview of such applications, we refer the reader to literature surveys on the topic (Meystre et al. 2008; Velupillai et al. 2015a; Wang et al. 2009).

Among all the clinical aspects that are of interest in these records, the patient timeline (Figure 1.1) is one of the most important. Being able to retrieve clinical timelines would allow for a better understanding of some clinical phenomena such as disease progression and longitudinal effects of medications (C. Lin et al. 2016a; W. Sun et al. 2013c). It would also allow to improve medical question answering and clinical outcome prediction systems. Accessing the clinical timeline is needed to evaluate the quality of the healthcare pathway by comparing it to clinical guidelines, and to highlight the steps of the pathway where specific care should be provided.

In this thesis, we focus on building such timelines by addressing two related NLP topics which are **temporal information extraction** and **clinical event coreference resolution**.

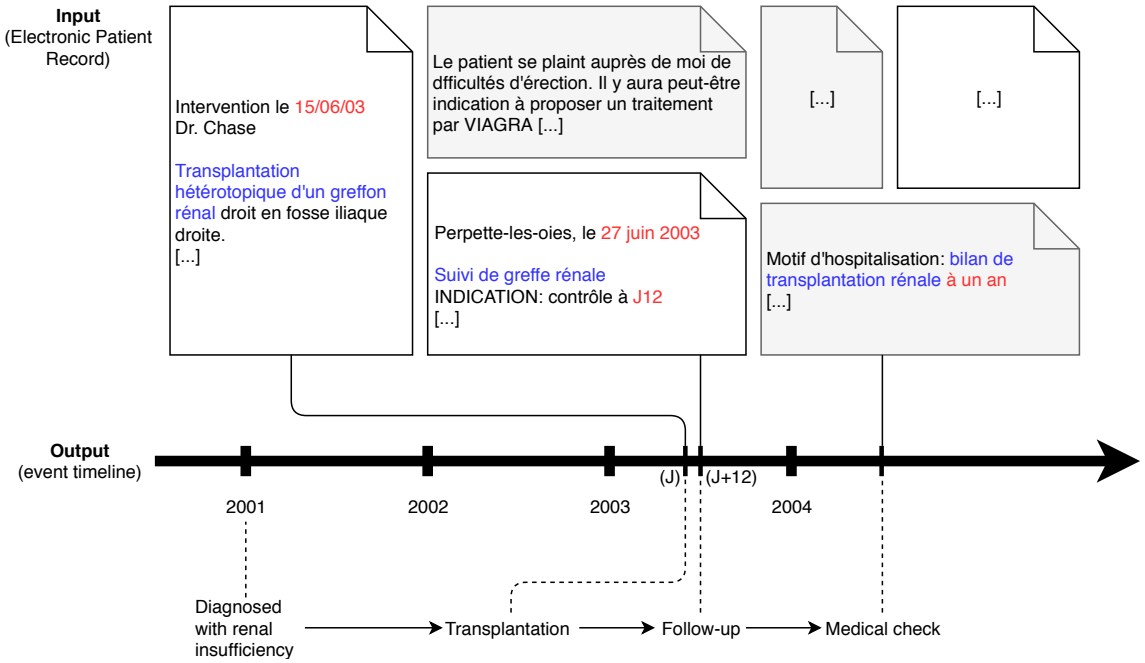

Figure 1.1: Example of timeline extracted from an electronic patient record.

1. https://www.i2b2.org

1.1 Temporal Information Extraction

Intuitively, building clinical timelines requires to extract relevant temporal information from text. Temporal information extraction has already a long history in NLP but was mainly addressed in the news domain. The first research efforts on the topic can be tracked down to the Message Understanding Conferences (MUCs) (Sundheim 1993). At that time, the scope of the effort was limited to extracting dates and times from news reports. A major milestone was reached in 2005 with the creation of the first full temporal annotation scheme called TimeML (Pustejovsky et al. 2005). This research effort allows to model precisely temporal information in text and has influenced most of the following research up to this day. The clinical domain started to develop an interest in temporal information in the early 2010's with the release of the i2b2 corpus (W. Sun et al. 2013b). The large majority of temporally annotated datasets in both the general and the clinical domains contain documents written in English. However, there are several efforts aiming at devising resources in other languages. For instance, Campillos et al. (2018) recently annotated a clinical corpus of documents written in French.

Temporal information extraction in the clinical domain can be broken down into two main steps. The first implies the extraction of *clinical event* mentions and *temporal expressions* from clinical narratives. The second involves the extraction of temporal relations. Example 1 contains one clinical event (*neck pain*) and one temporal expression (*July*) that are linked with a temporal relation (*begins-on*). Definition of events, temporal expressions and temporal relations vary across corpora but are often derived from the TimeML specification. Following this two-step process, temporal information acquired from documents can be aggregated into a timeline.

(1) She has been experiencing [$_{\text{EVENT}}$ **neck pain**] since [$_{\text{DATE}}$ **July**].
→ neck pain BEGINS-ON July

Feature-based approaches for automatic temporal information extraction have progressively been replaced by neural approaches over the last five years. However, initial hopes for a major performance improvement have rapidly faded as classical feature-based approaches still remain competitive (Bethard et al. 2015; Bethard et al. 2016; Bethard et al. 2017). There is an on-going effort in the NLP community for combining both approach types. Performance has been improved by including categorical features in neural network approaches. For instance, Dligach et al. (2017) used Part-Of-Speech (POS) tags as features for temporal relation extraction in clinical narratives.

1.2 Coreference Resolution

Another linguistic phenomenon to consider when building a patient timeline is coreference. Clinical reports contain multiple mentions of the same clinical events due to the fact that the medical staff needs to follow-up on these events. Intuitively, this phenomenon brings noise when building clinical timelines and needs to be handled appropriately.

Similarly to temporal information extraction, coreference resolution has been addressed primarily in the news domain and the first systematic evaluations can be traced back to the same conferences (Hirschman and Chinchor 1998; Sundheim 1995). In the clinical

domain, the i2b2 initiative is once again responsible for the diffusion of the first clinical dataset annotated with coreference in 2012 (Uzuner et al. 2012).

Coreference resolution in the clinical domain can be divided into two main steps. In the first one, potential coreferring text spans must be extracted. Usually, these elements fall into specific categories such as *medication* or *medical procedure*. Once these mentions have been extracted, one must figure out which ones are referring to the same real-world clinical events. In Example 2, all bracketed mentions are referring to the same real-world medical problem.

(2) The CXR revealed [**8 mm obstructing stone**] ... [**The renal stone**] was considered to be the cause of patient's symptoms ... We recommended surgical procedure to remove [**ureteropelvic stone**] ...

Research efforts for coreference resolution in the clinical domain implement feature-based approaches (Uzuner et al. 2012). However, similarly to temporal information extraction, neural approaches have become increasingly popular in the general domain (Lu and Ng 2018). Moreover, we note that hybrid approaches including categorical features in neural networks have allowed significant performance improvement. For instance, K. Lee et al. (2017) encoded speaker information, text genre and other features as dense representations in their neural-based coreference resolution approach.

1.3 Topic Interdependence

As we mentioned above, temporal information extraction and coreference resolution are two of the main topics that need to be addressed when considering the task of building clinical timelines. Event mentions need to be extracted and placed in time. Simultaneously, coreferring mentions need to be regrouped as they are referring to the same real-world event.

Beyond the obvious necessity of tackling both issues at the same time, we could argue that both topics are interdependent. Intuitively, two event mentions that are coreferent would share the same temporal information. This has a number of implications when considering that tasks are performed in a specific order (coreference resolution *then* temporal information extraction). Information brought by the coreference links could be used to increase the performance of a given temporal information extraction approach.

Conversely, two event mentions that share the same temporal location and the same meaning have high chances to be coreferent. The additional temporal information that would be available to a coreference resolution system could be valuable and bring a performance improvement.

Based on these observations, it becomes clear that temporal information extraction and clinical event coreference resolution need to be handled jointly not only because they are needed for clinical timeline extraction but also because there are somehow complements.

1.4 Research Questions

Temporal information extraction is a complex subject that needs carefully designed annotated corpora. These resources are mostly in English and the approaches that are developed

in the community are biased toward that language. This motivates our first research question: **is it possible to devise a generic approach for temporal information extraction that could be used in different languages?**

These corpora are often packing a rich set of categorical features describing the annotated entities. Moreover, there is a large number of clinical text preprocessing tools that allow to add more information about these entities. Given this context, **how could this diversity of categorical features be used in neural approaches?** This brings a follow-up question: **how does the use of these categorical features impact the performance of such approaches?**

Coreference resolution is an active topic in the clinical domain. As we mentioned above, both temporal information extraction and coreference resolution are interlinked. This leads to our fourth research question: **what sort of temporal information could be useful for coreference resolution?** Coreference resolution approaches are complex and it is very hard to improve the performance of simple vanilla models. In this context, **how could this temporal information be integrated into a neural approach for coreference resolution?**

1.5 Contributions

In the first part of this thesis, we address temporal information extraction from clinical narratives. We present four contributions to the topic:

- **A feature-based approach for narrative container extraction from clinical narratives**. We devise a competitive feature-based approach for temporal relation extraction. We test our system on an English dataset in the context of the 2016 edition of the Clinical TempEval shared task (Bethard et al. 2016).

- **An abstraction of our feature-based approach**. We perform an empirical evaluation on two corpora of documents written in English and in French. We show that a similar approach can be used for both languages.

- **A neural approach for end-to-end temporal information extraction that incorporates classical categorical features**. We address event, temporal expression and temporal relation extraction in clinical narratives. We test our approach on an English dataset in the context of the 2017 edition of the Clinical TempEval shared task (Bethard et al. 2017).

- **An empirical study on how categorical features impact the performance of our neural-based approach**. We use gold features available in the dataset but also predicted features obtained via the use of clinical text preprocessing tools.

The second part of this thesis is dedicated to clinical event coreference resolution. We present two contributions to the topic:

- **A neural-based approach for coreference resolution in the clinical domain** inspired by recent approaches in the general domain. We address coreference resolution on both gold mentions and predicted mentions.

- **A temporal feature derived from the temporal relation that exists between events and DCTs**. We test this feature in the context of an empirical study which aims at measuring how categorical features and neural network components such as attention mechanisms and token character-level representations influence the performance.

In addition to the above-mentioned contributions, we contribute to the community by devising two resources. First, **we convert the i2b2 task1c corpus to the CoNLL format**. This transformation could foster NLP research on the topic by allowing a better reproduction of results and by facilitating corpus processing. Second, we pack our neural sequence labeling module into a **open source tool called YASET** than can be used for any NLP sequence labeling tasks.

1.6 Outline

The thesis is divided into two parts. The first one is composed of three chapters and is related to temporal information extraction. We start by reviewing the literature on the topic and we present our contributions. The second part is composed of two chapters and is related to clinical event coreference resolution. Similarly, we begin with a survey of the literature on the topic followed by a presentation of our contributions.

Part I – Temporal Information Extraction from Clinical Narratives

- **Chapter 2 – It's About Time: Temporal Information Extraction from Text**. In this chapter, we review the different aspects of time from both linguistic and computational linguistic perspectives. We address the definition of the three primitives of time: events, times and temporal relations. We survey the different approaches that have been devised for temporal information extraction from text in both the general and the clinical domains.

- **Chapter 3 – Feature-Based Approach for Temporal Relation Extraction: Application to French and English Corpora**. We present a feature-based approach for temporal information extraction in clinical narratives. We evaluate our approach in the context of the 2016 edition of the Clinical TempEval shared task. Then, we investigate whether the same generic approach can be applied on two different languages. We evaluate our approach on two datasets of clinical documents written in English and French.

- **Chapter 4 – Neural Approach for Temporal Information Extraction**. We present a neural-based approach for temporal information extraction. We devise a neural architecture for clinical event, temporal expression and temporal relation extraction. We investigate how categorical features impact the performance of our model. Then, we evaluate our approach in the context of the 2017 edition of the Clinical TempEval shared task.

Part II – Clinical Event Coreference Resolution

- **Chapter 5 – Clinical Event Coreference Resolution**. This chapter presents an overview of the current state of the literature related to coreference. We present how the notion of coreference is handled in linguistics and computational linguistics. Specifically, we address the differences that exist between event coreference resolution in the general and in the clinical domains. We review the supervised machine learning approaches that have been devised in the literature.

- **Chapter 6 – Neural Entity-Based Approach for Coreference Resolution in the Clinical Domain**. We present a neural entity-based approach for coreference resolution inspired by recent efforts in the domain. We devise a temporal feature that we integrate into our approach. We perform an empirical study on how usual neural network components such as attention mechanisms or character-level representations impact the performance of our approach.

1.7 Published Work

The material presented in Chapter 3 is based on three publications: one at the 2016 edition of the SemEval workshop (Tourille et al. 2016b), one at the 2016 edition of the TALN conference (Tourille et al. 2016a) and one at the 2017 edition of the EACL conference (Tourille et al. 2017c).

The material presented in Chapter 4 is based on two publications: one at the 2017 edition of the Clinical TempEval workshop (Tourille et al. 2017a) and one at the 2017 edition of the ACL conference (Tourille et al. 2017b).

The presentation of our sequence labeling tool will be published in the proceedings of the 2018 edition of the LOUHI workshop (Tourille et al. 2018).

Publication List

1. Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier (June 2016b). "LIMSI-COT at SemEval-2016 Task 12: Temporal Relation Identification Using a Pipeline of Classifiers". In: *Proceedings of the 10th International Workshop on Semantic Evaluation* (San Diego, California, USA, June 16, 2016–June 17, 2016). Association for Computational Linguistics, pp. 1136–1142

2. Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier (July 2016a). "Extraction de Relations Temporelles dans des Dossiers Électroniques Patient". In: *Actes de la Conférence Traitement Automatique des Langues Naturelles 2016* (Paris, France, July 4, 2016–July 8, 2016). Association pour le Traitement Automatique des Langues, pp. 459–466

3. Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol (Apr. 2017c). "Temporal Information Extraction from Clinical Text". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Valencia, Spain, Apr. 3, 2017–Apr. 7, 2017). Association for Computational Linguistics, pp. 739–745

4. Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol (July 2017b). "Neural Architecture for Temporal Relation Extraction: A Bi-LSTM Approach for Detecting Narrative Containers". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vancouver, Canada, July 30, 2017–Aug. 4, 2017). Association for Computational Linguistics, pp. 224–230

5. Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol (Aug. 2017a). "LIMSI-COT at SemEval-2017 Task 12: Neural Architecture for Temporal Information Extraction from Clinical Narratives". In: *Proceedings of the 11th International Workshop on Semantic Evaluation* (Vancouver, Canada, Aug. 3, 2017–Aug. 4, 2017). Association for Computational Linguistics, pp. 597–602

6. Julien Tourille, Matthieu Doutreligne, Olivier Ferret, Nicolas Paris, Aurélie Névéol, and Xavier Tannier (Oct. 2018). "Evaluation of a Sequence Tagging Tool for Biomedical Texts". In: *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis* (Brussels, Belgium, Oct. 31, 2018). Association for Computational Linguistics

Part I

Temporal Information Extraction from Clinical Narratives

Chapter 2

It's About Time: Temporal Information Extraction from Text

2.1 Introduction

This chapter presents the different aspects of time in linguistics and computational linguistics. First, we survey how the notion of time is addressed in the linguistic literature (Section 2.2). Then, we discuss how the computational linguistics community has modeled this notion: we describe the main models (Section 2.3) and resources (Section 2.4) that are available for devising temporal information extraction approaches (Section 2.5).

2.2 Time in Natural Language: A Linguistic Perspective

Time can be described according to three primitives: *events*, *times* and *temporal relations* (Derczynski 2017). In this section, we review each of them from a linguistic perspective.

We focus our attention on the English language. Our choice is guided by the fact that a large number of annotated resources are available in that language and that we mostly worked on texts written in English. However, we acknowledge the fact that some language particularities may not be reflected in the overview presented in this section. For instance, *tense* and *aspect* are different in French and English. These disparities are reflected in annotation efforts (Bittar et al. 2011).

2.2.1 Temporal Expressions

Temporal expressions or *time expressions* are used in natural language to give information about *when* something happened, *how long* something lasted or *how often* something occurred (Marşic 2011). They may denote *calendar dates*, *times of day*, *durations* or *sets of recurring times*.

The text extent which signals the temporal expression is called a *lexical trigger*. Marşic (2011) identifies six types of lexical triggers: nouns (e.g. *century* or *year*), proper names (e.g. *Christmas* or *April*), adjectives (e.g. *past* or *current*), adverbs (e.g. *currently* or *weekly*), time patterns (e.g. *9:00* or *'80s*) and numbers (e.g. *4th* as in *John arrived on the 4th*).

Temporal expressions convey information about *position in time*, *duration*, *frequency* or *temporal relationship* (Biber et al. 1999). We discuss the first three in this section. The last one, temporal relationship information, will be discussed in Section 2.2.3.

Position in Time. Temporal expressions can be used to locate an event in time. In this case, they answer the *when* question. Such expressions can take the form of noun phrases and usually includes a determiner such as *that* (Example 3a). Noun phrases can also be used within prepositional noun phrases (Example 3b). Position in time can also be expressed by adverbs (e.g. *now*, *then*, *today*, *ago* or *yesterday*) as shown in Example 3c, using time points (Example 3d), or vaguely specified with time periods or intervals (Example 3e).

(3) a. Mary met him **that afternoon** (Marşic 2011)
b. The wedding was **on Thursday** (Marşic 2011)
c. John went for a walk **yesterday** (Marşic 2011)
d. We found the letter **at twelve noon** (Allen 1983)
e. We found the letter **yesterday** (Allen 1983)

There is an agreement in the literature to make the distinction between *absolute* and *relative* position in time (Derczynski 2017; Ferro et al. 2005; Mani and Wilson 2000). This consideration led to the creation of three categories of temporal expressions referring to a position in time in the literature:

- **Absolute**. This category regroups time expressions that refer to an absolute point in time. Enough information is contained in the expression to pinpoint the exact temporal location.

- **Deictic**. Time expressions that are interpreted relatively to the utterance situation are categorized as deictic. These expressions usually specify a temporal distance and a direction from the utterance time (e.g. *tomorrow*, *yesterday*).
- **Anaphoric**. Time expressions that are interpreted relatively to another time which is not the utterance time are considered anaphoric. These expressions have usually three parts: a temporal distance, a temporal direction and an anchor for both distance and direction (e.g. *that evening*, *a few hours later*).

Temporal Duration. Temporal expressions denoting a temporal duration answer the *how long* question. They can take the form of noun phrases or prepositional noun phrases (Example 4a and 5). Some studies underline that bare noun phrases can be considered as prepositional noun phrases lacking the preposition *for* (Example 4b, Marşic (2011)). Other studies describe durative temporal expressions as intervals bound by a start and an end point where the distance between the two is known (Derczynski 2017). Also, they cannot be placed on a calendar. These expressions are sometimes close to deictic expressions as shown in Example 5.

(4) a. They lived **several years** in Italy. (Marşic 2011)
b. His mother-in-law stayed (for) **three weeks**. (Marşic 2011)

(5) a. **Duration**: The Texas Seven hid out there **for three weeks**. (Ahn et al. 2005)
b. **Deictic**: California may run out of cash **in three weeks**. (Ahn et al. 2005)

Frequency in Time. Temporal expressions denoting a frequency in time answer the *how often* question. These expressions usually start with *every/each* (Example 6a). Frequency in time can also be expressed as a bare plural (Example 6b) or can be introduced by the prepositions *on*, *at* or *in* (Example 6c). Finally, it can also be expressed by adjectives and adverbs derived from time units (e.g. *hourly*, *monthly*, *annually*) as shown in Example 6d. All examples are extracted from Marşic (2011).

(6) a. Mary writes an article or a review **every month**.
b. **Saturdays** John goes to the theatre.
c. They reviewed their stock portfolio **on the first day of each month**.
d. A **monthly** newsletter is emailed to all customers.

2.2.2 Events

Events in linguistics are studied under the notion of *eventualities* (Bach 1986) or *situations* (Comrie 1976). The literature makes the distinction between *non-stative* and *stative* eventualities (Dowty 1979; Pustejovsky 1995; Vendler 1967). Stative eventualities involve events where there is no change during the time span denoted by the event. Non-stative eventualities involve a change of state.

Marşic (2011) highlights that event identification in text is not a simple task. Events can be realized by verbs (Example 7a), nouns (Example 7b), adjectives (Example 7c), verb phrases

(Example 7d), clauses (Example 7e), sentences (Example 7f) or semantic entities (Example 7g). On some occasions, agreeing on the number of events is even difficult (Example 7f). All examples are taken from Marşic (2011).

(7) a. In fiscal 1989, Elco **earned** $7.8 million, or $1.65 a share.
b. Ms. Atimadi says the **war** has created a nation of widows.
c. They say IRA commanders are **responsible** for the recent bomb attacks.
d. Rally's Inc. said it **has adopted a shareholders rights plan**.
e. The Federal Bureau of Investigation says **it received more than eight thousand reports of hate group crimes last year**
f. **Telerate's two independent directors have rejected the offer**.
g. We know that 3,000 teens start smoking each day, **although it is a fact that 90% of them once thought that smoking was something that they'd never do**.

Most linguistic studies on events are centered around the *verb*. Authors try to classify events according to how the events they denote take place in time (Marşic 2011). The phenomenon is captured by the notion of *lexical aspect* which is described according to three main dimensions (Bethard 2007; Marşic 2011):

- **Dynamicity**: it allows to distinguish between *static* events (e.g. *know* or *love*), that do not involve change and *dynamic* events that do involve change (e.g. *run* or *deliver*).
- **Durativity**: it allows to distinguish between *instantaneous* (or *punctual*) events, that are conceptually one point in time (e.g. *find* or *blink*) and *durative* events that last for a period of time (e.g. *desire* or *push a cart*).
- **Telicity**: *telic* events have an end point or are directed toward a goal (Marşic 2011) as shown in Example 8a. *Atelic* events do not have a goal or endpoint (Example 8b). The telic/atelic distinction has been coined with other pairs of terms in the literature: *bounded* vs. *non-bounded*, *culminating* vs. *non-culminating*, *delimited* vs. *non-delimited* or *definite* vs. *indefinite change of state* (Marşic 2011).

(8) a. **Telic event**: John **fixed** the roof. (Marşic 2011)
b. **Atelic event**: He **drove** for hours.

This three-dimension lexical aspect breakdown has been used in a large number of papers studying eventualities from a linguistic perspective. Most of them are based on the work of Vendler (1967) who makes the distinction between four event categories:

- **States**, that do not describe any change in the world and last for a period of time (e.g. *desire* or *believe*). They are *static* and *durative*.
- **Activities**, that are *dynamic*, *durative* and atelic (e.g. *swim*, *push a cart* or *running*).
- **Accomplishments**, that are *dynamic*, *durative* and *telic* (e.g. *draw a circle* or *recover*).
- **Achievements**, that are *dynamic*, *punctual* and *telic* (e.g. *recognize* or *reach the top*).

The four-way classification of Vendler (1967) formed the basis for following work on lexical aspect. For instance, Bach (1986) further subdivides the *states* category by introducing the notions of *dynamic states*, that can be expressed by verbs occurring with progressive tenses (Example 9a), and *static states* that hold permanently (Example 9b). Non-states are divided into *processes* (corresponding to Vendler's activities) and *events* (Vendler's accomplishments and achievements). These events can be *protracted* (accomplishments) or *momentaneous* (achievements). He also subdivides the *momentaneous* category between *culminations* (Example 9c) and *happenings* (Example 9d) according to whether the event involves a transition. A overview of other linguistic theories of lexical aspect can be found in Marşic's thesis (Marşic 2011).

(9) a. Mary is **feeling** sick.
b. John **knows** the answer.
c. John's father **died** a few years ago.
d. Mary **noticed** John's mistake.

2.2.3 Temporal Relations

There are two main areas of research involving temporal relations. First, researchers tried to identify and characterize the linguistic realizations of temporal relations in natural language. Second, there is a line of work, mostly in the field of knowledge representation, which tries to model the temporal relations using algebra in order to be able to reason about time.

Linguistic Realization

There are several mechanisms used in natural language to denote temporal relations between events and/or time expressions. All examples in this section are extracted from Marşic (2011).

Tense. According to K. Brown (2005), *tense is a grammatical category that serves to locate situations in time*. It allows to locate events in time relatively to the time of utterance, also called *speech time*. There are three main tenses in English:

- **Present tense** can be subdivided into timeless present (Example 10a), habitual present (Example 10b), instantaneous present (Example 10c), historic present (Example 10d) and present referring to the future (Example 10e).

  (10) a. Water **consists** of hydrogen and oxygen.
  b. They **visit** their parents every week.
  c. I **advise** you to quit.
  d. Just as John arrived, Mary **leaves** the house.
  e. The airplane **departs** at 9pm tomorrow.

- **Past tense** expresses that the event took place in the past (before speech time). Past tense can be anaphoric, thus events are interpreted relatively to other times expressed in the previous or current utterances as shown in Example 11.

(11) John **went** to the park and **saw** many squirrels.

- **Future Tense** expresses that the event takes place after the speech time. In English, future tense can be expressed by the modal auxiliary construction with *will*, *be going to* followed by the infinitive, simple present or *be to/be about to* followed by the infinitive.

Grammatical Aspect. It allows for the distinction between *completed* (perfective) and *ongoing* (imperfective or progressive) events. It can be combined with tenses (e.g. Past Perfect or Past Progressive). There is an overlap in meaning between tense and aspect which leads to potential confusion. The most idiomatic example is the selection between the Simple Past and Present Perfective. The first one indicates that the period of time has ended while the second indicates that the period still continue at the speech time and may continue in the future as shown in Example 12.

(12) a. **Simple Past**: John lived in London for two years.
b. **Present + Perfective**: John has lived in London for two years.

Time Adverbials. These adverbials convey temporal relations between the time they denote and the verbal event they depends on. They are realized by time expressions as shown in Example 13. Marşic (2011) makes the distinction between **time position adverbials** (Example 13a), **durative adverbials** (noun phrases and prepositional phrases introduced by *for*, *during*, *within*, *over*, *throughout*, *while*, *whilst*, *as long as*, *so long as*, *until*, *till*, *up to*, *to*, *since* or *from*) and **frequency adverbials** (Example 13b)

(13) a. Mary left **at 10:30 am**.
b. Mary is careful **whenever she crosses a street**.
c. Mary moved to France **after** she **graduated**.

Other Means. There are other means of expressing temporal relations in natural language. They can be expressed at the **syntactic** level as in Example 14 where the event *result* is qualified by a temporal expression *the Sunday Election*.

(14) They do not know the **result** of the **Sunday Election**.

They can also be inferred using **world knowledge**. This is the case for subevents as shown in Example 15 where world knowledge is necessary to interpret that the event *painted* is included in the event *redecorated*.

(15) John **redecorated** his house. He first **painted** the walls.

Causality is one other mean to express temporal relationship. If one event causes another, thus it is before the other as shown in Example 16.

(16) John **fell**. Mary **pushed** him.

The **narrative sequence**, i.e. the order of natural language utterances, can be used to infer temporal relations. In Example 17, utterance order allows us to place the event *cooked* after the event *went*.

(17) Mary **went** home. She **cooked** dinner and ate it in front of the TV.

Event coreference can also be used to infer temporal relations. If one event is mentioned several times, then temporal relations involving one mention also apply for other mentions.

Finally, **inference** (e.g. *transitivity*) allows for inferring temporal relations. If event A occurred before event B and event B occurred before event C then event A occurred before event C.

Temporal Reasoning

Once we have identified the explicit and implicit relations in text, we must devise a system for reasoning about them (e.g. for being able to infer other relations). This problem has been addressed in the literature under the notion of temporal algebra. A temporal algebra allows to *deduce relationships between events based on their connections to other times and events using a set of rules* (Derczynski 2017).

Early work in the domain can be traced back to Reichenbach (1947) who modeled events as points in time. In his model, an event duration is a series of points between a start and an end point. He introduces a set of three relations between event start and end points: *before*, *simultaneous* and *after*. The model has some drawbacks. Although it works well for events like *love* or *know* where one can interpret that the event is true for every point in the set, it does not work for event like *draw a circle* where at every point before the completion of the drawing, a circle has not yet been drawn (Bethard 2007). The model presented by Reichenbach (1947) is thus not well suited for modeling temporal relations in natural language. Following work on temporal algebra can be divided into three main categories.

Temporal Interval Logic. The most influential work in this domain was conducted by Allen (1983) who devised a model where events and times are represented as continuous intervals with start and end points. It is a graphical model where event and time expressions are the nodes and temporal relations are the edges that connect the nodes. Allen (1983) identifies thirteen interval–interval relations (i.e. edges) and provides an algorithm for relation inference (e.g. using temporal transitivity).

Galton (1990) identifies some flaws in Allen's model. The major drawback is that the theory does not account for *continuous change*. The author takes the example of a ball thrown vertically in the air. There are two intervals, one for the ball going up, one for the ball going down. There is a moment in time where the ball is neither in these intervals. Allen's theory would need to represent this moment as an interval but this is contradictory with the laws of physics. Jixin and Knight (1994) tried to remedy to this problem by proposing to represent points as zero-length intervals.

Semi-Interval Reasoning. Derczynski (2017) underlines that temporal logic intervals are not perfect. For instance they become intractable with medium and large scenarios and, as we discussed above, there are some problems when intervals have a duration of zero. Also, some scenarios does not fit well in Allen's interval logic. Consider Example 18. Boundaries of the event *control* are difficult to conceptualize. The start boundary is probably the utterance

time (or publication time in this case) but it is difficult to place its end boundary as the event may be finished at the time of reading or may continue in the future.

(18) Today, rebels still **control** the airfield and surrounding area. (Derczynski 2017)

Semi-interval reasoning is devised to handle such cases by allowing to define only one temporal boundary of the events. One example of such model is the one of Freksa (1992) who proposes a model which allows to tackle uncertainty about events and allows to capture information from text that may not describe completely all intervals.

Point-based Reasoning. Point-based reasoning allows to model events and times as individuals points rather than intervals. It is possible to decompose intervals with begin and end points. This model type suffers from the fact that annotating text using points rather than intervals complicates the annotation task (Derczynski 2017). Furthermore, temporal relations in text are better represented by interval or semi-interval logics rather that points. However, one major benefit of point-based reasoning is that it is very fast to process (Verhagen 2004).

2.3 Time in Computational Linguistics

Analyzing time from a linguistic perspective is helpful to understand the mechanisms involved in the realization of time in natural language. In order to being able to model time from a computational linguistics perspective, we must settle on a linguistic model and derive an annotation schema that will be used to annotate time in text.

In this section, we present the different approaches that have been devised in the computational linguistics literature to model the three primitives of time: temporal expressions, events and temporal relations.

2.3.1 Temporal Expressions: From TIMEX to TIMEX3

There are four main types of temporal expressions in the literature (Strötgen and Gertz 2016): *dates*, *times*, *durations* and *sets*. These expressions may be *explicit*, *implicit*, *relative* or *underspecified* (Strötgen and Gertz 2015). We note that this categorization corresponds roughly to the linguistic description from Section 2.2 in which temporal expressions may denotes calendar dates, times of day, durations or sets of recurring times. Furthermore, the linguistic distinction between absolute and relative position in time is kept.

There were several attempts to model temporal expressions in the computational linguistic literature. In this section, we describe the most influential models: the TIMEX model series.

TIMEX. This is one of the earliest attempt to create an annotation scheme for temporal expressions. It was developed for the MUC-5 conference (Sundheim 1993). The conference included a shared task on NER that involved the extraction and categorization of dates and times. The task was proposed again in following MUC conferences until MUC-7 (Chinchor 1998).

The goal of the task was to identify and annotate time expressions that denote *calendar dates* or *times* with one TIMEX tag. The tag had one *type* attribute that could take the value

date or *time*. There was no task related to time expression normalization. TIMEX extraction and classification were part of a bigger task related to slot filling in which participants were asked to assign times to events (e.g. rocket launching dates).

TIDES TIMEX2. The second version of TIMEX was developed under the Defense Advanced Research Projects Agency (DARPA) research project Translingual Information Detection, Extraction and Summarization (TIDES) and the Automatic Content Extraction (ACE) Program. The development of this specification spanned over five years between 2000 and 2005. The last and final version of the annotation scheme is described in Ferro et al. (2005). The TIDES TIMEX2 annotation scheme, materialized by the tag TIMEX2, aimed at annotate a wider range of English temporal expressions. The TIMEX2 tag includes six attributes:

- **VAL**: it contains the ISO-8601 normalized value of the temporal expression when it represents a point or interval on a calendar or a clock.
- **MOD**: it is filled in when the time expression includes a modifier (e.g. *no more than* or *approximately*).
- **ANCHOR_VAL** and **ANCHOR_DIR**: these two attributes are used together to indicate the orientation and anchoring of time expressions.
- **SET**: it is used to mark time expressions that are representing sets of time. Its only possible value is YES. The absence of the attribute implies that the time expression is not representing a set.
- **COMMENT**: this attribute may be used by annotators to justify decisions for ambiguous time expressions or to signal doubts during the annotation process.

TIMEX3. The third and last version of the TIMEX annotation scheme has been developed within the TimeML (Pustejovsky et al. 2005) and ISO-TimeML (Pustejovsky et al. 2010) projects which aimed at creating a formal specification for representing events, temporal expressions and temporal relations in natural language.

The TIMEX3 annotation scheme is heavily based on its predecessors. The main difference is that the model does now includes two new types: duration and set. Strötgen and Gertz (2016) reports other differences with the TIMEX2 annotation scheme. Events are no longer be parts of temporal expressions. This is also the case for specific pre- and post-modifiers or time expressions. Also, the TIMEX3 tag cannot be nested. Its main attributes are:

- **TYPE**: it allows to specify the type of the annotated temporal expression. Possible values are *date*, *time*, *duration* and *set*.
- **VALUE**: as for the TIMEX2 tag, it contains the ISO-8601 normalized value of the temporal expression.
- **MOD**: this attributes allows to specify a modifier when necessary.
- **QUANT** and **FREQ**: these two attributes allow to specify the quantity and frequency of time expressions denoting sets.

- **BEGINPOINT** and **ENDPOINT**: they are used when the related expression is a duration. They allow to specify begin and end points of the duration.
- **TEMPORALFUNCTION**: it is an optional attribute that allows to model anaphoric temporal expressions. It is used in conjunction with either **ANCHORTIMEID** whose value points to the anchor identifier or **VALUEFROMFUNCTION** whose value points to another temporal function.

Clinical Domain. Although the TIMEX3 annotation scheme was originally developed for the annotation of temporal expressions in various textual genres and domains, its attributes and their possible values may not reflect the particularities of a given domain. This is the case for clinical narratives for which several adaptations have been implemented in following work. Most of the adaptations involved the simplification of the attribute structure. W. Sun et al. (2013) omit the temporal function attributes to simplify the annotation process. Galescu and Blaylock (2012) only annotate type and value attributes. Styler IV et al. (2014) adopt the strategy of Galescu and Blaylock (2012) and annotate only type and value attributes. However they expand the type value set by including a new temporal expression type *prespostexp*, used to account for the clinical concepts "preoperative", "postoperative" and "intraoperative". Examples of such expressions are given in Examples 19 and 20 . Although annotating these expressions as TIMEX3 may seem odd at first, Styler IV et al. (2014) argue that these expressions designate *specific temporal spans related to an implicit EVENT, and thus, are TIMEX3s*.

(19) The patient exhibits **postoperative** [$_{\text{EVENT}}$ changes].

a. postoperative CONTAINS changes

(20) Patient is in [$_{\text{EVENT}}$ recovery], no **post-operative** [$_{\text{EVENT}}$ nausea].

a. post-operative CONTAINS nausea

Tapi Nzali et al. (2015) devised an approach for automatic extraction of temporal expressions across domains in French narratives. Studying three corpora (news-based, historical and clinical), they show that time expression distribution is domain-specific. Among other observations, the authors highlight that the clinical corpus has a high proportion of *Set* compared to the two other domains.

2.3.2 Events: Task-Dependent Modeling

Event models vary according to the task they are devised for. Besides the *temporal information extraction* task, which is the one we are concerned with in this thesis, we identify three other tasks in the literature: *event extraction*, *slot filling* and *topic detection and tracking*.

Event Extraction

In event extraction, also called *event detection*, the goal is to extract and classify event mentions and their arguments from text and regroup mentions according to the event they are referring to. An event is defined as "a specific occurrence of something that happens, often a change of state, involving participants".

The task was formalized during the ACE 2005 shared task (Linguistic Data Consortium 2005) on event extraction. The ACE program aimed at providing new datasets and annotation schemas for the development of new information extraction methods. Objects of interest included *entities*, *times*, *values*, *relations* and *events*.

In the context of the shared task, an event mention is a span of text, also called an *extent*, which is usually a sentence. Each event mention has an anchor which is the most representative word of the event mention. Within the event mention, there could be zero or more arguments. In the ACE 2005 shared task, an event is an actual real-world event, i.e. the collection of event mentions referring to it. The ACE 2005 shared task limited the scope to specific event categories and argument positions.

System performance was evaluated with two methods. The first one was designed as an end-to-end evaluation that takes into account the NER phase (entities, time expressions, values and events). For the second one, participants were given the gold named entities and were asked to perform argument extraction.

This type of modeling inspired later work in both general (Mitamura et al. 2015) and biomedical domains. In the successive versions of the BioNLP shared task (Kim et al. 2009; Kim et al. 2011; Kim et al. 2016; Kim et al. 2013), participants were challenged on biomedical event extraction. The task involved extracting proteins, their types and their arguments. These tasks were part of bigger tasks such as protein coreference resolution, entity relation extraction or gene renaming tasks. Moreover, several tasks were concerned with the detection of event negation and speculation.

Slot Filling and Knowledge Base Population

In slot filling based shared tasks, the goal is not anymore the annotation of event mentions and anchors in text but rather to identify knowledge about entities and events. Participants are given a fixed inventory of relations and attributes and are asked to fill in these slots with values extracted from the text. Slots can be related to the *type*, the *agent*, the *time and place* or the *effect*. This task can be considered as something hybrid between relation extraction and question answering.

The first instance of this task can be traced back to the MUC conferences (Grishman and Sundheim 1996). Targeted topics changed regularly: naval sightings and engagements (MUC-1 and MUC-2), terrorist attacks (MUC-3 and MUC-4), joint ventures (MUC-5), succession events (MUC-6) or airplane crashes, and rocket/missile launches (MUC-7).

More recently, the NLP community started to focus on knowledge base population. The Text Analysis Conferences (TAC) introduce in 2014 a shared task (Mitamura et al. 2015) where the goal is to fill event slots with relations such as org:founded_by (who) or org:date_founded (when).

Topic Detection and Tracking

The Topic Detection and Tracking (TDT) task was a DARPA-sponsored shared task (Allan et al. 1998a; Allan et al. 1998b). The main objective of the task was to find and follow events in a stream of broadcast news stories. The first iterations of the task were concerned with *topics* but the following years, the task switched to *events*, i.e. things that happen at a point in time. There were several subtasks:

- **Segmentation**: participants were asked to segment a stream of text into stories.
- **Detection**: the objective is to identify the events discussed in the stories. Two tracks were organized: *retrospective* and *on-line* detection. In both cases, the task was modeled as a clustering problem were participants must regroup stories according to the event they are discussing.
- **New event detection**: the goal is to mark stories that are discussing an event for the first time.
- **Tracking**: participants are asked to associate stories with known events.

TDT is hard for both the organizers and the participants and was not further developed in later years until recent years. The recent development of micro-blogging platforms such as Twitter led a regain of interest for the task among the NLP community. Atefeh and Khreich (2015) presents a survey of models for event detection in Twitter. The authors identify tasks that are similar to those which were put into place for TDT in traditional media: detection of *unspecified* vs. *specified* (or *planned*) events and the *retrospective* vs. *on-line* detection.

Temporal Information Extraction

The last task involving event detection is temporal information extraction. This is the task we are concerned with in this thesis. In this task, the event is considered as an atomic object, generally represented by a word in a text. The goal is not to extract arguments as in previous models but rather to place an event in time relatively to other events and time expressions. The task is usually divided in three blocks:

- **Event extraction**: the objective is to extract event extents which are usually limited to one token.
- **Time expression extraction**: the objective is to extract time expressions usually modeled as TIMEX3.
- **Temporal relation extraction**: the objective is to extract temporal relations between events and/or temporal expressions but also between events and document creation times.

The most prominent shared tasks organized in the community around temporal relation extraction are the TempEval campaigns (UzZaman et al. 2013; Verhagen et al. 2007; Verhagen et al. 2010) which offer to work on news-based corpora (cf. Section 2.4.2). The annotation schema used to annotate these datasets is based on ISO-TimeML (Pustejovsky et al. (2010), cf. Section 2.4).

Temporal information extraction was originally addressed in the news domain. The specification TimeML (Pustejovsky et al. 2005) and the proof of concept corpus TimeBank (Pustejovsky et al. 2003) annotated with this schema paved the way for the development of new methods for temporal information extraction.

Clinical Domain. In the meantime, the clinical community started expressing a strong interest in temporal information extraction from clinical narratives (Bramsen et al. 2006a; Hripcsak et al. 2009). However, it came quickly to light that the notion of event is different between the general and the clinical domains.

First, medical staff is not interested in all events that appear in the text but only those which are medically relevant (Galescu and Blaylock 2012; Styler IV et al. 2014a).

Second, the forms of events, i.e. the linguistic realization of these medical events, differs from the ones found in the general domain. The ISO-TimeML specification mentions that events may be expressed by verbs, nouns and adjectives. The majority of nouns in this case are nominalizations of verbal events (e.g. *the detonation*). Similar cases can be found in the clinical domain (e.g. *intubation*) but most events in the clinical sense, are noun phrases that would not be annotated according to the ISO-TimeML specification. Galescu and Blaylock (2012) identify this phenomenon while annotating temporal relations between medical *problems*, *tests* and *treatments*. In a majority of cases, these entities are not eventive in the sense of ISO-TimeML but are considered as representative of the event they are the most closely related. Styler IV et al. (2014) explains that this interpretation of disorder, medication and procedure as rightful events comes from the fact that the medical staff does not discuss them without an associated (implicit) event. In the light of this observation, Styler IV et al. (2014) infer that all entities falling into the following Unified Medical Language System (UMLS®) categories can be considered as events: *Disorder*, *Chemical/Drug*, *Procedure* and *Sign/Symptom*.

2.3.3 Temporal Relations

The last piece of information to be modeled in order to have a fully functional annotation scheme is the temporal relation set. One of the first attempt to model temporal relations was done during the MUC-7 shared task (Chinchor 1998), where participants were asked to provide a launch date for rocket launching events. Later, Katz and Arosio (2001) annotated a set of fifty sentences with intra-sentence temporal relations. The corpus was also annotated with syntactic information. Relations were drawn between verbs and utterance times as well as between pairs of verbs. There were only two relations in the set: *precedence* and *inclusion*.

Setzer and Gaizauskas (2002) annotated a trial corpus of six newswire articles using the Sheffield Temporal Annotation Guidelines (STAG) (Setzer 2001). They devised an iterative annotation process where the first step consists in annotating events, time expressions, signals and temporal relations that are explicit in the text or syntactically implicit. Then, they derive all inferrable annotations based on this first phase. They repeat the two steps until every pair is annotated. One major drawback emerging from that annotation effort is the low inter-annotator agreement. The authors identify several causes: the underspecified annotation guidelines, the lack of annotator training and the task difficulty.

Pustejovsky et al. (2005) and Pustejovsky et al. (2010) presented the ISO-TimeML Specification which has become the basis of most following temporal annotation efforts. The annotation scheme allows to annotate events, time expressions, signals, and temporal relations between these entities. The temporal relation set is derived from Allen's interval relations: *before*, *after*, *includes*, *is included*, *during*, *during inv*, *simultaneous*, *iafter*, *ibefore*, *identity*, *begins*, *ends*, *begun by* and *ended by*. They annotated the TimeBank corpus (Pustejovsky et al.

2003) as proof of concept, which was used in a simplified version in the TempEval shared tasks (UzZaman et al. 2013; Verhagen et al. 2007; Verhagen et al. 2010)

Clinical Domain. In the clinical domain, several temporal annotation guidelines have been devised upon the work of Pustejovsky et al. (2010) and led to the creation of related corpora. Among these efforts, the THYME-TimeML specification (Styler IV et al. 2014a) proposes a full temporal annotation scheme adapted to the clinical domain. The authors annotated a corpus of 1,200 documents that have been used in the Clinical TempEval challenges (Bethard et al. 2015; Bethard et al. 2016; Bethard et al. 2017).

One major modification of TimeML brought by THYME-TimeML lies in the temporal relation set. Instead of considering classic Allen's relations, the authors embrace the narrative container concept (Pustejovsky and Stubbs 2011). According to Styler IV et al. (2014), a narrative container is a temporal bucket in which several events may fall. These containers may be realized by temporal expressions, durative events or abstract medical concepts. Another way to visualize the narrative container concept is to think of it as the imbrication of temporal intervals. Example 21, taken from Styler IV et al. (2014), illustrates the concept. In this example, the event *colonoscopy* happened on *January 7, 2010*. Thus, the temporal expression associated with the calendar date is marked as containing the event *colonoscopy*. All other events mentioned in the example are included in the event *colonoscopy*.

(21) [EVENT **Colonoscopy**] ([TIMEX3 **January 7, 2010**]): Fair/adequate [EVENT **prep.**], Limited [EVENT **Colonoscopy**] to the distal sigmoid due to an obstructive [EVENT **lesion**]. Diminutive [EVENT **polyps**] of the rectosigmoid.

a. [TIMEX3 **January 7, 2010**] CONTAINS [EVENT **Colonoscopy**]
b. [EVENT **Colonoscopy**] CONTAINS [EVENT **prep.**]
c. [EVENT **Colonoscopy**] CONTAINS limited [EVENT **colonoscopy**]
d. [EVENT **Colonoscopy**] CONTAINS [EVENT **lesion**]
e. [EVENT **Colonoscopy**] CONTAINS [EVENT **polyps**]
f. [EVENT **Colonoscopy**] CONTAINS [EVENT **removed**]

Styler IV et al. (2014) explain that using the narrative container concept instead of classical relations allows for better annotation quality by improving the inter-annotator agreement while limiting under- and over- annotation effects that are common in temporal annotation efforts. The authors add that this approach mimics closely *the general structure of story-telling* in both the general and clinical domains. Medical staff tend to cluster discussions of medical events around a given date or other event. Placing events in narrative containers and linking these containers with a few relations allows to draw quickly a useful understanding of the overall clinical timeline. Events within clusters are not ordered but the ordering can be found a posteriori with domain knowledge. The authors argue that this annotation strategy offers a good balance between temporal informativeness and annotation quality.

2.4 Resources for Temporal Information Extraction

In this section, we present resources that have been developed in computational linguistics for temporal information extraction. First, we review the two most influential annotation

schemes used in the literature to build temporally annotated corpora. Second, we describe the main annotated corpora that exist in both the general and the clinical domains.

2.4.1 Full Annotation Schemes

There are two main annotation schemes that have been developed in computational linguistics. ISO-TimeML (Pustejovsky et al. 2010) is a general markup language for annotating temporal expressions, events and temporal relations in text. THYME-TimeML (Styler IV et al. 2014a) is an adaptation of ISO-TimeML to the clinical domain.

ISO-TimeML

ISO-TimeML (Pustejovsky et al. 2010) is an ISO standard for temporal information markup in text. It is a standardization of TimeML (Pustejovsky et al. 2005), a general-purpose markup language for time. This specification allows to annotate events, time expressions, relations between events and/or time expressions and the relative ordering of events.

Temporal expressions are captured by TIMEX3 tags (described in Section 2.3.1). Minor differences with the original TIMEX3 annotation scheme includes the limitation of possible modifier values to a restricted set and the addition of an attribute which allows to specify the function of the temporal expression within the document (*creation-time*, *modification-time*, *publication-time*, *release-time*, *reception-time*, *expiration-time* or *none*).

Events are modeled with the combination of EVENT and MAKEINSTANCE tags. The authors distinguish between event triggers and event instances. The MAKEINSTANCE tag holds most of the attributes and creates the actual realization of the event. Such scheme allows to model complicated examples such as the one presented in Example 22 where two event instances are needed.

(22) John taught on Monday and Tuesday.
<EVENT eid="e1">taught</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1" tense="PAST" aspect="NONE" pos="VERB"/>
<MAKEINSTANCE eiid="ei2" eventID="e1" tense="PAST" aspect="NONE" pos="VERB"/>

According to the annotation guidelines (Saurí et al. 2006), the authors consider events as *a cover term for situations that happen or occur*. Only one word should be annotated, usually the head of the minimal chunk expressing the event (e.g. a verb, a noun or an adjective). Event may be realized by verbs (Examples 23a and 23b), nominalizations (Example 23c), adjectives (Example 23d), predicatives clauses (Example 23e) or prepositional phrases (Example 23f). All examples are taken from Saurí et al. (2006).

(23) a. A fresh flow of lava, gas and debris **erupted** there Saturday.
b. Prime Minister Benjamin Netanyahu called the prime minister of the Netherlands **to thank** him for thousands of gas masks his country has already contributed.
c. Israel will ask the United States to delay a military **strike** against Iraq until the Jewish state is fully prepared for a possible Iraqi **attack**.
d. A Philippine volcano, **dormant** for six centuries, began exploding with searing gases, thick ash and deadly debris.

e. "There is no reason why we would not **be prepared**", Mordechai told the Yediot Ahronot daily.

f. All 75 people **on board** the Aeroflot Airbus died.

The main event instance attributes are:

- **CLASS**: it allows to distinguish between classes of events. An event can be of the type *occurrence*, *perception*, *reporting*, *aspectual*, *state*, *i-state* or *i-action*.
- **TENSE** and **ASPECT** allows to capture common distinctions between categories of verbal phrases. The TENSE attribute can take the following values: *past*, *present*, *future*, *none*, *infinitive*, *prespart* or *pastpart*. The ASPECT attribute can take the following values: *progressive*, *perfective*, *imperfective*, *prefective-progressive*, *imperfective-progressive*, *none*. Non finite verbs are marked with either *infinitive*, *prespart* or *pastpart* values.
- **POS**: this attribute allows to specify the syntactic category of the text extent (usually one token) marked as an event. It can take the value *adjective*, *noun*, *verb*, *preposition* or *other*.
- **POLARITY**: this attribute allows to specify if the event is negated or not. It can take the values *neg* or *pos*.
- **MODALITY**: the attribute allows to mark event modality. Examples of possible values are *must* or *should*.
- **CARDINALITY**: this attribute is used if there is a modifier denoting a cardinality. For instance, Example 22 can also be modeled with a unique MAKEINSTANCE element whose CARDINALITY attribute will be 2.

Temporal relations can happen between two events, two times or between an event and a time. They are modeled as TLINKs with several attributes:

- **RELTYPE**: it allows to specify the temporal relation holding between the entities. Possible values are *before* or *after* (Example 24a), *includes* or *is-included* (Example 24c), *during* or *during-inv* (Example 24d), *simultaneous*, *iafter* or *ibefore* (Example 24b), *identify*, *begins* or *begun-by* (Example 24e), *ends* or *ended-by* (Example 24f). All examples are taken from Saurí et al. (2006).

(24) a. The police looked into the **slayings** of 14 women. In six of the cases suspects have already been **arrested**.

b. All passengers **died** when the plane **crashed** into the mountain.

c. John **arrived** in Boston **last Thursday**.

d. James was **CTO** for **two years**.

e. John was **in the gym** between **6:00 p.m.** and 7:00 p.m.

f. John was **in the gym** between 6:00 p.m. and **7:00 p.m.**

- **EVENTINSTANCEID** or **TIMEID**: identification label of the source entity (event or time expression) involved in the relation.

- **RELATEDTOEVENTINSTANCE** or **RELATEDTOTIME**: identification label of the target entity (event or time expression) involved in the relation.

Besides TLINKs which are used to mark temporal relation between entities, there are two other link types that are used to annotate modality and aspect:

- **SLINK** tags are used to mark subordination links between two events. This relation can be *modal*, *factive*, *counter-factive*, evidential, *negative-evidential* or *conditional*.
- **ALINK** tags are used to mark the relation between an aspectual event and its argument event. Possible values are *initiation*, *culmination*, *termination*, *continuation* or *reinitiation*.

Finally, **SIGNAL** tags are used to mark text extents that make explicit the relation holding between two entities. Signals can take the form of temporal prepositions (e.g. *on*, *in* or *at*), temporal conjunctions (e.g. *before* or *after*), prepositions signaling modality (e.g. *to*) or special characters (e.g. the character / which denotes a range as in 1998/1999). Signal may be referenced in relations expressed with TLINKs, SLINKs or ALINKs.

THYME-TimeML

THYME-TimeML (Styler IV et al. 2014a) is a temporal annotation scheme developed specifically for the clinical domain. As its name suggests, it is strongly inspired by the ISO-TimeML specification described in the previous section. Three types of elements are annotated: clinical events, temporal expressions and temporal relations between these two entity types as well as between events and document creation times. All examples in this section are taken from Styler IV et al. (2014).

As previously mentioned in Section 2.3.2, Styler IV et al. (2014) extend the definition of an event proposed in Pustejovsky et al. (2010) by including all pertinent events for the clinical timeline. In this context, events can be diagnostics, diseases, procedures and tumors. Also, AJCC tumor type codes (American Join Committee on Cancer Staging Codes) are annotated. These codes are events in the sense that they are assigned to a patient at a given point in his medical history. According to ISO-TimeML, they would not be annotated. However, they represent valuable information for medical staff. In this context, verbs of discussions such as *talk*, *agree* and *repeat* are also annotated due to the fact that they are important for legal reasons. Moreover, the annotation specification formalizes the notion of *entities as events* that has been devised in previous research efforts (e.g. in A. Roberts et al. (2009)). Thus treatments (Example 25a) and diseases (Example 25b) can be considered as events.

(25) a. [$_{\text{EVENT}}$ **Levaquin**] 750 mg p.o. q. day (will restart today).
b. The [$_{\text{EVENT}}$ **CT**] showed a small rectal [$_{\text{EVENT}}$ **abcess**].

Styler IV et al. (2014) made several modifications to the ISO-TimeML specification to simplify the annotation process and better fit the clinical domain. To improve the inter-annotator agreement, event modality is not modeled with SLINK anymore but with three event attributes.

The **contextual modality** attribute can take four values. An event is considered as *actual* if it did happen in the real world (Example 26a). Conditional or hypothetical events will be

marked as *hypothetical* (Example 26b). Some events, such as diagnostics, can be expressed with caution by the medical staff in order to avoid legal repercussions. These events will be marked as *hedged* (Example 26c). Finally, events discussing a treatment or a disease from a generic point of view will be marked as *generic* (Example 26d).

(26) a. His anterior chest rash has not [$_{\text{ACTUAL}}$ **reoccurred**].
b. We suspect either [$_{\text{HYP.}}$ **achalasia**] or [$_{\text{HYP.}}$ **pseudoachalasia**] here.
c. The patient may have undergone a mild [$_{\text{HEDGED}}$ **stroke**]
d. I explained that BRAF [$_{\text{GEN.}}$ **mutations**] have no predictive value with regard to cetuximab [$_{\text{GEN.}}$ **sensitivity**].

The **contextual aspect** attribute allows to distinguish between intermittent events (e.g. *vomiting*), constant events (e.g. *fewer*) and new events (e.g. discovering a new tumor). The attributes will respectively take the values *intermittent* (Example 27a), *n/a* and *novel* (Example 27b).

(27) a. He reports occasional bright red [$_{\text{INT.}}$ **bleeding**] from the rectum.
b. The newest [$_{\text{NOVEL}}$ **MRI**] revealed a previously undiscovered mass.

Finally, the **permanence** attribute allows to express the difference between the medical concepts *acute* and *chronic*. A disease will be considered chronic if it is uncurable. In other cases, it will be marked as finite. Although this feature was originally bound to be annotated, the authors dropped it after a few annotation iterations due to the fact that the task demands a lot of medical knowledge.

Besides event modality, there are other event properties marked in the corpus. The **type** attribute allow to distinguish between three event types. *Aspectual* events encode aspectual information about other events (Example 28a). *Evidential* events allow to link an information source (e.g. radiography) to an observation/diagnostic based on this source (e.g. broken rib) as in Example 28b. Other events will be marked as *n/a*.

(28) a. The rash has not [$_{\text{ASPECTUAL}}$ **reappeared**] and we will monitor closely.
b. Her CT-scan [$_{\text{EVIDENTIAL}}$ **showed**] a small mass in the right colon.

The **polarity** attribute encodes the fact that an event took place in the real world (Example 29a) or did not (Example 29b).

(29) a. The patient has [$_{\text{POSITIVE}}$ **hepatosplenomegaly**].
b. No evidence for new suprasellar [$_{\text{NEGATIVE}}$ **mass**].

The **degree** attribute allows to nuance polarity of an event. Thus, an event can be a little positive (Example 30a) or almost positive (Example 30b). Similarly, an event may be considered as little negative or almost negative.

(30) a. There is a small amount of bright T1 [$_{\text{LITTLE}}$ **signal**].
b. Abdominal tenderness has nearly [$_{\text{MOST}}$ **disappeared**].

Temporal expressions are modeled using TIMEX3 tags. The annotation scheme uses all the types defined in ISO-TimeML and adds a specific *prepostexp* category for temporal expressions such as *preoperative*, *intraoperative* and *postoperative*.

TIMEX3 tags have only one attribute *class* which can take several values. The attribute will take the value *date* (Example 31a) if the temporal expression denotes a calendar date. It can also take the value *time* if the expression denotes a specific time in the day (Example 31b). The expression will be marked as a *duration* if it denotes a duration (Example 31c). The expression can also be marked as a *quantifier* (Example 31d) or a *set* (Example 31f) whenever it combines both a duration and a quantity. Finally, the attribute will take the value *prepostexp* if the temporal expression denotes the temporal phenomenon presented above (Example 31e).

(31) a. I would probably restart her furosemide [DATE **tomorrow morning**].
b. Following the patient's latest seizure, [TIME **20 minutes ago**], we are re-evaluating her medications.
c. In [DURATION **the last week**], his pain has significantly worsened.
d. The patient vomited [QUANTIFIER **twice**] before the surgery.
e. The patient exhibits [PREPOSTEXP **post-exposure**] changes.
f. Mirtazapine REMERON 7.5-mg tablet 1 tablet by mouth [SET **every bedtime**].

As we mentioned in Section 2.3.3, Styler IV et al. (2014) embrace the narrative container concept to annotate temporal relations between events and/or temporal expressions. However, other types of relations are annotated to model temporality within containers. An entity can be placed *before* an other entity (Example 32a). Both entities can *begin* at the same time (Example 32b) or *ends* at the same time (Example 32c). Finally, two entities can temporally *overlap* with each other (Example 32d). Only linguistically apparent relations are annotated.

(32) a. She **vomited** shortly before **surgery**.
b. She has had abdominal **cramping** since **January**.
c. She has had no **bleeding** since her **stitches** were **removed**.
d. The patient's first **MI** occurred while she was undergoing **chemotherapy**.

The second type of temporal relation is the relation existing between events and document creation times. Hence, an event can happen *before* (Example 33a) or *after* (Example 33b) the document creation time. It may have started before the document creation time and still be on-going at that time (*before-overlap*, Example 33c). Finally, it can be considered as true at the document creation time (*overlap*, Example 33d). Document creation time is the same time at which the clinical examiner has seen the patient even if the document has been written after the meeting.

(33) a. This is unchanged and may be related to treatment [BEFORE **changes**].
b. The patient will [AFTER **return**] tomorrow for [AFTER **labs**] and [AFTER **exam**].
c. She has not [BEF./OVER. **seen**] a cardiologist.
d. The patient [OVERLAP **continues**] to [OVERLAP **do**] well as an outpatient.

2.4.2 Corpora and Associated Shared Tasks

Several corpora for temporal information extraction have been devised in both general and clinical domains. In this section, we describe the resources based on ISO-TimeML and THYME-TimeML specifications, which were discussed in the previous section. Table 2.1 compares these resources along four axes: languages, document types, annotation schemes and sizes.

The TimeBank and the AQUAINT TimeML Corpora

The TimeBank corpus (Pustejovsky et al. 2003) is a proof-of-concept resource for the ISO-TimeML specification. It contains 183 documents ($\approx$ 61,000 tokens) annotated according to the specification. It includes biographies, description of events and broadcast news and newswire.

The AQUAINT TimeML Corpus, also known as the *Opinion Corpus*, is a small resource containing 73 news reports ($\approx$ 38,000 tokens) annotated according to the ISO-TimeML specification.

The work of Pustejovsky et al. (2003) inspired other researchers in creating temporally annotated corpora in other languages based on the same specifications. Among them, Bittar et al. (2011) led an annotation effort on a comparable French corpus. Asahara et al. (2014) and Caselli et al. (2011) did the same with Japanese and Italian texts. Other efforts include the translation of the TimeBank corpus in Portuguese (Costa and Branco 2012) and Romanian (Forascu and Tufiș 2012).

The TempEval Corpora

The TempEval corpora are a series of resources used in the TempEval shared tasks (UzZaman et al. 2013; Verhagen et al. 2007; Verhagen et al. 2010). The corpora were annotated according to a simplified version of ISO-TimeML. A version of TimeBank following this new annotation scheme was provided to the participants as training corpus. The temporal relation set comprises six relations: *before*, *after*, *overlap*, *before-or-overlap*, *overlap-or-after* and *vague*.

The first edition of the evaluation campaign (Verhagen et al. 2007) focused on temporal relation extraction between EVENTs and TIMEX3s within the same sentence, between EVENTs and document creation times and between two main EVENTS of two consecutive sentences.

The second and third editions of TempEval (UzZaman et al. 2013; Verhagen et al. 2010) offered a task related to EVENT and TIMEX3 extraction and further subdivided the temporal relation extraction task. Participants were asked to extract temporal relations between EVENTs and TIMEX3s which are syntactically dominated by the EVENTs, between EVENTs and document creation times, between two main EVENTS of two consecutive sentences and between EVENTs which are in a syntactic dependency relation.

The Clinical E-Science Framework (CLEF) corpus

The CLEF corpus (A. Roberts et al. 2008) is a set of documents from the Royal Marsden Hospital. It contains clinical narratives, histopathology reports and imaging reports. The

Corpora	Languages	Document Type	Annotation Scheme	Size (Approx number of documents or tokens)
TimeBank	EN	news, biographies, event descriptions	ISO-TimeML	183 doc.
FR-TimeBank	FR	news	Adapted ISO-TimeML	109 doc.
TimeBankPT	PT	news, biographies, event descriptions	ISO-TimeML	183 doc. (Translation of TimeBank)
BCCWJ-TimeBank	JP	news, blogs, books, magazines	Adapted ISO-TimeML	1,000,000 tok.
Romanian TimeBank	RO	news, biographies, event descriptions	ISO-TimeML	183 doc. (Translation of TimeBank)
Ita-TimeBank	IT	news	Adapted ISO-TimeML	150,000 tok.
AQUAINT	EN	news	ISO-TimeML	73 doc.
TempEval-1	EN	news	Modified ISO-TimeML	183 doc. (same as TimeBank)
TempEval-2	CN, EN, IT, FR, KO, SP	news	Modified ISO-TimeML	23,000 tok. (CN); 63,000 tok. (EN); 27,000 tok. (IT); 19,000 tok. (FR); 14,000 tok. (KO); 68,000 tok. (SP)
TempEval-3	EN, SP	news	Modified ISO-TimeML	768,075 tok. (EN); 67,819 tok. (SP)
CLEF	EN	clinical narratives, histopathology reports, imaging reports	Custom	50 doc.
Galescu & Blaylock	EN	discharge summaries	modified ISO-TimeML	7,701 tok.
i2b2	EN	discharge summaries	THYME-TimeML beta version	310 doc.
THYME	EN	clinical reports pathology reports	THYME-TimeML	1,200 doc.
MERLoT	FR	discharge summaries, physician letters, procedure reports, prescriptions	modified ISO-TimeML	500 doc.

Table 2.1: Publicly available corpora annotated with temporal information based on either the ISO-TimeML or THYME-TimeML specifications. We compare these resources along four axes: languages, document types, annotation schemes and sizes

corpus contains documents from two patient records. The document selection process aimed at selecting documents that are representative of the various document types and lengths. Also, each of the two patient records annotated with temporal information contains nine narratives, one radiology report, seven histopathology reports and the associated structured data.

These documents have been annotated with entities but also with semantic and temporal relations. The latter are modeled as CTlinks (CLEF Temporal Links) between TLCs (Temporally Located CLEF entities). The annotation scope is limited to relations that occur between an event and a time expression. Considered events include investigations, interventions and conditions. Temporal expressions are modeled as TIMEX3 elements. There are several temporal links: *after*, *ended by*, *begun by*, *overlap*, *before*, *none*, *is included*, *unknown* and *includes*. The other relation type annotated is the one existing between TLCs and document creation times. The relation type set is the same as the one used for the other temporal relation.

The Corpus of Galescu and Blaylock (2012)

Galescu and Blaylock (2012) annotated a subset of the corpus used during the 2010 edition of i2b2/VA challenge on relations (Uzuner et al. 2011). The authors chose only the documents that were not temporally altered by the deidentification process. They extracted the sections related to *History of Present Illness* from 97 discharge summaries from Partners Healthcare, resulting in a corpus comprising 44 sections (410 sentences and 7,701 tokens).

The authors annotated clinical events, time expressions and temporal relations between these entities. The scope of the annotation effort is limited to intra-sentence relations. Considered events include *problems*, *tests* and *treatments*.

Temporal expressions are annotated as TIMEX3. Similarly to other annotation efforts, the authors annotate only the *type* and *value* attributes. The type attribute can take the following values: *date*, *time*, *duration* or *set*. One TIMEX3 is added to each document to represent the document creation time.

Temporal relations are annotated as TLINKs. They can occur between events, time expressions and between events and time expressions. Relations are the same as the ones used during the TempEval shared tasks: *before*, *after*, *overlap*, *before-overlap*, *overlap-or-after* or *vague*.

The Informatics for Integrating Biology & the Bedside (i2b2) Corpus

The i2b2 corpus (W. Sun et al. 2013a) is a temporally annotated corpus of 310 discharge summaries ($\approx$ 178,000 tokens) from Partners Healthcare and the Beth Israel Deaconess Medical Center. The resource was used during the 2012 edition of the i2b2/VA shared task on temporal information extraction. The annotation guidelines used during its creation are based on ISO-TimeML and on an earlier version of THYME-TimeML.

Clinically relevant events are annotated with EVENT tags. These events include clinical concepts (*problems*, *tests* and *treatments*), clinical departments (e.g. *surgery* or *main floor*), evidentials (i.e. events that indicate a source of information such as the word *complained* in the patient *complained* about), occurrences (i.e. events that happen to the patient, such as *admission*, *transfer* or *follow-up*). Events have several attributes: type (problems, test or

treatment), polarity (positive or negated) and modality (indicates whether an event *happen*, *is proposed*, *is mentioned as conditional* or as *possible*.)

Time expression annotation follows the TIMEX3 annotation scheme. The authors added a section time to track the section creation date. For instance, the section time for the *clinical history section* is the date of admission whereas the section time for the *hospital course section* is the discharge date. TIMEX3s have three attributes: type (*date*, *time*, *duration* or *set*), value (ISO-8601) and modifier (*exact* or *approximate*)

Temporal relations are annotated with TLINKs. They occur between two TIMEX3s, two EVENTs or between a TIMEX3 and a EVENT. The initial set of relation types included *before*, *after*, *simultaneous*, *overlap*, *begun-by*, *ended-by*, *during* and *before-overlap*. However, the authors noticed that the inter-annotator agreement was low for several relation types. They decided to merge *before*, *ended-by* and *before-overlap* into the single class *before*. They performed the same for *begun-by* and *after* which were merged into the single class *after*. Finally *simultaneous*, *overlap* and *during* were merged as *overlap*.

The 2012 i2b2 shared task offered three tracks: EVENT and TIMEX3 extraction, TLINK extraction (gold entities are provided) and end-to-end extraction where participants must accomplish the two first tasks.

The Temporal Histories of Your Medical Event (THYME) Corpus

The THYME corpus (Styler IV et al. 2014a) is set of clinical documents annotated as proof-of-concept for the THYME-TimeML specification. It contains two sets of 600 documents from brain cancer and colon cancer patients at the Mayo clinic. These notes have been deidentified. Each set is divided in three parts: train (50%), dev (25%) and test (25%).

The corpus was used during the Clinical TempEval shared tasks (Bethard et al. 2015; Bethard et al. 2016; Bethard et al. 2017). In the two first editions, only the colon cancer part was used. In the last edition of the shared task, participants were asked to perform domain adaptation by training on colon cancer related documents and testing on brain cancer related documents. At this occasion, two tracks were proposed, one without any document from the target domain (unsupervised domain adaptation) and the other with 30 documents from the target domain (supervised domain adaptation).

Several subtasks were presented to the participants: time expression extraction, event extraction and temporal relation identification (between events and/or time expressions but also between events and document creation times). Also, two tracks were offered for temporal relation extraction, one where gold time expressions and events were given to the participants and one where participants were asked to perform all three tasks (Bethard et al. 2015; Bethard et al. 2016).

The Medical Entity and Relation LIMSI annOtated Text (MERLoT) Corpus

The MERLoT corpus (Campillos et al. 2018) is a corpus of 500 clinical notes written in French from a hepato-gastro-enterology and nutrition ward. The documents have been de-identified and pseudonymized using the Medical Information Anonymization (MEDINA) tool (Grouin and Névéol 2014). Documents are segmented into zones to distinguish between headers and footers from the core medical content. There are several types of documents: discharge summaries, physician letters, medical procedure reports and prescriptions. The corpus is

extracted from a larger collection of documents. Sampling has been done by preserving document type proportion.

The corpus has several layers of annotations. The annotation scheme contains 12 entities: *anatomy*, *biological process or function*, *chemical and drugs*, *concept or idea*, *devices*, *disorder*, *genes or proteins*, *hospital*, *living beings*, *medical procedure*, *person*, *sign or symptom*, *temporal expression*. A subset of these entities are considered as events: *biological process or function*, *chemical or drug*, *concept or idea*, *disorder*, *medical procedure* and *sign or symptom*. Events are characterized by several attributes that take the form of entities linked with a relation to the events. For instance, aspectual text elements (e.g. *started on* or *interrupted*) may be linked to events via relations (e.g. *start* or *stop*). We refer the reader to the original paper for an overview of all entities and relations involving events (Campillos et al. 2018).

The temporal annotation scheme follows mostly the ISO-TimeML specification. Time expressions are encoded as TIMEX3 entities. However, the SIGNAL tag described in Pustejovsky et al. (2010) is not used in this annotation effort. Instead, the text extent that is considered as SIGNAL in ISO-TimeML is embedded in the time expression. Time expressions have one attribute type which can take the following values: *date*, *time*, *duration* or *frequency*

Two types of temporal relations are annotated within the corpus. The first one concerns relations between events and/or temporal expressions. The relation set comprises six relation types: *before*, *begins on*, *during*, *ends on*, *overlap* and *simultaneous*. The second one concerns relations between events and document creation times. It is realized as an event attribute which can take the following values: *before*, *before-overlap*, *overlap*, *after*.

Other Clinical Corpora

Several other clinical corpora can be found in the literature. Harkema et al. (2005) devised an annotation scheme based on ISO-TimeML and annotated events related to X-rays and CT-scans in 446 clinical documents.

Mowery et al. (2008) annotated 24 clinical reports with temporal information. They do not follow the ISO-TimeML specification and annotate directly events with temporal information.

Savova et al. (2009) annotated a corpus of 5,000K tokens following an annotation scheme based on ISO-TimeML. All clinical and non-clinical events are annotated. The authors annotate relations between events and document creation times and event-event relations. They also annotate event modality and aspect.

2.5 Approaches for Temporal Information Extraction

Several approaches have been devised in the literature for temporal information extraction ranging from rule-based methods to fully supervised machine learning approaches. In this section, we review these approaches in both the general and clinical domains. We start with time expression extraction. Then, we discuss event extraction. Finally we review methods for temporal relation extraction.

2.5.1 Temporal Expression Extraction

Rule-based Approaches. Rule-based approaches have been very efficient for temporal expression extraction due to the fact that the diversity of realization in text is rather small. One of the most popular rule-based system is Heideltime (Strötgen and Gertz 2013), a multilingual, cross-domain, temporal tagger. It extracts time expressions as TIMEX3 entities and provides their normalization to ISO-8601 format when applicable. Heideltime targets four document types: narrative-style, news-style, English colloquial and scientific writing. The tool is build as a Unstructured Information Management Architecture (UIMA) module and can therefore be used in NLP pipelines. It may also be improved by the community which can provide new resources for a language or a domain (e.g. Moriceau and Tannier (2014) provide a French model). The system obtained competitive results in several shared tasks such as the SemEval 2013 subtask on temporal expression extraction with a f-score of 90.30%.

Other rule-based systems include SUTime (A. X. Chang and Manning 2012), a temporal tagger similar to Heideltime. The system is part of the CoreNLP text processing pipeline (Manning et al. 2014). TempEx (Mani and Wilson 2000) was one of the first time expression tagger. It extracts and normalizes time expressions according to the TIMEX2 specification. It served as the base for GUTime, which is an extension of TempEx. It extracts and normalizes time expressions based on the TIMEX3 specification. Finally, Chronus (Negri and Marseglia 2005) was developed for the ACE TERN 2004 competition on time expression extraction and normalization.

Hybrid Approaches. Several other models implement hybrid approaches. The ATT system (Jung and Stent 2013) used hand-crafted features for the TempEval time expression extraction task. The authors modeled the problem as a sequence labeling task and trained a Conditional Random Field (CRF) model to label each token in the corpus. They obtained a f1-score of 85.60%. Bethard (2013) used a similar approach on the same shared task. He exploited a simple set of features and trained a Support Vector Machine (SVM) instead of a CRF.

Clinical Domain. Almost all approaches devised for temporal expression extraction in the clinical domain are hybrid. K. Roberts et al. (2013) used HeidelTime and Llorens et al. (2012) system outputs do devise features that are used in a CRF and a SVM for TIMEX3 span extraction and attribute classification in the 2012 i2b2/VA shared task on temporal information extraction. The same approach is implemented in MedTime (Y.-K. Lin et al. 2013) where the authors devised features from Heideltime output and UMLS®. The system achieved a score of 88.0% in i2b2 2012 temporal relation challenge for the time expression extraction task.

Tapi Nzali et al. (2015) investigated temporal expression extraction in French across three domains (news, historical and medical). They devised a CRF-based system that takes domain independent features as input. They also used the output of Heildeltime in the feature set. Among other results, they showed that adapting preprocessing (e.g. tokenization) to the targeted domain yields a significant performance improvement.

Velupillai et al. (2015) built a UIMA pipeline using ClearTK (Bethard et al. 2014) and SVM classifiers for TIMEX3 extraction in the 2015 edition of Clinical TempEval. The authors used

simple features including POS tags, orthographic, and gazeteer information based partly on Heideltime output.

H.-J. Lee et al. (2016) used a HMM-SVM sequence tagger (Joachims et al. 2009) to extract TIMEX3 spans and attributes. They used a combination of several features including lexical, syntactic, discourse-level, word representation and gazeteers features. They also derived features from the output of SUTime (A. X. Chang and Manning 2012).

Khalifa et al. (2016) used the output of cTAKES to generate morphological, lexical and syntactic-level features. They tested two machine learning algorithms: CRF and SVM for TIMEX3 span extraction and attribute classification.

Cohan et al. (2016) implemented a system based on CRF to extract TIMEX3 spans. They devised morphological features and also made use of brown clustering. Attributes were extracted with a logistic regression classifier using similar features as the one used for span detection.

2.5.2 Event Extraction

In this section, we focus on event extraction in the context of temporal information extraction as defined in Section 2.3.2.

Supervised Machine Learning Approaches.

Most approaches for event extraction are data-driven and use supervised machine learning algorithms. Bethard and Martin (2006) presented STEP (System for Textual Event Parsing), a system for TimeML event recognition based on a SVM. The authors used a rich set of features including textual, morphological, syntactic, temporal and WordNet features (hypernymy relations). March and T. Baldwin (2008) presented a system for event recognition and classification based on a SVM. Features include tokens and POS tags in a window around the considered word. The authors also implemented feature reduction by removing stop words.

TempEval Shared Tasks. There were significant advances during the TempEval shared tasks. Jung and Stent (2013) investigated the usefulness of various features for the task. They made use of simple common features such as POS tags, tokens and lemmas but also semantic role labels. Bethard (2013) proposed the same approach that has been used for temporal expression extraction and classification. Based on a SVM classifier, the author used a small set of features derived from POS tags, tokens and syntactic constituency parses. The model is integrated in the tool ClearTK. Kolya et al. (2013) combined features extracted from Wordnet (hypernyms, hyponyms and other semantic relations) with features derived from semantic role labels for training a CRF. The authors did not obtain good performance on the task. NavyTime (Chambers 2013) extracted a minimalist set of features derived from tokens, POS tags and syntactic parses to train a maxent classifier. Finally, KUL (Kolomiyets and Moens 2013) is a system based on a multi-label logistic regression classifier. The features are derived from dependency and constituency parse trees.

Clinical Domain. In the clinical domain, most systems used a sequence labeling machine learning algorithm for event span extraction and a SVM for attribute classification. K. Roberts

et al. (2013) derived features from Brown clustering (P. F. Brown et al. 1992) on large biomedical and non-biomedical corpora and used them for both event span extraction and event attribute classification tasks. Y.-K. Lin et al. (2013) extracted features from Wikipedia and Metamap (Aronson 2001). Kovačević et al. (2013) derived semantic features from cTAKES (Savova et al. 2010).

In the Clinical TempEval campaigns, Velupillai et al. (2015) used the same UIMA module created for temporal expression extraction. H.-J. Lee et al. (2016) used a the same architecture as the one use for time expression extraction (HMM-SVM sequence tagger) and used similar features. For event attribute classification, they trained three SVM classifiers for each of the three attributes (modality, degree and polarity) with similar features extracted in a window of 5 words around the event. Khalifa et al. (2016) used the output of cTAKES to generate morphological, lexical and syntactic-level features. They tested two machine learning algorithms: CRF and SVM for EVENT span extraction and attribute classification. Finally, Leeuwenberg and Moens (2017) used a SVM classifier for span and attributes extraction. They rely on a small set of features based on POS and token form. They considered single tokens as event candidates.

2.5.3 Relation Extraction

Temporal relation extraction is mostly performed via supervised machine learning approaches although some rule-based methods gave interesting results on TempEval corpora

Rule-Based Approaches. There are a few rule based systems. Mani et al. (2003) devised a rule-based model to anchor event to times and obtained good result on the TimeBank corpus. Hagège and Tannier (2007) obtained the best performance on event-event relations during the first edition of TempEval with a rule-based system based on the custom parser XIP.

Data-driven Approaches. The most effective algorithms for temporal relation extraction implement data-driven approaches. Boguraev and Ando (2005) presented an algorithm for temporal relation identification and classification on the TimeBank corpus. The model jointly predicts both link and labels based on features derived from a finite state parser. One limitation of the work is that the authors considered only EVENT–TIMEX3 links. The authors noticed that limiting the distance between two relation arguments to four tokens gave the best results. Mani et al. (2006) proposed a supervised approach for EVENT–EVENT and EVENT–TIMEX3 relation extraction. The features are derived from the text and entity attributes. The authors also devised simple features describing how well tenses and aspects of both relation elements are compatible. Hepple et al. (2007) experienced with several classifiers and several attributes in order to find to which extent attributes contribute to correctly classify temporal relations during the first edition of TempEval. They found that tense and aspect were not as useful in EVENT–TIMEX3 relation classification as they are for EVENT–EVENT relation classification.

Clinical Domain. In the clinical domain, K. Roberts et al. (2013) used a SVM-ranker to detect links between pairs of entities and a multi-class SVM to assign a type to these links. Cherry et al. (2013) further divided the two i2b2 tasks in four subtasks: anchoring EVENT

to the section time, extraction intra-sentence EVENT–TIMEX3 relations, extracting inter-sentence OVERLAP relations between EVENTS and determining causal relation induced TLINKS. They devised a set of features including surface, syntax, semantic and structural features and use a Maximum Entropy algorithm to learn a model. The idea of subdividing the tasks was also adopted by other teams (Grouin et al. 2013; Xu et al. 2013). Y.-C. Chang et al. (2013) used features devised from a rule based approach for TLINK extraction in a MaxEnt component. Nikfarjam et al. (2013) implemented the same approach but used a SVM classifier instead of a MaxEnt.

Concerning the Clinical TempEval shared tasks, Velupillai et al. (2015) used token-level features partly derived from cTAKES output. Both temporal relation extraction tasks were addressed by using a CRF. H.-J. Lee et al. (2016) divided the narrative container identification task into six problems according to whether the link is between two EVENTs or an EVENT and a TIMEX3 and whether the link concerned entities within one sentence, in adjacent sentences or separated with more than two sentences. They trained a SVM classifier to identify if a pair of ordered entities is linked with a narrative container. In order to take into account the imbalance of the training set, they used the transitive closure of containment relations and filtered candidate pairs that are unlikely to have a TLINK. They devised a set of filters based on the modality and document creation time relation attributes of the entities and on whether the entities are in the same section. They also applied cost-sensitive learning and assigned a weight to each class during learning. The feature are derived from cTAKES and include POS tags of the entities, tense of the sentence verb, section information and sentence type. The authors also added the distance between the two entities when applicable.

Khalifa et al. (2016) used the output of cTAKES to generate morphological, lexical and syntactic-level features. They tested two machine learning algorithms: CRF and SVM for document creation time relation classification. For narrative container relation extraction, they trained four models for intra- and inter-sentence EVENT–EVENT, EVENT–TIMEX3 relations. They did not apply any filter for negative instance sub-sampling. Cohan et al. (2016) used a logistic regression classifier to assign a document creation time relation to each EVENT entity. They used similar features as the ones used for EVENT and TIMEX3 attribute classification and add specific features derived from the sentence and nearby time and date mentions. For narrative container relation identification, the authors used the semantic frames of the sentence. By doing so, they only consider intra-sentence relations. They derived features from a semantic role labeler and a dependency parser.

Temporal relation extraction has also been addressed using neural networks. Dligach et al. (2017) tested Convolutional and Recurrent Neural Networks for containment relation extraction. They focus on intra-sentence relations. They tested several input sequences: token, POS sequences and a combination of both. They showed that using only POS tags instead of tokens for EVENT–TIMEX3 relation extraction impact only a little the performance of the system. Furthermore, they showed that Recurrent Neural Networks (CNNs) give better performance that LSTMs on this task. They also highlighted that the neural model is not able to give better performance than a traditional feature based system and concluded that the model may not be able to generalize well on the input for this relation type.

Global Constraint Approaches. The vast majority of approaches make independent decisions on pair of entities and do not take into account the dependencies of the local decisions

in the temporal graph. Chambers and Jurafsky (2008) devised a system capable of generating consistent temporal graph. They limited the scope to *before* and *after* relations. They performed training set enhancement with closure and folding over an extended set of relations. Yoshikawa et al. (2009) used Markov logic network to model global constraints for relation typing. Bramsen et al. (2006) used Integer Linear Programmingng (ILP) for global optimization of the temporal graph. They restrained the relation set to *before* and *after* relations. Denis and Muller (2011) used point algebra instead of interval algebra to model the relations within the TimeBank corpus. This allows for temporal relation set reduction while keeping most of the information. The authors used ILP to find a optimal temporal graph.

Chapter 3

Feature-Based Approach for Temporal Relation Extraction: Application to French and English Corpora

The material presented in this chapter is based on three publications: one at the 2016 edition of the SemEval workshop (Tourille et al. 2016b), one at the 2016 edition of the TALN conference (Tourille et al. 2016a) and one at the 2017 edition of the EACL conference (Tourille et al. 2017c).

3.1 Introduction

In this chapter, we focus on the extraction of temporal relations between clinical events, temporal expressions and document creation times. More specifically, we address intra- and inter-sentence narrative container relation extraction between events and/or temporal expressions and DCT relation extraction between events and documents. We only consider the situation where gold events and temporal expressions are given.

For containment relation extraction, the objective is to identify temporal inclusion relations between pairs of entities (event and/or temporal expression) formalized as narrative container relations (cf. Chapter 2, Section 2.3.3). For DCT relation extraction, the objective is to temporally locate events according to the DCT of the documents in which they are mentioned. DCT relations include *before*, *before-overlap*, *overlap* and *after*.

Feature-based approaches have proven to be competitive with recent neural models (Bethard et al. 2016; Bethard et al. 2017). In this chapter, we describe a feature-based machine learning approach for containment and DCT relation extraction. We evaluate our approach on the THYME corpus in the context of the 2016 edition of the Clinical TempEval shared task (Bethard et al. 2016).

Most approaches presented in the literature have been devised for processing clinical documents written in English. However, temporal information extraction in clinical text is an active research area in many other languages (Névéol et al. 2018). In the second part of this chapter, we adapt our approach for clinical narratives written in French and experiment on the MERLoT corpus.

The remainder of this chapter is organized as follows. In Section 3.2, we describe the datasets that were used in our experiments. Specifically, we address the differences between the two annotation sets. In Section 3.3, we present our approach for containment and DCT relation extraction. Section 3.4 presents our participation to the 2016 edition of Clinical TempEval. In Section 3.5, we present our experiments using a similar model for both English and French clinical texts. Finally, we close the chapter with a conclusion (Section 3.6).

3.2 Data

In this chapter, we use the MERLoT (Campillos et al. 2018) and THYME (Styler IV et al. 2014a) corpora. As we described in Chapter 2, the definition of a clinical event is slightly different in each corpus. According to the annotation guidelines of the THYME corpus, a clinical event is anything that could be of interest on the patient's clinical timeline. It could be for instance a *medical procedure*, a *disease* or a *diagnosis*. For the MERLoT corpus, clinical events are described according to UMLS® Semantic Groups and Semantic Types (McCray et al. 2001). Several categories are considered as events: *disorder*, *sign or symptom*, *medical procedure*, *chemical and drugs*, *concept or idea* and *biological process or function*.

The THYME corpus comprises several attributes for clinical events. However, not all of them were used during the Clinical TempEval challenges. In the version provided to the participants, there were five attributes given to each event: *Contextual Modality* (actual, hypothetical, hedged or generic), *Degree* (most, little or n/a), *Polarity* (pos or neg), *Type* (aspectual, evidential or n/a) and *DocTimeRel* (before, before-overlap, overlap or after). Events in the MERLoT corpus carry only one *DocTime* attribute (before, before-overlap, overlap or after).

Both corpora are annotated with temporal expressions that take the form of TIMEX3 elements. Each TIMEX3 carries one *type* attribute. For the THYME corpus this attribute can take the following values: date, time, duration, quantifier, prepostexp or set. For the MERLoT corpus, the possible values are: date, time, duration or frequency. More information about THYME and MERLoT entity attributes can be found in Chapter 2, Section 2.4.

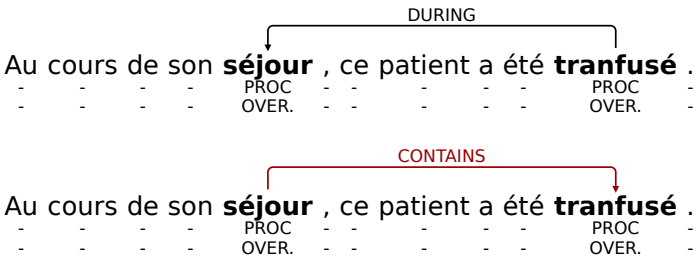

Figure 3.1: Example of the transformation of a DURING relation into a CONTAINS relation in the MERLoT corpus.

We noted earlier that the THYME and MERLoT corpora do not have the same temporal relation set (cf. Chapter 2, Section 2.4). The THYME corpus is annotated following the THYME-TimeML specification and therefore embraces the concept of narrative container to annotate temporal relations between medical events and/or TIMEX3. The MERLoT corpus does not explicitly cover narrative container relations. However, some transformations can be undertaken to build an equivalence with the THYME corpus. In this work, we consider that *during* relations are equivalent to *contains* relations (Figure 3.1). In addition, we also consider that *reveals* and *conducted* relations imply *contains* relations. Furthermore, the MERLoT corpus does not cover inter-sentence relations (relations that can spread over multiple sentences), which is the case for the THYME corpus.

The second type of relation is the one existing between events and the DCTs. Both corpora are annotated with this relation which takes the form of an attribute on event elements. Possible values are: before, before-overlap, overlap and after.

In the rest of this chapter, we will refer to clinical events as EVENTs and to containment relation as CONTAINS relations. Corpus statistics for both MERLoT and THYME corpora are presented in Table 3.1 and Table 3.3. For the THYME corpus we present separated counts for each subpart of the corpus, including the 30-document part related to brain cancer patients that was released during the 2017 edition of Clinical TempEval. For the MERLoT corpus, we present the number of contains relations created through the transformation process described above. Only a subset of all documents are annotated with CONTAINS relations in the THYME corpus. We report statistics on this corpus part in Table 3.2.

	Colon			Brain
	Train	Dev	Test	30-doc
# of documents	293	147	151	30
DOCTIME	322	168	168	7
EVENT	38,937	20,974	18,990	2,557
SECTIONTIME	284	123	154	8
TIMEX3	3,833	2,078	1,952	350
CONTAINS	11,248	6,226	5,930	788

Table 3.1: THYME corpus descriptive statistics (all documents).

	Colon			Brain
	Train	Dev	Test	30-doc.
# of documents	195	98	100	20
DOCTIME	219	115	114	5
EVENT	31,902	17,245	15,503	2,124
SECTIONTIME	278	122	153	4
TIMEX3	3,756	2,035	1,917	306
CONTAINS	11,248	6,226	5,930	788

Table 3.2: THYME corpus descriptive statistics including only the documents annotated with narrative container relations.

# of documents	500
EVENT	18,127
TIMEX3	3,940
CONTAINS	4,295

Table 3.3: MERLoT corpus descriptive statistics

3.3 Model Overview

Document Creation Time Relation Subtask. We treated the subtask as a supervised classification problem where each EVENT is classified into four categories (*before*, *before-overlap*, *overlap* or *after*). The features used for the classification are described in the appropriate sections below.

Containment Relation Subtask. Similarly to the DCT relation subtask, we cast the containment relation subtask as a supervised classification problem and more particularly, as a binary classification task applied to pairs of EVENT and/or TIMEX3 entities. As we highlighted in the previous section, the MERLoT corpus does not cover inter-sentence relations. However these relations are annotated in the THYME corpus and were included in the Clinical TempEval challenge tasks (Bethard et al. 2015; Bethard et al. 2016; Bethard et al. 2017). This inclusion of both types of relations brings a difficulty related to training instance creation. Indeed, considering all possible pairs of entities for building the training set without any scope restriction would lead to unbalanced training examples where the negative examples largely outnumber the positive examples. Hence, some choices have been made to reduce the number of negative examples.

The analysis of the training corpus shows that a large majority – around 76% – of the CONTAINS relations are intra-sentence relations, which means that the problem of their scope actually occurs for one quarter only. The remaining relations, called inter-sentence relations, spread over at least two sentences. Given this context, we have built two separate classifiers, the first one for the intra-sentence relations, the second one for the inter-sentence relations. This distinction has two main advantages: first, it reduces drastically the number

win.[a]	# of rel.[b]	total[c]
1	13,304	13,304 (76.30%)
2	1,463	14,767 (84.69%)
3	752	15,519 (89.00%)
4	497	16,016 (91.85%)
5	364	16,380 (93.94%)
6	151	16,531 (94.80%)

[a] Sentence window
[b] Number of CONTAINS relations
[c] Cumulative count of CONTAINS relations

Table 3.4: CONTAINS relations according to sentence window size in the THYME corpus. Window of size 1 corresponds to the intra-sentence level.

of negative examples, which produces better results as we observed on our development set; second, the intra-sentence classifier can benefit from a larger and richer set of features coming from sentence-level linguistic analyzers.

Concerning the inter-sentence relations, considering all pairs of EVENT and/or TIMEX3 would still give us a very large amount of negative examples. We first observed that all of them were contained within sections (no relation overlaps section boundaries). Within the scope of a section, we further noticed that inter-sentence relations within a 3-sentence window covered approximately 89% over all existing relations. A wider window would bring too much noise while giving us a very small bump on coverage. Table 3.4 shows the number of covered relations according to the size of the window, expressed in the number of sentences. The first line corresponds to the intra-sentence level (window=1). These statistics are computed on the *train* and *dev* parts of the THYME corpus. Following these observations, we decided to limit the scope of inter-sentence relations to a 3-sentence window without crossing any section boundary.

To further reduce the number of candidates for both inter- and intra-sentence classifiers in the MERLoT and THYME corpora, we transformed the 2-category problem (*contains vs. no-relation*) into a 3-category classification problem (*contains*, *is-contained* or *no-relation*). Instead of considering all permutations of events within a sentence or a sentence-window, we considered all pairs of events from left to right, changing when necessary the *contains* relations into *is-contained* relations. This strategy allowed us to divide by a factor of two the number of candidates.

Some entities are more likely to be containers. By example TIMEX3 entities are, by nature, potential containers. This is also the case for some clinical events. For instance, a *surgical operation* may contain other events such as *bleeding* or *suturing*. It will not be the same with the two latter in most cases. Following this observation, we have built a model to classify entities as being a potential *container* or *not*. As we will show in Section 3.4.3, this classifier obtains a high accuracy. We used its output as feature for our intra- and inter-sentence classifiers.

Finally, we developed a rule-based module to capture specific inter-sentence CONTAINS relations within the THYME corpus. There are some strong regularities in the handling of laboratory results where the first temporal expression contains all the results, which are expressed with EVENTs. The module we have built aims at capturing these inter-sentential regularities with the use of rules.

To summarize, our system is composed of four modules:

1. **Container detection module**: entities are classified according to whether or not they are the source of one or more CONTAINS relations;

2. **Intra-sentence relation module**: combinations of entities within sentences are considered (relations *contains*, *is-contained* or *no-relation*);

3. **Inter-sentence relation module**: combinations of entities within a 3-sentence window are considered. We use the same relation classes as those used for intra-sentence relations;

4. **List detection module**: specific laboratory results written as lists are handled via manual rules.

Lexical Feature Representation. We implemented two strategies to represent the lexical features in both subtasks. In the first one, we used the plain forms of the different lexical features. In the second strategy, we substituted the lexical forms with word embeddings. For English, these embeddings have been computed on the MIMIC III corpus (A. E. W. Johnson et al. 2016). Concerning the French language, we used the whole collection of raw clinical documents from which the MERLoT corpus has been built. In both cases, we computed[1,2] the word embeddings using the word2vec (Mikolov et al. 2013) implementation of gensim (Řehůřek and Sojka 2010). We used the max (Section 3.5) or the mean (Section 3.4) of the vectors for multi-word units. Lexical contexts are thus represented by 200-dimensional vectors. When several contexts are considered (e.g. right and left), several vectors are used.

3.4 Evaluation on the THYME corpus

We evaluated our approach in the context of the second phase of the 2016 edition of the Clinical TempEval shared task (Bethard et al. 2016). In this phase of the competition, participants were provided with gold EVENT and TIMEX3 entities and were asked to extract CONTAINS and DCT relations. Our submission used the full pipeline presented in the previous section, including the inter-sentence relation modules.

1. Parameters used during computation (Section 3.4): algorithm=CBOW; min-count=5; vector-size=200; window=20.
2. Parameters used during computation (Section 3.5): algorithm=CBOW; min-count=5; vector-size=200; window=10.

3.4.1 Preprocessing and Feature Extraction

We applied a four-step preprocessing on the 440 texts that were provided for the subtasks. First, we used NLTK (Bird and Loper 2004) to segment the texts into sentences with the *Punkt Sentence Tokenizer* pre-trained model for English provided within the framework.

The second step consisted of parsing the resulting sentences. For this task, we used the BLLIP Reranking Parser (Charniak and M. Johnson 2005) and a pre-trained biomedical parsing model (McClosky 2010). This step gave parse trees, POS and Coarse-grained Part-Of-Speech (CPOS) tags.

In the third step, we lemmatized the corpus using BioLemmatizer (Liu et al. 2012), a tool built for biomedical literature processing. We used the POS tags from the previous step as parameters for the lemmatization.

The last step consisted in using Metamap (Aronson 2001) to detect biomedical events and linking them, after disambiguation, to their related Semantic Types and Semantic Groups. Semantic Types and Groups are sets of subject categories (organized as a tree) that are used to categorize concepts in the UMLS® Metathesaurus. There are currently 133 types and 133 groups (Bodenreider 2004). We chose to keep biomedical entities that had a span overlapping with at least one EVENT of the gold standard.

For both tasks, we used a combination of structural, lexical contextual features yielded from the corpora and the preprocessing steps. We selected these features based on previous research efforts on temporal information extraction (Bethard et al. 2015; UzZaman et al. 2013; Verhagen et al. 2007; Verhagen et al. 2010). These features are presented in Table 3.5.

3.4.2 Algorithm Selection

A grid search strategy was applied to select the most appropriate machine learning algorithm and its hyperparameters. When using plain lexical forms of the tokens, three algorithms were considered in our search: Random Forests, Linear SVM (liblinear) and SVM with a RBF kernel (libsvm). For the second strategy, using pretrained word embeddings, we only considered the Linear Support Vector Machine for the CONTAINS relation extraction task and Random Forests for the DCT relation extraction task.

For both strategies, 5-fold cross-validation was used to choose the algorithm and its hyperparameters. Preliminary experiments on reducing the feature set revealed that selecting the most informative features may result in a higher performance. Following this observation, we implemented statistical feature selection as part of the grid search for the strategy using plain lexical forms of the tokens, reducing progressively the number of attributes, using ANOVA F-test.

The machine learning algorithms used for the final submission are presented in Table 3.6 together with their parameters and the percentage of the feature space kept after statistical feature selection. We used the Scikit-learn (Pedregosa et al. 2011) machine learning library for implementing our classification models and performing statistical feature selection.

3.4.3 Results

We present the cross-validation accuracies of our DCT and CONTAINER models over the development corpus in Table 3.7. For the DCT model, we obtain high performance with plain

Feature	DCT	CONTAINER[c]	CONTAINS	
			Intra	Inter
Entity type		✓	✓	✓
Entity form	✓	✓	✓	✓
Entity attributes	✓	✓	✓	✓
EVENT Semantic Types and Semantic Groups[b]	✓	✓	✓	✓
Entity Lemmas	✓	✓	✓	
Entity POS and CPOS tags	✓	✓	✓	
Do the entities contain other entities?				✓
Do the entities are contained by other entities?				✓
Container model output for the two considered entities			✓	
Syntactic paths between the two considered entities[a]			✓	
Sentence entity forms	✓	✓		
Sentence entity types	✓	✓	✓	
Sentence entity attributes	✓		✓	
Sentence EVENT Semantic Types and Semantic Groups	✓	✓	✓	
Sentence entity POS and CPOS tags	✓	✓		
Sentence entity lemmas	✓	✓		
Sentence token lemmas	✓	✓		
Sentence token POS and CPOS tags	✓	✓		
Section entity forms	✓			
Section entity types	✓			
Section entity lemmas	✓			
Section EVENT Semantic Types and Semantic Groups	✓			
Section entity POS and CPOS tags	✓			
Section entity attributes	✓			
Center context entity types				✓
Center context entity attributes				✓
Center context EVENT Semantic Types and Semantic Groups				✓
Center context entity container model outputs				✓
Sentence position in section	✓			✓
Sentence position in document	✓			
Number of containers at the sentence level			✓	
Number of entities between the considered entities			✓	✓
Number of tokens between the entities			✓	
Number of entities before and after at the sentence level		✓		
Number of entities before and after at the section level	✓			
Number of entities before and after at the document level	✓			

[a] Several paths are considered when the entities spread over more than one token.
[b] Semantic Types and Semantic Groups of the medical entities that have been detected by Metamap and that share a span overlap with the considered EVENTs.
[c] Classifier that predicts whether an entity is the source of one or more CONTAINS relations.

Table 3.5: Features used by our classifiers.

Strategy	Classifier	Algorithm	Parameters	% feat.[a]
Plain text	CONTAINER[c]	SVM (RBF)	C=10, gamma=0.01	60
	CONTAINS INTRA	SVM (RBF)	C=10, gamma=0.01	60
	CONTAINS INTER	SVM (RBF)	C=1000, gamma=0.01	100
	DCT	SVM (Linear)	C=1, tol[b]=0.0001, normalization=l2, loss function=hinge	100
Word embeddings	CONTAINER[c]	LinearSVM	C=1, tol[b]=0.01, normalization=l2, loss function=hinge	100
	CONTAINS INTRA	SVM (Linear)	C=1, tol[b]=0.01, normalization=l2, loss function=squared hinge	100
	CONTAINS INTER	SVM (Linear)	C=1000, tol[b]=0.01, normalization=l2, loss function=hinge	100
	DCT	Random Forests	max features=auto, criterion=entropy, estimators=100	100

[a] Percentage of feature space kept for final submission (using ANOVA F-test)
[b] Tolerance for stopping criteria
[c] Classifier that predicts whether an entity is the source of one or more CONTAINS relations.

Table 3.6: Machine learning algorithms and hyperparameters used for our final submission to the 2016 edition of the Clinical TempEval shared task (Bethard et al. 2016).

lexical features. However, using words embeddings gives lower performance. Concerning the CONTAINER model, we obtain high performance with both strategies.

We submitted two runs with our system for the CONTAINS and DCT relation extraction tasks, one for each strategy (plain text or word embeddings). The results for both subtasks are presented in Tables 3.8 and 3.9.

Model	Plain text	Word embeddings
DCT	0.873	0.778
CONTAINER[a]	0.917	0.924

[a] Classifier that predicts whether an entity is the source of one or more CONTAINS relations.

Table 3.7: DCT and CONTAINER model accuracies on the development corpus.

Strategy	ref.	pred.	corr.	P	R	F1
Plain text	18,990	18,989	14,603	0.769	0.769	0.769
Word embeddings	18,990	18,989	15,317	0.807	0.807	**0.807**

Table 3.8: DCT relation extraction subtask: evaluation script output. We report the number of gold standard relations (ref.), the number of predicted relations (pred.), the number of correct predictions (corr.), precision (P), recall (R) and f1-score (F1).

Strategy	ref.	pred.	corr.	P	R	F1
Plain text	5,894	3,755	2,642 2,570	0.704	**0.436**	**0.538**
Word embeddings	5,894	2,544	1,911 1,889	**0.751**	0.320	0.449

Table 3.9: CONTAINS relation extraction subtask: evaluation script output. We report the number of gold standard relations (ref.), the number of predicted relations (pred.), the number of correct relation with and without temporal closure (corr.), precision (P), recall (R) and f1-score (F1).

Concerning the DCT relation extraction subtask, we obtained above-median scores (median score: 0.724) for both runs. The second run, which relies on word embeddings to represent the lexical features of the EVENT entities, achieves better performance. These results are consistent with what was observed during the cross-validation process using the development set. The fact that the second strategy achieves the best performance is however in contradiction with the scores obtained during cross-validation, where plain text features performed best.

In the CONTAINS relation extraction subtask, we obtained above-median F1 for the first run (plain lexical features) and median scores for the second run (word embeddings) (median score: 0.449). Using plain lexical features gives us a more balanced system than using

word embeddings. With a F1 of 0.538, our system achieves performance close to the best system (0.573), thus validating our modeling choices. These results are consistent with those we obtained when testing against the development part of the corpus. The reasons for the decrease in recall when using the second strategy are however unclear and need further investigation.

Overall, our two submitted runs achieved good performance and ranked third and fifth on twenty submitted runs. We reproduce the score table presented in Bethard et al. (2016) in Appendix A (Table A.1).

3.4.4 Discussion

Our feature-based approach for DCT and CONTAINS relation extraction allows for high performance in both tasks. The fact that plain text lexical features gives a lower performance for the DCT relation extraction subtask is surprising as the top 2 best systems of the shared task used similar features and obtained scores above 0.80 F1 (Khalifa et al. 2016; H.-J. Lee et al. 2016). However, these approaches limit the size of the window in which features are extracted ([-5;+5]). Our approach did not implement any filtering, which may introduce noise in the classifier.

Although we obtained a high performance for CONTAINS relation extraction, the gap between our best run and the best performing system of the shared task remains large (0.035 F1). We identify several areas of improvement. First, we did not implement any filtering on entity pairs. For instance, H.-J. Lee et al. (2016) discarded event pairs that have contradictory modality and doctimerel attribute values. They also apply heuristic rules to further filter pairs when considering inter-sentence relations. Implementing a similar approach could increase the overall performance of our system.

Second, our machine learning approach did not use any weighting scheme to account for class imbalance within the corpus. The THYME corpus is highly unbalanced, especially for inter-sentence relations. This can result in a biased model toward the majority class, the negative one in our case.

Finally, applying a classical grid search approach for hyperparameter optimization can lead to an under-efficient model. The optimal value for a given hyperparameter can fall between its possible values defined in the range. Random search (Bergstra and Bengio 2012) or a tree-structured parzen estimator approach (Bergstra et al. 2011) may be more suitable as they allow to not make any assumptions on hyperparameter value sets besides their boundaries.

3.5 Temporal Relation Extraction in the Clinical Domain: Adapting the Approach to French Clinical Text

The main motivation for this research effort is to evaluate whereas our feature-based approach can be used for other languages than English, provided that the different language-sensitive resources along our preprocessing pipeline are replaced by equivalent resources in the target language. We experiment on the THYME and MERLoT corpora.

Similarly to our participation to Clinical TempEval, we focused on temporal relation extraction and use the gold entities provided within the two corpora. We discarded inter-sentence containment relations as they are not annotated in the French dataset. The MERLoT corpus

has been transformed into a comparable corpus according to the process described at Section 3.2. The number of DCT relations per class for both corpora is presented at Table 3.10.

	THYME (en)	MERLoT (fr)
Before	29,170	1,936
Before-Overlap	4,240	2,643
Overlap	37,091	12,211
After	8,400	1,337

Table 3.10: MERLOT (fr) and THYME (en) corpora: DCT relation distribution.

3.5.1 Preprocessing and Feature Extraction

The THYME corpus was preprocessed using cTAKES (Savova et al. 2010), an open source natural language processing system for the extraction of information from electronic health records. We extracted several features from the output of cTAKES: sentences boundaries, tokens, POS tags, token types and Semantic Types of the entities that have been recognized by cTAKES and that have a span overlap with at least one EVENT entity of the THYME corpus.

Concerning the MERLoT corpus, no specific NLP pipeline exists for French clinical texts; we thus used Stanford CoreNLP system (Manning et al. 2014) to segment and tokenize the text. We also extracted POS tags. As the corpus already provides a type for each EVENT, there is no need for detecting other clinical information.

For both DCT and CONTAINS relation extraction tasks, we used a combination of structural, lexical and contextual features yielded from the corpora and the preprocessing steps. The choice of feature is inspired by research efforts in the temporal information extraction domain (Bethard et al. 2015; UzZaman et al. 2013; Verhagen et al. 2007; Verhagen et al. 2010). These features are presented in Table 3.11.

3.5.2 Experimental Setup

We divided randomly the two corpora into train and test set following the ratio 80/20. As we mentioned in the Section 3.4.4, using a traditional grid-search approach for hyperparameter optimization may result in an under-efficient model. Following this observation, we performed hyper-parameter optimization using a tree-structured parzen estimator approach (Bergstra et al. 2011), as implemented in the library Hyperopt (Bergstra et al. 2013), to select the hyper-parameter C of a Linear SVM, the lookup window around entities and the percentile of features to keep. For the latter we used the ANOVA F-value as selection criterion. We used the SVM implementation provided within Scikit-learn. In each case, we performed a 5-fold cross-validation. For the container classifier and contains relation classifier, we used the f1-score as performance evaluation measure. Concerning the DCT classifier, we used the accuracy.

Feature	DCT	CONTAINER[c]	CONTAINS
Entity type	✓	✓	✓
Entity form	✓	✓	✓
Entity attributes	✓	✓	✓
Entity position (within the document)	✓	✓	✓
Container model output			✓
Document Type[a]	✓	✓	✓
Contextual entity forms	✓	✓	✓
Contextual entity types	✓	✓	✓
Contextual entity attributes	✓	✓	✓
Container model output for contextual entities			✓
POS tag of the sentence verbs	✓	✓	
Contextual token forms (unigrams)	✓	✓	
Contextual token POS tags (unigrams)	✓	✓	
Contextual token forms (bigrams)[b]	✓	✓	
Contextual token POS tags (bigrams)[b]	✓	✓	

[a] Information available only for the MERLoT corpus.
[b] Only when using plain lexical forms.
[c] Classifier that predicts whether an entity is the source of one or more CONTAINS relations.

Table 3.11: Features used by our classifiers.

3.5.3 Results

Cross-validation results on the training corpus are presented in Table 3.12. DCT and CONTAINS relation extraction task results on the test set are presented respectively in Table 3.13 and Table 3.14. For both tasks, we present a baseline performance. For the DCT relation extraction task, the baseline predicts the majority class (*overlap*) for all EVENT entities. For the CONTAINS relation extraction task, the baseline predicts that all EVENT entities are contained by the closest TIMEX3 entity within the sentence in which they occur.

3.5.4 Discussion

There is a gap of 0.04 in performance between the French (0.83) and English (0.87) corpora for the DCT relation extraction task. We notice that results per category are not homogeneous in both cases. Concerning the MERLoT corpus, the score obtained for the category *overlap* is better (0.90) than the score obtained for *before-overlap* (0.69), *before* (0.69) and *after* (0.73). Concerning the THYME corpus, the performance for the category *before-overlap* (0.66) is clearly detached from the others which are grouped around 0.85 (0.88 for *before*, 0.84 for *after* and 0.89 for *overlap*). This may be due to the distribution of categories among the corpora. Typically, the performance is lower for the categories where we have a lower number of training examples (*before-overlap* for the THYME corpus and categories other than *overlap* for the MERLoT corpus).

Concerning the CONTAINS relation extraction task, results are separated by a 10 point gap (0.65 for the MERLoT corpus and 0.53 for the THYME corpus). Results obtained for the THYME corpus are coherent with those presented in Section 3.4 on the Clinical TempE-

	DCT		CONTAINER		CONTAINS		CONTAINS w/o CONTAINER	
Corpus	Plain	W2V	Plain	W2V	Plain	W2V	Plain	W2V
MERLOT (fr)	0.830 (0.008)	0.785 (0.006)	0.837 (0.004)	0.776 (0.014)	0.827 (0.007)	0.799 (0.012)	0.724 (0.011)	0.670 (0.016)
THYME (en)	0.868 (0.002)	0.797 (0.006)	0.760 (0.007)	0.678 (0.031)	0.751 (0.003)	0.702 (0.013)	0.589 (0.006)	0.468 (0.018)

Table 3.12: Cross-validation results over the training corpus for all tasks. We report f1-score for CONTAINER and CONTAINS tasks and accuracy for DCT task. We also report standard deviation for all models (in brackets).

	MERLOT (fr)			THYME (en)		
	P	R	F1	P	R	F1
baseline	0.67	0.67	0.67	0.47	0.47	0.47
before-overlap	0.68	0.69	0.69	0.73	0.60	0.66
before	0.81	0.60	0.69	0.88	0.88	0.88
after	0.79	0.69	0.73	0.84	0.84	0.84
overlap	0.88	0.92	0.90	0.88	0.90	0.89
micro-average	0.83	0.84	0.83	0.87	0.87	0.87

Table 3.13: DR task results over the test corpus. We report precision (P), recall (R) and f1-score (F1) for all relation types.

	MERLOT (fr)			THYME (en)		
	P	R	F1	P	R	F1
baseline	0.43	0.15	0.22	0.55	0.06	0.11
no-relation	0.99	1.00	0.99	0.96	0.98	0.97
contains	0.75	0.57	0.65	0.61	0.47	0.53
micro-average	0.98	0.98	0.98	0.93	0.94	0.93

Table 3.14: CR task results over the test corpus. We report precision (P), recall (R) and f1-score (F1) for all relation types.

val 2016 evaluation corpus[3]. We increased the recall value in comparison to their results (from 0.436 to 0.47) but this measure is still the main point to improve.

The scores obtained by our CONTAINS relation classifier without the use of the CONTAINER classifier output suggest that this feature is valuable for the classification decision with a drop ranging from -0.103 F1 to -0.234 F1.

More globally, the best results of the Clinical TempEval shared task were 0.843 (accuracy) for the DCT relation extraction task and 0.573 (f1-score) for the CONTAINS relation extraction task, which are comparable to our results (0.87 for the DCT task and 0.53 for the CONTAINS task).

Results presented in Table 3.12 indicates that replacing lexical forms by word embeddings seems to have a negative impact on performance in every case. This performance drop need further investigation.

As for the difference of performance according to the language, several parameters can affect the results. First, the sizes of the corpora are not comparable. The THYME corpus is bigger and has more annotations than the MERLoT corpus. Second, the quality of annotations is more formalized and refined for the MERLoT corpus. This difference can influence the performance, especially for the CONTAINS relation extraction task. Third, the lack of specialized clinical resources for French can negatively influence the performance of all classifiers.

Concerning the quality of annotations, it has to be pointed out that Inter-Annotator Agreement (IAA) for temporal relation is low to moderate: in the MERLoT corpus, IAA measured on a subset of the corpus is 0.55 for *During* relations, 0.32 for *conducted* relations and 0.64 for *reveals* relations. In the THYME corpus, IAA for *contains* relation is 0.651. The inter-annotator agreement is comparable in both languages, and suggests that temporal relation extraction is a difficult task to perform, even for humans.

Overall, we managed to obtain comparable results for both English and French datasets, suggesting that our approach can be applied to at least these two languages by replacing language sensitive resources in the preprocessing pipeline.

3.6 Conclusion

In this chapter, we have presented a feature-based approach focusing on the extraction of temporal relations between clinical events, temporal expressions and document creation times from clinical notes written in English and in French. This approach, based on feature engineering, obtained competitive results with the state-of-the-art at the time of publication and led to two main observations.

Our feature engineering approach can be applied with comparable results to two different languages, English and French in our case, by changing language dependent resources in the pipeline.

In the next chapter, we investigate the use of neural networks for temporal information extraction from clinical narratives. Specifically, we try to combine the benefits of both feature-based and neural approaches by incorporating categorical features in our model. We focus on both entity and relation extraction.

3. Similarly to our evaluation corpus for English, the Clinical TempEval 2016 evaluation corpus was extracted from the THYME corpus but the two corpora are different.

Chapter 4

Neural Approach for Temporal Information Extraction

The material presented in this chapter is based on three publications: one at the 2017 edition of the Clinical TempEval workshop (Tourille et al. 2017a), one at the 2017 edition of the ACL conference (Tourille et al. 2017b) and one at the 2018 edition of the LOUHI workshop (Tourille et al. 2018).

4.1 Introduction

In this chapter, we investigate the two tasks related to temporal information extraction in clinical narratives: entity extraction (event and temporal expression) and temporal relation extraction (DCT and CONTAINS relations). More specifically, we devise a supervised approach using neural networks for entity and relation extraction, and a supervised approach based on a linear SVM for event attribute classification.

As we mentioned in the introduction of this thesis, annotated corpora are often packing a rich attribute set describing the entities. In this chapter, we investigate how these features can be used in neural approaches and how they will impact the performance of such approaches. Our neural-based classifier leverages categorical features extracted from the corpus itself and from the output of cTAKES. These features are used in combination with classical word and character-level embeddings. We begin by studying how categorical features can be used to further improve the performance of our containment relation extraction module. Then, we evaluate our approach in the context of the 2017 edition of the Clinical TempEval shared task (Bethard et al. 2017).

The remainder of this chapter is organized as follows. In Section 4.2, we quickly describe the dataset that was used in our experiments. In Section 4.3, we present the models for entity, attribute and temporal relation extraction. We also describe how input embeddings are built. Section 4.4 describes our work on the influence of categorical features on the performance of our system. Section 4.5 presents the model evaluation in the context of the 2017 edition of the Clinical TempEval shared task. We conclude in Section 4.6.

4.2 Data

In this chapter, we use exclusively the THYME corpus for our experiments. Similarly to Chapter 3, we focus on the version of the THYME corpus that was proposed to the participants during the Clinical TempEval challenges (Bethard et al. 2015; Bethard et al. 2016). This version comprises the train, dev and test subparts of the colon cancer section of the corpus.

EVENTs are given five attributes: *Contextual Modality*, *Degree*, *Polarity*, *Type* and *DocTimeRel*. TIMEX3s elements have only one *Class* attribute. The temporal relation set is limited to the containment relation, expressed as CONTAINS relations in the dataset. Furthermore, the challenge addressed both intra- and inter-sentence relation extraction. More information about the corpus can be found in Chapter 2. Corpus statistics can be found in Chapter 3, Table 3.1.

4.3 Model Overview

Our model is composed of three components: entity extraction (Section 4.3.1), EVENT attribute and DCT relation extraction (Section 4.3.2), and containment relation extraction (Section 4.3.3).

4.3.1 Entity Extraction

Our approach relies on LSTMs (Hochreiter and Schmidhuber 1997) and is inspired by recent research efforts in sequence labeling for NER (Huang et al. 2015; Lample et al. 2016; X. Ma and Hovy 2016). The architecture of our model is presented in Figure 4.1. For a given sequence of tokens, represented as vectors, we compute representations of left and right contexts of the sequence at every token (including the token itself). These representations are computed using two LSTMs (forward and backward LSTM in figure 4.1). Then, these representations are concatenated and linearly projected to a n-dimensional vector representing the number of categories. Finally, we add a CRF layer to take into account the previous label during training and prediction. Token labels follow the IOB (*Inside*, *Outside*, *Begin*) annotation scheme commonly used for NER.

Noticing a lack of efficient neural sequence labeling tools in the community, we decided to pack our module into a open source tool called YASET (Tourille et al. 2018). This tool reaches state-of-the-art performance on various NER tasks. It is available online[1].

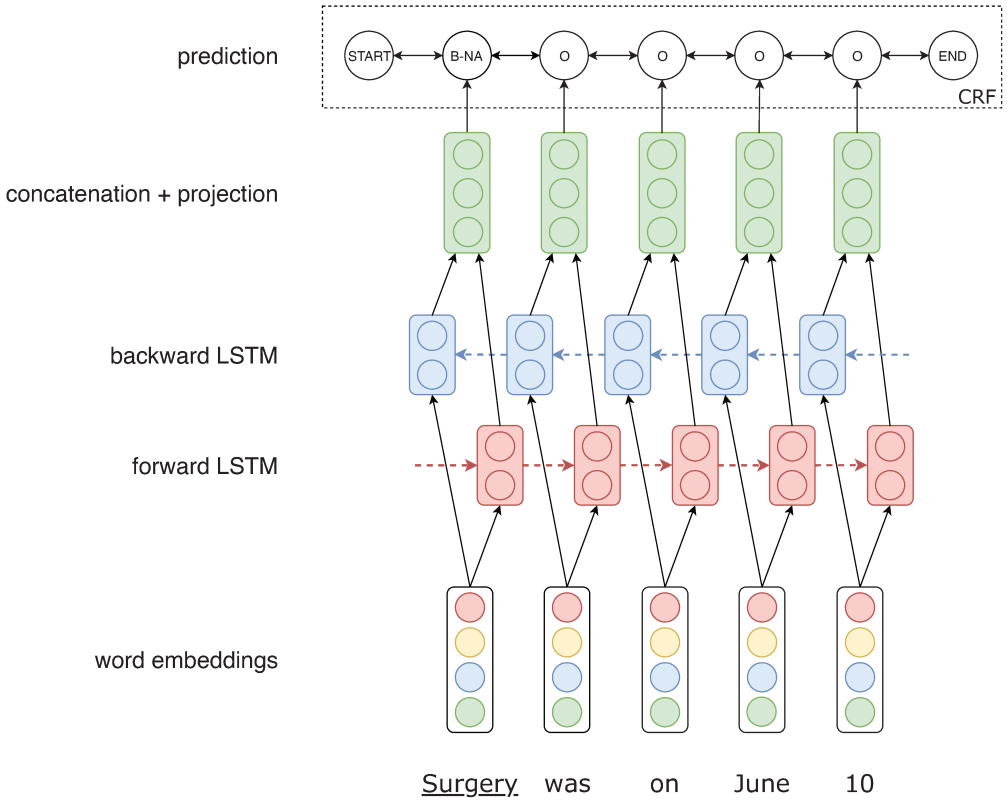

Figure 4.1: Neural architecture for entity extraction. In this example, each token of the sentence is assigned a label following the IOB format. In the example, the event *Surgery* is marked as a medical event with the type *N/A*.

1. https://github.com/jtourille/yaset

4.3.2 Event Attribute and Document Creation Time Extraction

Feature-based approaches have proven to be efficient for attribute classification on the THYME corpus (Bethard et al. 2015; Bethard et al. 2016). These approaches rely on handcrafted features extracted from the sentence context. Following these previous research efforts, we devised a feature-based approach using a SVM.

We treated each EVENT attribute extraction subtask as a supervised classification problem. We built a common pipeline for all attributes based on a linear SVM (Figure 4.2). First, a feature vector is extracted around each entity in a window of size w. Then, we retain only the k best features, ranked via their ANOVA F-score. The last step is the classification. The classifier has two hyperparameters. The parameter C allows for penalizing more or less classification mistakes. The parameter l allows to choose the loss function (*hinge* or *squared hinge*). Hyperparameter optimization is addressed by using a tree-structured parzen estimator approach (Bergstra et al. 2011) as implemented in Hyperopt[2] (Bergstra et al. 2013). The same pipeline is used for the DCT relation extraction subtask.

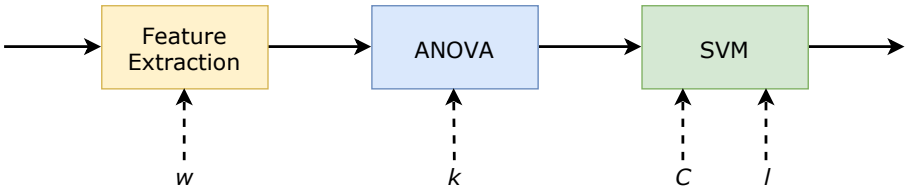

Figure 4.2: Pipeline for entity attribute classification.

4.3.3 Containment Relation Extraction

Similarly to the entity extraction module, our approach relies on LSTMs. The architecture of our model is presented in Figure 4.3. For a given sequence of tokens separating two entities (EVENT and/or TIMEX3), represented as vectors, we compute a representation of the context between the two concerned entities by going from left to right in the sequence (forward LSTM in Figure 4.3).

As LSTMs tend to be biased toward the most recent inputs, this implementation would be biased toward the second entity of each pair processed by the network. To counteract this effect, we compute the reverse representation with an LSTM reading the sequence backwards, from right to left (backward LSTM in figure 4.3). By doing so, we keep as much information as possible about the two entities.

The two final states are then concatenated and linearly transformed to a n-dimensional vector representing the number of categories (concatenation and projection in Figure 4.3). Finally, a softmax function is applied.

Similarly to the approach described in Chapter 3, we build two separate classifiers for intra- and inter-sentence relations and limit the scope of the inter-sentence classifier to relations that do not span over more than three sentences. For inter-sentence relations, a special token SENT is added to the input sequence to mark sentence boundaries.

2. https://github.com/hyperopt/hyperopt

Furthermore, as we did in the previous chapter, we transform the 2-category problem (*contains*, *no-relation*) into a 3-category problem. For each combination of EVENT and/or TIMEX3 from left to right, three cases are possible:

- the first entity temporally *contains* the second entity,
- the first entity *is* temporally *contained* by the second entity,
- there is no temporal containment relation between the entities.

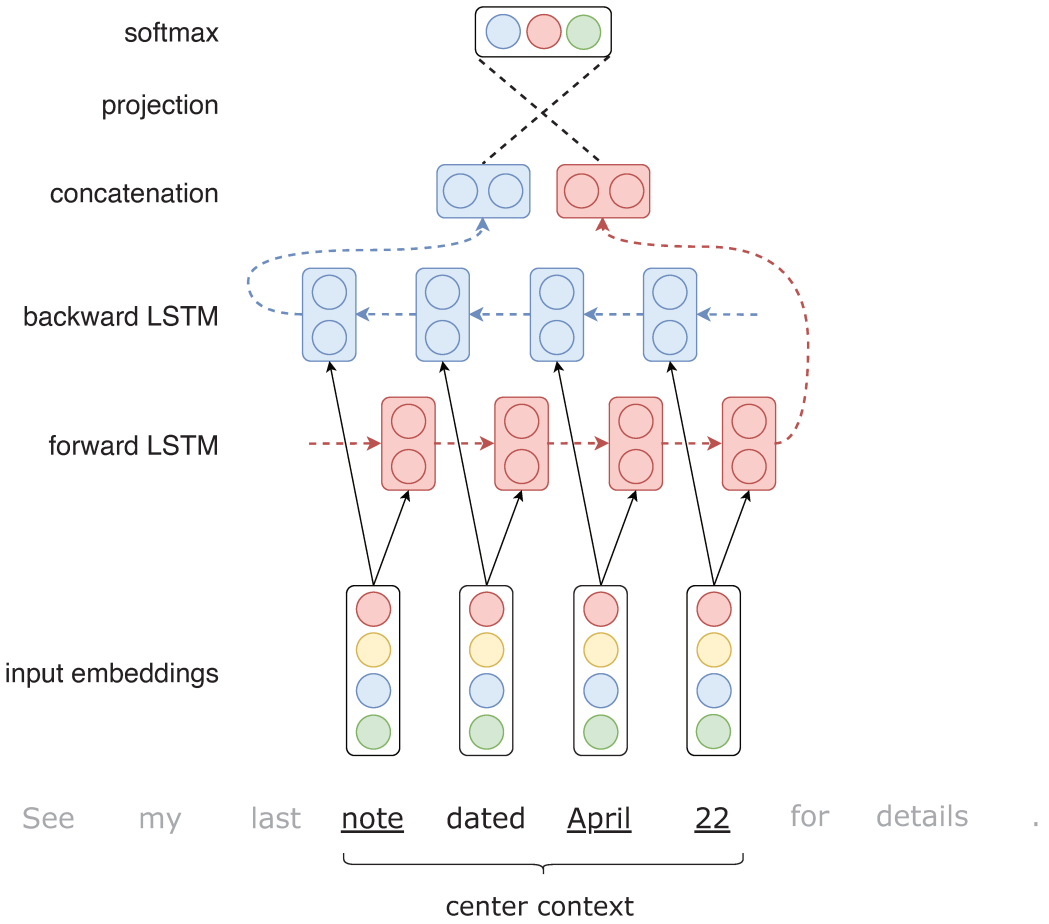

Figure 4.3: Neural architecture for containment relation extraction.

4.3.4 Input Word Embeddings

Vectors representing tokens are built by concatenating a character-based embedding, a word embedding and one embedding per categorical feature. While the word embedding is a classical option in the context of neural models, using embeddings for categorical features is a way of integrating in such model features that have been demonstrated as useful in previous work. Finally, temporal clues such as verbs, and more particularly their tense, which are important in assessing if one entity temporally contains another, are taken into account by our character-based representation of tokens.

An overview of the embedding computation is presented in Figure 4.4. Following Lample et al. (2016), the character-based representation is constructed with a Bi-LSTM. First, a random embedding is generated for every character present in the training corpus. Token characters are then processed with a forward and backward LSTM similar to the one we use in our general architecture. The final character-based representation is the result of the concatenation of the forward and backward representations.

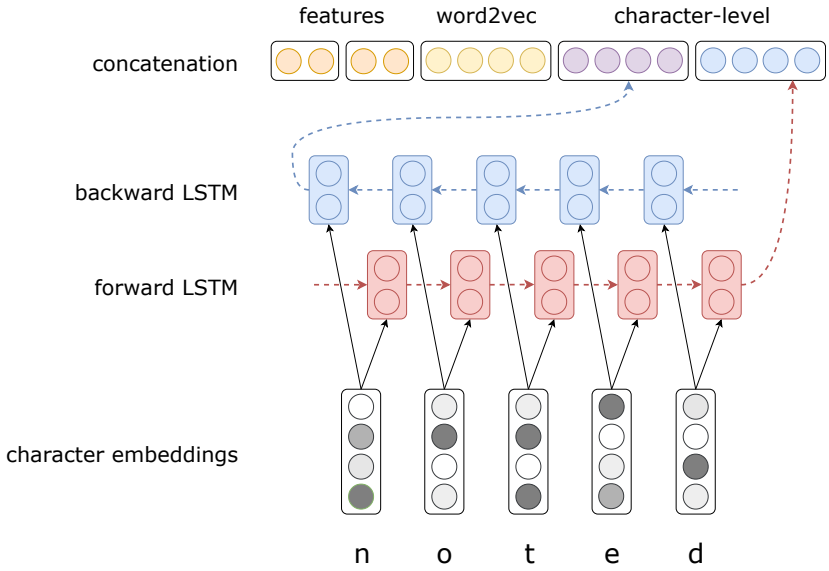

Figure 4.4: Neural architecture for word embedding creation.

4.4 Influence of Categorical Features

Our first set of experiments aims at measuring how categorical features influence the performance of our CONTAINS relation extraction approach.

4.4.1 Preprocessing

Our experimental setup mimics the conditions of the second phase of the 2016 edition of the Clinical TempEval challenge where participant were provided with gold entities (EVENTs and TIMEX3s). We focus on CONTAINS relation extraction and consider that the DCT relation is given as gold attribute. We experiment on the colon cancer part of the THYME corpus. More specifically, we use the train and dev part of the corpus section as training corpus. We measure the performance on the test part of the corpus with the official evaluation script provided during the challenges.

We preprocessed the corpus using cTAKES. We extracted sentence and token boundaries, as well as token types and semantic types of the entities that have a span overlap with a least one gold standard EVENT of the THYME corpus. This information was added to the set of gold standard attributes available for EVENTs in the corpus. An overview of the attributes available for each token is presented in Table 4.1.

Source	Attribute	Values
Corpus	Contextual Modality	Actual, Hypothetical, Hedged, Generic or no-value
	Degree	Most, Little, N/A or no-value
	Polarity	Pos, Neg or no-value
	Type	Aspectual, Evidential, N/A or no-value
	DocTimeRel	Before, Before-Overlap, Overlap, After or no-value
	Entity	EVENT, TIMEX3 or no-entity
cTAKES	Entity Type[a]	DiseaseDisorderMention, LabMention, MedicationEventMention, MedicationMention, ProcedureMention, SignSymptomMention or no-value
	Semantic Type[a]	list of semantic types extracted from the training corpus or no-value

[a] If the token is not part of an EVENT span, the value is automatically *no-value*.

Table 4.1: Attributes available for each token of the corpus after preprocessing.

4.4.2 Experimental Setup

We implemented the two neural models described in the previous section using TensorFlow 0.12 (Abadi et al. 2015). We trained our network with mini-batch Stochastic Gradient Descent (SGD) using Adam with a batch-size of 256. The learning rate was set to 0.001. The hidden layers of our forward and backward LSTMs have a size of 512. We kept 10% of the training corpus for a development corpus and we implemented early stopping with a patience of 10 epochs without performance improvement. Finally, we used dropout training to avoid overfitting. We applied dropout on input embeddings with a rate of 0.5.

We experimented with three configurations. In the first one, we used only word embeddings and character embeddings. In the second one, we added the feature embeddings related to the Gold Standard (GS) attributes. Finally, in a third experiment, we added the feature embeddings related to cTAKES. For each experiment, we report precision (P), recall (R) and f1-score (F1) computed with the official evaluation script[3] provided during the Clinical TempEval challenges.

4.4.3 Results

Results of the experiments are presented in Table 4.2. For comparison, we report the baseline provided as reference during the Clinical TempEval shared tasks, the results of the best system of the Clinical TempEval 2016 challenge (H.-J. Lee et al. 2016) and the best scores obtained after the challenge (C. Lin et al. 2016b) on the test portion of the corpus. Both H.-J. Lee et al. (2016) and C. Lin et al. (2016) rely on SVM classifiers using hand-engineered linguistic features.

	P	R	F1
baseline (closest)	0.459	0.154	0.231
H.-J. Lee et al. (2016)	0.588	0.559	0.573
C. Lin et al. (2016)	0.669	0.534	0.594
No features	0.646	0.568	0.605
+ GS features	**0.687**	0.549	0.610
+ cTAKES features	0.657	**0.575**	**0.613**

Table 4.2: Experimentation results. We report precision (P), recall (R) and f1-score (F1) for each configuration of our model, for the best system of the Clinical TempEval 2016 challenge (H.-J. Lee et al. 2016) and for the best result obtained so far on the corpus (C. Lin et al. 2016b).

4.4.4 Discussion

All three of our models perform better in terms of f1-score than H.-J. Lee et al. (2016) and C. Lin et al. (2016). Our two best models also outperform Leeuwenberg and Moens (2017), who report an f1-score of 0.608 using a structured perceptron. Interestingly, their model did

3. https://github.com/bethard/anaforatools

not distinguish between intra- and inter- sentence relations, but instead considered that related entities had to occur within a window of 30 tokens. We see that the addition of attribute embeddings slightly improves the overall performance of our system (+0.008). Adding the embeddings of GS features contributes to the major part of this improvement but tends to increase the imbalance between recall and precision. On the contrary, while the attribute embeddings related to cTAKES seem to have little impact on the overall performance, they tend to restore more balanced precision and recall.

The results for respectively intra- and inter-sentence relations are presented in Table 4.3. Similarly to our global results, the intra-sentence classifier benefits from the addition of feature embeddings with a small increase for GS features and only a very little improvement for cTAKES features.

The inter-sentence classifier exhibits the same trend: GS features do improve the performance. However, adding cTAKES features degrades it slightly (-0.013).

The closest work compared to ours is clearly Dligach et al. (2017) as it also heavily relies on neural models for extracting temporal containment relations between medical events. Dligach et al. (2017) tested both CNN and LSTM models and found CNN superior to LSTM. However, this work addressed intra-sentence relations only. Moreover, its LSTM model was not a Bi-LSTM model as ours and it did not include character-based or attribute embeddings. Finally, it distinguished EVENT–TIMEX3 and EVENT–EVENT relations while we have only one model for the two types of relations.

4.4.5 Perspective

From a global perspective, the work we have presented in this section shows that in accordance with a more general trend, our neural model for extracting containment relations clearly outperforms classical approaches based on feature engineering. However, it also shows that incorporating classical features in such a model is a way to improve it, even if all kinds of features do not contribute equally to such improvement. A more fine-grained study has now to be performed to determine the most meaningful features in this perspective and to measure the contribution of each feature to the overall performance, with a specific emphasis on character-based embeddings.

The place where these categorical features are embedded in our network should also be questioned. Including features related to the concerned entities in higher-level layers of our architecture could improve the performance. This possibility will be investigated in future research efforts.

Beyond a further analysis of the characteristics of our model, we are interested in three main extensions. The first one will investigate whether training two models, one for EVENT–TIMEX3 relations and one for EVENT–EVENT relations, as done by Dligach et al. (2017), is a better option than training one model for all types of containment relations as presented herein. The second extension consists in transposing the model we have defined in this work for English to French, as done in Chapter 3 for a more traditional approach based on a feature engineering approach. In the last one, we plan to explore additional strategies for CONTAINS relation extraction. For instance, adding a feature predicting whether a given EVENT entity is a container or not has proved to be useful in our feature-based approach (Chapter 3), but was not implemented in our system due to time constraints.

	Intra-sentence classifier						Inter-sentence classifier					
	ref	pred	corr	P	R	F1	ref	pred	corr	P	R	F1
No Features	4365	4529	3035	0.670	0.681	0.675	743	895	377	0.421	**0.498**	0.456
+ GS	4365	4253	2980	**0.701**	0.661	0.680	743	692	349	**0.504**	0.462	**0.482**
+ cTAKES	4365	4780	3170	0.663	**0.704**	**0.683**	743	628	305	0.486	0.408	0.443

Table 4.3: Results obtained by the intra-sentence and inter-sentence classifiers for each model of this paper. We report the number of gold standard relations (ref), the number of relations predicted by our system (pred), the number of true positives (corr), precision (P), recall (R) and f1-score (F1).

4.4.6 A Word on Temporal Coherence

Our neural approach for CONTAINS relation extraction only considers pairs of entities when making classification decisions. This could result in incoherent temporal graphs. We evaluated the degree of incoherence resulting from our approach by counting the number of cycles in the temporal graphs. Our experiments on intra-sentence CONTAINS relation extraction during development suggested that temporal graph incoherence is only minor with an average number of cycles per documents close to zero in a given iteration.

However, robust temporal graph consistency is important for clinical staff that rely on extracted information to make informed decisions. Hence, we will investigate global approaches for temporal graph creation (Bramsen et al. 2006b; Denis and Muller 2011). These approaches allows to take into account all classification decisions to generate the final temporal graph.

4.5 Evaluation on the THYME Corpus: Domain Adaptation for Temporal Information Extraction

In this second set of experiments, we evaluate our approach in the context of the 2017 edition of the Clinical TempEval challenge (Bethard et al. 2017). Similarly to the previous edition, the participants were asked to perform entity and containment relation extraction on the THYME corpus (Styler IV et al. 2014a). However, this edition included the notion of domain adaptation. Source and target domains were different. Systems were trained on documents related to *colon cancer* and were tested on *brain cancer* documents. Domain linguistic variation can decrease the performance of a given approach. As annotating a corpus is expensive and time consuming, there is a strong need for developing domain adaptation approaches for clinical NLP (Miller et al. 2017a; Miwa and Ananiadou 2015; Zhang et al. 2015).

Two experimental setup were proposed by the organizers. In the first one, no target domain annotations were provided (unsupervised domain adaptation). In the second one, a set of 30 documents related to the target domain was given to the participants (supervised domain adaptation).

4.5.1 Preprocessing

For unsupervised domain adaptation, participants were given the three parts of the colon cancer section of the THYME corpus for developing their systems (train, dev, test). For supervised domain adaptation, 30 documents extracted from the train part of the brain cancer section were given to the participants. Performance was measured on the test part of the brain cancer section.

We preprocessed the corpus using cTAKES 3.2.2. We extracted sentence and token boundaries, as well as token types and semantic types of the entities that have a span overlap with a least one gold standard EVENT of the THYME corpus. This information was added to the set of gold standard attributes available for EVENTs in the corpus. We also preprocessed the corpus using HeidelTime 2.2.1 and used the results to further extend our feature set.

4.5.2 Architecture Description

We implemented a pipeline approach using the three modules described above (Section 4.3). For entity extraction, we integrated the attributes *Type* for EVENT and *Class* for TIMEX3 in the IOB scheme.

Concerning the attribute classification, we trained a separate classifier for each of the three remaining attributes and the DCT relation based on lexical, contextual and structural features extracted from the documents:

- EVENT type attribute,
- EVENT plain lexical form,
- EVENT position within the document,
- POS tags of the verbs within the right and left contexts of the considered entity,
- EVENT POS tag,
- *Type* or *Class* of the other entities that are present within the left and right contexts,
- token unigrams and bigrams within a window around the entity.

Input vectors are built differently depending on the subtask. For the entity extraction subtask, vectors representing tokens are built by concatenating a character-based embedding and a word embedding. Whether we are dealing with EVENT or TIMEX3 entities, we add one embedding per cTAKES attribute or one embedding representing the TIMEX3 class as detected by HeidelTime. Concerning the containment relation subtask, input vectors are built by concatenating a character-based embedding, a word embedding, one embedding per gold standard attribute and one embedding for the type of DCT relation.

4.5.3 Domain Adaptation Strategies

We implemented two strategies for unsupervised domain adaptation. In the first strategy, we blocked further training of the pretrained word embeddings during network training. Since a large number of medical events mentioned in the test set are not seen during training, we believe that our system should rely on untuned word embeddings to make its prediction.

In the second strategy we randomly replaced tokens that composed EVENT entities by the *unknown* token[4]. Given the fact that our word embeddings are pretrained on the MIMIC III corpus (A. E. W. Johnson et al. 2016) and on the colon cancer part of the THYME corpus, a number of tokens (and therefore EVENTs) of the test part of the corpus may not have a specific word embedding. By replacing randomly EVENT tokens, we force our networks to look at other contextual clues within the sentence. Both strategies were applied on EVENT entity and CONTAINS relation extraction subtasks.

Supervised domain adaptation was addressed by implementing two strategies. In the first one, we mixed the 30 texts about brain cancer to the 591 texts about colon cancer. In the second one, we randomly chose 30 texts related to colon cancer and combined them to the 30 texts about brain cancer, resulting in a balanced training corpus. Both strategies were applied on EVENT, TIMEX3 and CONTAINS extraction subtasks.

4. Replacement probability = 0.2.

4.5.4 Network Training

We trained our networks with mini-batch SGD using Adam (Kingma and Ba 2015) with a batch-size of 256. The learning rate was set to 0.001. Hidden layers of our forward and backward LSTMs have a size of 256. We kept 10% of the training corpus for a development corpus and we implemented early stopping with a patience of 10 epochs without performance improvements. We use dropout training to avoid overfitting. We applied dropout on input embeddings with a rate of 0.5.

4.5.5 Results

Our model ranked first in almost all categories. Reproductions of the challenge result tables are provided in Appendix B. Table B.1 presents the results on the TIMEX3 extraction task. Although our approach has a lower performance on f1-score and precision, we obtained the best recall across all configurations. The results for EVENT extraction are presented in Table B.2. Our approach obtained the best f1-score across all configurations. Concerning the temporal relation extraction task (DCT and CONTAINS relations), our system performed best in three out of four configurations (Table B.3).

Results for our four runs are presented in Table 4.4. The two strategies implemented for unsupervised domain adaptation yield similar results (0.01 difference in f1-score at most), with only a very slight advantage for the strategy blocking further training of the word embeddings (STATIC strategy in the table).

For supervised domain adaptation, the two strategies also yield close results (0.04 difference in f1-score) for the EVENT and temporal relation extraction subtasks. However, the strategy consisting in taking all available annotations (ALL strategy in the table) outperforms slightly the training on a balanced corpus, especially for the extraction of CONTAINS relations. The same strategy seems to perform much better for the TIMEX3 entity extraction subtask where the gap in f1-score reaches 0.06. This superiority agrees the general observation that the size of the training corpus has often a greater impact on results than its strict matching with the target domain. Overall, in both phases (supervised and unsupervised domain adaptation) and for all strategies, results are competitive for entity and temporal relation extraction.

4.5.6 Discussion

The performance obtained by our system relies in part on corpus tayloring. Some sections of the test corpus related to *medication* and *diet* are not to be annotated according to the annotation guidelines. However, these sections are not formally delimited within the documents. To avoid annotating them during test time, we developed a semi-automatic approach for detecting these sections and put them aside.

Other aspects linked to the corpus limit the performance. Some sections should not be annotated as they are duplicate of other sections found in the corpus as a whole. However, we have no information on how to formally identify these sections. Furthermore, a number of temporal expressions are annotated as SECTIONTIME or DOCTIME entities. Detecting TIMEX3 entities instead decreases the precision of our model.

	Phase 1						Phase 2					
	STATIC			REPLACE			ALL			30-30		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EVENT Span	0.622	0.843	**0.716**	0.606	0.841	0.705	0.691	0.854	**0.764**	0.660	0.865	0.749
EVENT Modality	0.553	0.749	**0.636**	0.537	0.745	0.624	0.628	0.775	**0.694**	0.598	0.784	0.679
EVENT Degree	0.616	0.834	**0.708**	0.600	0.831	0.697	0.682	0.843	**0.754**	0.652	0.854	0.739
EVENT Polarity	0.603	0.816	**0.693**	0.588	0.815	0.683	0.676	0.835	**0.747**	0.644	0.844	0.731
EVENT Type	0.608	0.823	**0.699**	0.592	0.821	0.688	0.675	0.834	**0.746**	0.641	0.841	0.728
EVENT All attributes	0.374	0.507	**0.431**	0.365	0.507	0.425	0.468	0.578	**0.517**	0.440	0.577	0.500
TIMEX3 Span	0.421	0.660	0.514	0.421	0.660	0.514	0.510	0.671	**0.579**	0.452	0.621	0.523
TIMEX3 Class	0.401	0.630	0.490	0.401	0.630	0.490	0.487	0.641	**0.553**	0.430	0.591	0.498
DCT Relation	0.443	0.599	**0.509**	0.436	0.604	0.506	0.535	0.661	**0.591**	0.511	0.670	0.580
CONTAINS	0.280	0.396	**0.328**	0.264	0.408	0.320	0.244	0.438	**0.316**	0.211	0.422	0.282

Table 4.4: Results obtained by our system across our four runs. We report precision (P), recall (R) and f1-score (f1). The best f1-score performance in each phase is bolded.

Concerning the domain adaptation strategies, results of the shared task confirm that a domain shift decreases significantly the performance of a given approach. Although we obtained the best score in the shared task, our performance remains much lower than the best performance obtained during the previous edition of the shared task (Bethard et al. 2016). For instance, H.-J. Lee et al. (2016) obtained 0.479 F1 for end-to-end CONTAINS relation extraction while our approach gives 0.33 F1 in the unsupervised domain adaptation track.

Interestingly, TIMEX3 extraction has a very low f1-score in comparison to the best performance obtained when no domain adaptation is implied (Bethard et al. 2016). H.-J. Lee et al. (2016) obtained a score of 0.772 F1 while the best score obtained in the unsupervised domain adaptation track was 0.51 F1 (MacAvaney et al. 2017). This suggests that temporal expressions are either expressed differently or that their sentence context is not the same within the two corpus sections (colon cancer and brain cancer). EVENT extraction suffers from the same type of performance drop. Overall, our strategies may not be suitable for this type of domain shift. We will investigate other approaches for domain adaptation in future work.

4.6 Conclusion

In this chapter, we presented an hybrid approach for temporal information extraction in clinical narratives. We tackled both entity and temporal relation extraction tasks. Our approach relies on classical word embeddings and on categorical features embedded as dense vector in our network. We tested the effect of those features while experimenting on the THYME corpus and showed that they allow for a performance increase.

Then, we evaluated the approach in the context of the 2017 edition of the Clinical TempEval shared task, which proposed two tracks related to domain adaptation. Our system was ranked first in almost all categories. The two strategies that we implemented for unsupervised domain adaptation during the shared task performed similarly. However, for the supervised domain adaptation phase, the strategy which consist in training on a balanced dataset gives lower performance. This confirms the general observation that the size of the training corpus has a great impact on neural network performance.

Compared to our feature-based approach, the model presented in this chapter allows for a significant performance increase, suggesting that neural approaches are more efficient for temporal information extraction. However, by including categorical features within our network, we managed to improve the performance of our system. This suggests that hybrid approaches are a promising research area and should be investigated in the future.

Part II

Clinical Event Coreference Resolution

Chapter 5

Clinical Event Coreference Resolution

5.1 Introduction

This chapter discusses event coreference resolution in the clinical domain. We survey how the notion of coreference is apprehended from a linguistic perspective (Section 5.2). We present the terminology used by the NLP community (Section 5.3) and go over the differences between event coreference in the general and the clinical domains (Section 5.4). We detail the annotated corpora in the clinical domain (Section 5.5). We present the different approaches developed in the literature (Sections 5.6, 5.7 and 5.8). Finally, we describe the approaches that have been developed for coreference resolution in the clinical domain

(Section 5.9) and present the evaluation metrics used to measure coreference system performance (Section 5.10). We close the chapter with a conclusion (Section 5.11).

5.2 Anaphora and Coreference: A Linguistic Perspective

The terms *anaphora* and *coreference* are not well defined in the NLP community and are often used confusingly, even synonymously in some cases. In this section, we briefly go over the different concepts behind these terms from a linguistic perspective.

An *anaphoric* expression depends on the *linguistic context* to be correctly interpreted. Personal pronouns such as *she* or *he* (Example 34a) but also demonstrative pronouns such as *that* (Example 34b) are examples of such expressions. In other words, the interpretation of an anaphoric expression depends on the entities previously mentioned or inferred from what have been said (Poesio 2016). This phenomenon is called *anaphora*.

(34) a. Barbara is coming tonight. [She] will bring a cake.
b. The train [that] departed at 3pm is arrived.

Besides the linguistic context, the *visual context*, also called *visual deixis*, regroups entities that are not mentioned in the linguistic context but are shared by the participants of a given conversation (H. H. Clark and Marshall 1981). For instance, *the seat* in Example 35, is only known by the participants to the conversation and was not mentioned previously in the linguistic context.

(35) Could you bring back [the seat] inside?

More generally, we say that these expressions depend on the *discourse situation* or *utterance situation* in order to be interpreted (Barwise and Perry 1981). The set of entities introduced in the discourse situation is called *Universe of Discourse* in the *Discourse Representation Theory* (DRT) (Kamp and Reyle 1993). The interpretation of references to the visual context has been the focus of many research efforts (e.g. in Landragin et al. (2002)). However, this question is out of scope in our thesis and we will focus on anaphora.

Poesio (2016) identifies several expression types which depend on the linguistic context. They includes noun phrases as illustrated above but also *pro-verbs* (Example 36a), *ellipsis* (Example 36b) or *full verbal expressions* (Example 36c). In the last example, the time at which Kim *listened to her messages* is determined by the discourse situation. Examples are extracted from Poesio (2016).

(36) a. Kim is making the same mistakes that I [did].
b. Kim brought the wine, and Robin [ ] the cheese.
c. Kim arrived home. She [listened to the messages] on her answering machine.

Research in computational linguistics mainly focuses on anaphoric nominal expressions although some work has been done on ellipsis resolution (e.g. in Dalrymple et al. (1991)). These nominal expressions can be divided in four categories, according to their semantic function (Poesio 2016):

1. **Referring**: noun phrases that introduce a new entity in discourse or that rely on previously introduced entities for their interpretation.
2. **Quantificational**: noun phrases that denote a relation *part/whole* between two sets of objects. In the Example 37, the expression *few trains* is a subset of the set of *trains*.

   (37) [Few trains] arrived in time. (Poesio 2016)

3. **Predicative**: noun phrases that denote properties of objects (Example 38).

   (38) Bob is [a 43-year-old man].

4. **Expletive**: expressions that are used to fill in a verbal argument (e.g. *it* and *there* in Example 39).

   (39) a. It rains.
   b. There is a cat in the tree.

Predicative noun phrases are less dependent on the universe of discourse than other types of noun phrases (Poesio 2016). The interest given to predicative noun phrases in the literature lies in the fact that they do not introduce new entities in the universe of discourse and will therefore not be used as anchors for anaphoric expressions. This distinction is reflected in recent coreference annotation guidelines, for instance in the corpus OntoNotes (Hovy et al. 2006) used in recent shared tasks on coreference resolution (Pradhan et al. 2012; Pradhan et al. 2011; Pradhan et al. 2007).

Most of the work on anaphora resolution targets referring noun phrases and the selection of an anchor for anaphoric expressions. Referring noun phrases include *pronouns* (reflexive, definite, indefinite and demonstrative), *nominals* and *proper names*. Proper names are a special type of referring noun phrases. They are not dependent on the linguistic context. They are directly referring to an object that is encoded in their semantics. In that respect, their interpretation is not similar from the one of others noun phrases such as pronouns and nominals (Poesio 2016).

Anaphoric expression forms vary across languages. For instance, reflexives and personal pronouns can be realized as incorporated pronouns as in Example 40. In several languages such as Italian and Japanese, one argument of an anaphoric reference can sometimes be omitted. We called this phenomenon **zero anaphora** (Example 41). All examples are taken from Poesio (2016).

(40) a. Italian: Giovanni i e' in ritardo così mi ha chiesto se posso incontrar[lo] al cinema.
   b. English: John i is late so he i asked me if I can meet [him] at the movies.

(41) a. Italian: [Giovanni] andò a far visita a degli amici. Per via, Φ comprò del vino.
   b. Japanese: [John]-wa yujin-o houmon-sita. Tochu-de Φ wain-o ka-tta.
   c. English: [John] went to visit some friends. On the way, [he] bought some wine.

Up to this point, all examples were build around the notion of **identity of reference** between the referring expression and its anchor, meaning that the referring expression and its anchor refer to the same object. This is a particular case of anaphora, which is called **coreference**. One important thing to notice is that coreference does not implies anaphora. Proper names are the most idiomatic examples. Two mentions of the same person would be coreferring to the same real-world person, but do not depend on the linguistic context to be interpreted. Similarly, anaphora does not imply coreference, several other relation types exist:

- **Identity of sense**: this is the case for indefinite pronouns such as *one* or *another* which refer to a different object of the same type (Example 42). This can also be the case for definite pronouns, as for the paycheck pronouns (Example 43).

  (42) Sally admired Sue's jacket, so she got [one] for Christmas. (Garnham 2001)

  (43) The man who gave his paycheck to his wife is wiser that the man who gave [it] to his mistress. (Karttunen 1976)

- **Bound anaphora**: this is the case when the anchor is a quantified expression and the referring expression is a pronoun. In that case, the pronoun behave like a variable in a loop that gets called over the set of objects referenced by the anchor as in Example 44.

  (44) No Italian ever believes that the referee treated [his] team fairly. (Poesio 2016)

- **Associative anaphora**: the relation between the referring expression and its anchor is one of *part/whole* or *set/subset* as in Example 45. This relation type is also called **bridging anaphora**.

  (45) We saw a flat yesterday. [The kitchen] is very spacious but [the garden] is very small. (Poesio 2016)

Determining the correct relation that lies between an anaphoric expression and its anchor is not an easy task. Recasens et al. (2011) point out that even the coreference (i.e. identity of reference) can sometimes be nuanced. For instance, in Example 46, 47 and 48, the authors described the relations that link the mentions as relations of **near-identity**. Recasens et al. (2011) argue for a continuum ranging from identity to non-identity.

(46) For centuries here, [the people] have had almost a mystical relationship with Popo, believing the volcano is a god. Tonight, [they] fear it will turn vengeful. (Recasens et al. 2011)

(47) On homecoming night [Postville] feels like Hometown, USA, but a look around [this town of 2000] shows it's become a miniature Ellis Island. [This] was an all-white, all-Christian community …For those who prefer [the old Postville], Mayor John Hyman has a simple answer. (Recasens et al. 2011)

(48) "[Your father] was the greatest, but he was also one of us," commented an anonymous old lady while she was shaking Alessandro's hand—[Gassman]'s best known son. "I will miss [the actor], but I will be lacking [my father] especially," he said. (Recasens et al. 2011)

Intuitively, entities mentioned in the previous utterances of a given discourse are prominent for the interpretation of referring expressions. This observation has led to the **discourse-model hypothesis** and to **the dynamic models of discourse interpretation** (Garnham 2001; Karttunen 1976). The discourse model hypothesis explains that context dependent expressions are processed and interpreted with respect to the current state of the universe of discourse. This led to two main considerations in the literature (Poesio 2016). First, the universe of discourse is constantly updated and this update potential has to be accounted for in the models. Second, the objects included in the discourse model are not limited to the ones being mentioned. A number of entities can be constructed or inferred. In Example 49, the constructed set comprising *John* and *Mary* is the anchor of the pronoun *they*. Propositions or abstracts concepts can also be inferred such as the fact that *the court does not believe someone* in Example 50. Finally, entities may have been introduced implicitly like in Example 51 where the expression *the government* refers to *the government of Korea*.

(49) John and Mary came to dinner last night. [They] are a nice couple. (Poesio 2016)

(50) We believe her, the court does not, and [that] resolves the matter. (Poesio 2016)

(51) For the Parks and millions of other young Koreans, the long-cherished dream of home ownership has become a cruel illusion. For [the government], it has become a highly volatile political issue. (Poesio and Vieira 1998)

5.3 Definitions and Terminology: The NLP imbroglio

In the previous section, we stated that an anaphoric expression depends on the entities introduced in the universe of discourse in order to be interpreted. When the relation between the anaphoric expression and its anchor is one of identity, we say that they are coreferring.

However, the terms anaphora and coreference have been misused in the NLP literature. The term coreference was introduced during the first MUC shared task (Sundheim 1995) on entity coreference resolution. The term did not convey the same concept that we defined in the previous section. Deemter and Kibble (2000) highlighted several conception mistakes. First, non-referring expressions were used in anaphoric relations and annotators were asked to mark bridging relations as coreference relations (Example 52).

(52) a. No solution emerged from our discussions.
b. Whenever a solution emerged, we embraced it.

Deemter and Kibble (2000) also point out that predicative anaphoric expressions were marked as coreferent with their anchor. This has some implications regarding *change over time*. In the annotation guidelines (Hirschman and Chinchor 1998) for the MUC-6 and MUC-7 shared tasks, it is stated that two mentions should be marked as coreferent if the text asserts them to be coreferential at any time. Also, *Henry Higgins, sales director of Sudsy Soaps* and *president of Dreamy Detergents* should be marked as coreferent in Example 53 according to the annotation guidelines. Since *coreference implies equivalence*, the annotation implies that the sales director of Sudsy Soaps and the president of Dreamy Detergents is the same person, which is not correct.

(53) Henry Higgins, who was formerly sales director of Sudsy Soaps, became president of Dreamy Detergents.

The main conclusion of Deemter and Kibble (2000) analysis is that the relation annotated in MUC is different from the coreference relation discussed in linguistics. Anaphora, predication and coreference are annotated together under the same term coreference. The authors recommend that the next annotations guidelines reflect the distinction between them.

The observations formulated by Deemter and Kibble (2000) have been taken into account in recent coreference annotation efforts. The coreference annotations of the OntoNotes corpus (Pradhan et al. 2007) makes the distinction between identical and appositives coreference links. Furthermore, coreference annotations does not include generic, underspecified or abstract entities. However the distinction between anaphora and coreference remains loose in the sense that a coreference link will be set between two proper names as well as between a pronoun and its anchor. Both cases will be considered as examples of anaphoric coreference.

In this thesis, we borrow the terminology of the NLP community and slightly abuse the terms coreference and anaphora. Subsequently, several terms need to be defined. Consider the following example:

(54) [John]$_{e1}$ went to [[Bob]$_{e2}$'s house]$_{e3}$.
[He]$_{e1}$ gave [him]$_{e2}$ back [[his]$_{e2}$ umbrella]$_{e4}$.

Following Pradhan et al. (2011), we define the **coreference resolution task** as identifying all mentions of entities or events in text and clustering them into equivalence classes. This task does not imply to determine to which real-world entity or event the mentions are referring to, but whether the mentions are referring to same entity or event.

The text spans that are considered as potential entity or event mentions within a coreference resolution system are called **mentions**, **candidates** or **markables**. In Example 54, all bracketed expressions are candidates.

As we will explain later in this chapter, coreference resolution systems usually order the mentions according to their position in the text. Typical approaches process the mentions in order and try to draw an anaphoric coreference link between the **active mention**, i.e. the mention currently in focus and one of its **antecedents**, i.e. the mentions physically located before the active mention. The active mention is also called the **anaphor** in the literature, even if the mention itself is not anaphoric.

Following the literature, two mentions m, n are coreferent if and only if Referent(m) = Referent(n). The coreference relation is reflexive, symmetric and transitive.

A **cluster of mentions** referring to same entity is called a **coreference chain** in the literature. The term comes from the fact that if we were to sort all mentions belonging to the same cluster according to their position in a document, the resulting form would looks like a chain where each mention excepted the first one is linked to the previous one via an anaphoric coreference link. Note that coreference chains are also called **entities** in the literature.

5.4 Event Coreference Resolution

Until now, we have been talking mostly about entity coreference resolution where the objective is to determine whether two entity mentions refer to the same real-world entity. Event

coreference resolution targets events instead of entities. It is a growing field of research in NLP which is considered as more complex and more challenging than entity based coreference resolution (Lu and Ng 2018). Consider the following example:

(55) Georges Cipriani [$_{\text{EVT-1}}$ **left**] a prison in Ensisheim in northern France on parole on Wednesday. He [$_{\text{EVT-2}}$ **departed**] the prison in a police vehicle bound for an open prison near Strasbourg. (Lu and Ng 2018)

The two event mentions in Example 55 are coreferent for two reasons. First, the subtype of the events are compatible. They are both referring to a *movement* which implies the *transport* of a *person*. Second, their arguments are also compatible. The first event mention has three arguments: a person argument (*Georges Cipriani*), an origin argument (*a prison*) and a time argument (*Wednesday*). The second event mention has also three arguments: a person argument (*He*), an origin argument (*the prison*) and an instrument argument (*a police vehicle*). The two overlapping arguments are compatible (i.e. they are entity-coreferent). Both constraints (subtype and argument constraints) should be met in order for two event mentions to be coreferent (Lu and Ng 2018).

In order to perform event coreference resolution, a typical event coreference system must implement several steps. First, it must extract entities and perform entity coreference resolution. Then, event mentions should be extracted and entities should be classified as arguments. Finally, event coreference resolution may be done.

Intuitively, event coreference resolution is harder than its counterpart on entities due to error propagation in the pipeline. Moreover, as we have seen in Chapter 2, event mentions have various linguistic realizations ranging from noun phrases to verb phrases, while entity mentions are mostly noun phrases and pronouns.

Clinical Domain. Event coreference resolution in the clinical domain is different for several reasons. First, as we have seen in Chaper 2, clinical events differ from general domain events where the linguistic object of interest is mostly the verb and its derivatives (e.g. nominalizations). In the clinical domain, events may be realized by noun phrases. This is the case for the treatment and the two problems mentioned in Example 56. Thus, the task of finding markables in the clinical documents may be cast as a NER problem. This difference must be reflected in the clinical event coreference systems.

(56) She received a 7 day course of [$_{\text{TREATMENT}}$ **amoxicillin**] for [$_{\text{PRONOUN}}$ **which**] [$_{\text{PROBLEM}}$ **Enterococcus**] was sensitive but [$_{\text{PROBLEM}}$ **Klebsiella**] unknown.

Second, clinical events do not usually imply syntactically explicit arguments in the text. In Example 57a, the three problems mentioned in the sentence have a similar argument structure. They have a location argument (*Youville Hospital*), a date argument (*2020-03-10*) and an instrument argument (*MRI*). However, all three arguments are difficult to retrieve with a syntactic analysis of the sentence.

(57) a. [$_{\text{TEST}}$ **MRI lumbar spine**] (Youville Hospital 2020-03-10): [$_{\text{PROB.}}$ **L3-4 osteomyelitis**], [$_{\text{PROB.}}$ **discitis**], and [$_{\text{PROB.}}$ **epidural abscess**].

b. [$_{\text{PROB.}}$**Degenerative changes**] including [$_{\text{PROB.}}$**multilevel central spinal stenosis**].

Sometimes, the event arguments are not even in the same sentence. This is the case for the event mentions in Example 57b. The sentence is located next to the sentence presented in Example 57a. All event mentions are also related to the MRI done in Youville Hospital.

Third, event arguments may be implicit. This is the case for all event mentions in Example 57 which all have an implicit person argument, i.e. the patient. Note that sometimes the person argument may be explicit, as shown in Example 58. Fortunately, clinical documents concern only one patient, so this implicit argument does not play such an important role for coreference resolution.

(58) **She** also described [$_{\text{PROB.}}$**sob**], [$_{\text{PROB.}}$**nausea**], [$_{\text{PROB.}}$**worsening lower ext edema**].

Finally, another specificity of clinical event coreference resolution emerges when considering the downstream applications where the task could be useful. In our case, the medical staff is interested in reconstructing the patient clinical timeline based on the patient clinical documents. In this context, the temporal location of clinical events seems to play an important role for coreference resolution. Intuitively, two event mentions sharing the same semantic meaning and the same temporal location have a high probability to be coreferent. In other words, each event mention has an implicit temporal argument. Once retrieved, this argument could be used for coreference resolution.

These domain characteristics (e.g. event mention forms or absence of arguments) seems to make the clinical event coreference resolution task more similar to entity based coreference resolution. This is reflected in the clinical NLP literature where systems and methods follow the entity based coreference resolution methods developed in the general domain.

5.5 Annotated Corpora

There are two open annotated corpora available for event coreference resolution in the clinical domain. The biomedical and the general domains are also active areas for event coreference resolution. However we will not review their associated corpora in this section. An overview of these resources can be found in K. B. Cohen et al. (2017) and Lu and Ng (2018).

The i2b2 Corpus. The i2b2 corpus (Uzuner et al. 2012) is the first publicly available dataset annotated with coreference in the clinical domain. The corpus was created to foster the development of new methods for coreference resolution in clinical text. The corpus builds upon existing annotation efforts made in the context of the i2b2 challenges. Thus, it reuses previous layers of annotations.

The corpus is composed of two sub-corpora: the i2b2/VA corpus and the Ontology Development and Information Extraction (ODIE) corpus. Both corpora provide annotated clinical documents (discharge summaries and progress notes) from various institutions.

Event types annotated in each sub-corpora are not identical. The ODIE corpus (Savova et al. 2011) contains 164 documents and includes ten entity categories: anatomical site, disease or syndrome, indicator/reagent/diagnostic aid, laboratory or test result, none, organ or tissue function, other, people, procedure and sign or symptom. The i2b2/VA (Uzuner et al. 2011) corpus contains 814 documents and annotates a smaller set of entities: problem, person, pronoun, test and treatment. Statistics about the two corpora can be found in Uzuner

et al. (2012). Interestingly, the corpus addresses both entity and event based coreference by annotating coreference chains related to people. However the clinical documents involve a limited set of persons beside the patient himself. Therefore, there are a small number of person coreference chains involving a large number of mentions, mostly pronouns.

The THYME corpus. Besides the temporal annotations (Styler IV et al. 2014a), approximately 300 colon cancer documents are annotated with within-document coreference chains (Miller et al. 2017b). The main difference with the i2b2 corpus lies in the markable definition. The THYME corpus considers that all nouns, noun phrases (including relative clauses), nominal modifiers, pronouns and nominalized verbs can be considered as markables. Moreover, the annotations do not cover singletons (non-coreferent mentions) and markables are not typed. Annotators were asked to annotate *the longest and the most specific span*, including determiners and modifying information.

Another difference with other general and clinical domain coreference corpora is that the THYME corpus covers other relations besides the identity and appositive relations. It includes part/whole (Example 59a) and set/subset (Example 59b) relations.

(59) a. [$_{M1}$ **The SFA**] has severe stenotic disease. [$_{M2}$ **The popliteal segment**] has moderate disease without stenosis.

b. [$_{M1}$ **Laboratory studies**]. [$_{M2}$ **Mammogram**]. [$_{M3}$ **CT Angiogram**].

5.6 A Word on Mention Extraction

There are two variants of the coreference resolution task. In the first one, gold mentions are given and used in the coreference systems. Thus, the task does not involve mention extraction and/or classification but only determining the coreference links between mentions. In the second one, systems must start from scratch, without any mentions to reason with. This version of the coreference resolution task has become more and more popular over the years as the first one is considered too simple and does not reflect the true difficulty of the task.

Coreference resolution therefore implies mention extraction as a first step. Most approaches in the general domain use rule-based systems based on the syntactic analysis of sentences (e.g. in K. Clark and Manning (2016); Wiseman et al. (2015)). For instance, the Berkeley Coreference System (Durrett and Klein 2013) builds on H. Lee et al. (2011) to extract mentions based on sentence parse trees and the output of a NER system. The extraction is sometimes followed by a filtering step to remove unlikely coreferent mentions such the pleonastic pronoun *it*. Several systems train an anaphoricity classifier to determine whether a mention is anaphoric (Björkelund and Nugues 2011; Rahman and Ng 2009).

Rule-based approaches are also popular in the clinical domain. Grouin et al. (2011) use an analysis engine based upon regular expressions of words, rules and lexicons. They create a lexicon based on the UMLS® by selecting the relevant concept via its semantic types. Miller et al. (2017) use the output of cTAKES (Savova et al. 2010) to extract candidate mentions based on the syntactic analysis of sentences.

For both the general and the clinical domains, mention extraction systems are tuned to maximize recall. This leads to the creation of large number of candidates which have to be pruned somehow during the following processing steps.

5.7 Early Approaches for Coreference Resolution

Until the apparition of corpora in the mid 1990s and the introduction of shared tasks on the subject (Hirschman and Chinchor 1998; Sundheim 1995), most of the work on coreference resolution was theoretically inspired and rule-based.

Hobbs (1978) presented a syntax-based approach for pronoun resolution which is still often used as a baseline. The algorithm builds on the observation made in the literature that syntactic and morphosyntactic information plays a role for interpretation. This information is often modeled as constraints. For instance gender, number or person constraints.

Hobbs' algorithm traverses the surface parse tree breadth-first, from left to right. It goes backwards one sentence at a time and looks for the correct antecedent which matches the pronoun constraints (gender and number). As Poesio et al. (2016) underlines, Hobbs (1978) is one of the first who tries to implement a formal evaluation. The algorithm is tested with 100 examples.

Several approaches were based on common sense knowledge. Charniak (1972) proposed a model called DSP (Deep Semantic Processing) that takes hand-coded assertions for a group of sentences and applies deductive inferences that resolve anaphoric references. Although the system is one of the first proposing to use inferences for anaphora resolution, it suffers from several problems including lack of evaluation (Poesio et al. 2016b).

A number of research efforts are build around the notion of salience. The idea of salience is to account for the recency of entities introduced in discourse (Grosz and Sidner 1986). Hobbs (1976) showed that 90% of pronoun anchors are in the same sentence as the pronoun and 98% are in the same or previous sentences. However, choosing the closest matching antecedent do not give good results.

This observation led to the development of a framework where the notion of focus is introduced. Grosz and Sidner (1986) proposed two focus levels. The *global focus* specifies the articulation of a given discourse into segments. In other words, discourses are segmented according to topics. The second level is called the *local focus*. At this level, the authors model the fact that depending on the position in the text or in the conversation, the entities have a variable salience. The most prominent entities will be preferred in the case of pronominal anaphora.

This notion of local focus was later divided into two sub-concepts. The *discourse focus* is concerned by the discourse topic whereas the *actor focus* is concerned by the preference for pronouns in subject positions to refer to antecedents in subject positions.

This approach for accounting for the salience of entities in discourse inspired researchers in developing the *centering theory* (Grosz et al. 1995). According to this theory, every utterance updates the local focus by introducing mentions of discourse entities. These entities are ranked at each utterance and the first ranked is called the *Preferred Center*. This corresponds to the actor focus mentioned above. The theory accounts also for the discourse focus by acknowledging the existence of an object playing this role, called the *backward looking center*. There is a large number of models based on the theories of salience in the literature (Brennan et al. 1987; Lappin and Leass 1994; Strube 1998; Strube and Hahn 1999).

5.8 Supervised Approaches for Coreference Resolution

Supervised approaches for coreference resolution have given interesting results over the last two decades. We identify four main models in the literature. As it is (almost) always the case for any categorization, some models fit with difficulty in our schema, either because they are crossing several categories or because they are somehow unique. For a different view of the domain we refer the reader to other reviews that have been published on the topic (Lu and Ng 2018; Martschat 2017; Ng 2010; Ng 2017; Poesio et al. 2016a; Zheng et al. 2011).

Each model presented in the section is first described by taking the most representative research effort from the literature. Then, following Ng (2016), we describe the different model variants according to four main axes: the learning algorithm, the instance creation method, the features and the clustering algorithm.

5.8.1 Mention-Pair Model

The mention-pair model was first introduced by McCarthy and Lehnert (1995) and Aone and Bennett (1995) who focused on organizations involved in business joint ventures in news articles written in English and Japanese. The model gained popularity when Soon et al. (2001) proposed a generalization of the model by including all noun phrases.

Soon et al. (2001) cast the task as a classification problem where a classifier, a decision tree learner (Quinlan 1993), is trained to decide whether a given pair of noun phrases is coreferent or not. In this model, each pair is represented as a feature vector. Soon et al. (2001) relied on a small set of twelve features including distance, string matching and agreement features, which were, for most of them, already used in previous work (Cardie and Wagstaf 1999; Fisher et al. 1995).

This first classification step is followed by a clustering phase where local pairwise classification decisions are clustered together to form proper coreference chains. However, these local decisions can be contradicting. For instance, consider Example 60. A contradiction could arise if the pronoun *he* was wrongly classified as coreferent with both *John* and *Bob* mentions, but the two latter were classified as non-coreferent.

(60) [John]$_{e1}$ went to [Bob]$_{e2}$'s house. [He]$_{e1}$ have [him]$_{e2}$ back [his]$_{e2}$ umbrella.

Furthermore, even if we ignore the potential contradictions, several antecedents could have been marked as coreferent with a given mention. In this case, how do we choose the correct antecedent?

To answer this question, Soon et al. (2001) implemented *closest-first* clustering. In this strategy, the model chooses the closest antecedent which was classified as coreferent with the active mention. Although the strategy allows to build a coherent set of coreference chains, it does not resolve contradictions that could arise at the classifier level.

One major issue with mention pair models lies in the creation of the training instances. The distribution of positive and negative instances is highly unbalanced since the majority of mention pairs are not coreferent. Soon et al. (2001) tackled this issue by implementing a heuristic rule. Positive instances are created between anaphoric noun phrases and their closest preceding antecedent. All antecedents occurring between the anaphoric noun phrases and their closest antecedents are used to build negative instances.

Learning algorithm Several machine learning algorithms have been used to learn the pairwise classifier: decision tree learners (Quinlan 1993) (e.g. in McCarthy and Lehnert (1995); Soon et al. (2001)), memory-based learners (Daelemans and Bosch 2005) (e.g. in Hoste (2005); Recasens and Hovy (2009)), perceptrons (e.g. in Bengtson and Roth (2008); Stoyanov et al. (2009)), support vector machines (Cortes and Vapnik 1995) (e.g. in Rahman and Ng (2011); Xu et al. (2012)), maximum entropy learners (Bergert et al. 1996) (e.g. in Kehler et al. (2004)) or the RIPPER rule learner (W. W. Cohen 1995) (e.g. in (Hoste 2005; Ng and Cardie 2002a; Ng and Cardie 2002b; Ng and Cardie 2002c)).

Features The small feature set used by Soon et al. (2001) was improved and extended by Ng and Cardie (2002). The authors added a set of 41 features including new mention comparisons, grammatical constraints and semantic features. Later, the availability of large corpora allowed the inclusion of lexical features (Bengtson and Roth 2008; Björkelund and Nugues 2011)

Several research efforts considered the use of world knowledge in the model. Ponzetto and Strube (2006) used the category network of Wikipedia to build a taxonomy and built semantic relatedness features based on it. Rahman and Ng (2011) derived features based on YAGO (Suchanek et al. 2007).

Clustering A large number of research papers focused on improving the clustering step of the initial mention-pair model. While Soon et al. (2001) performed closest-first clustering, Ng and Cardie (2002) implemented *best-first* clustering where the active mention is linked to the best scoring antecedent (among those who have a coreference score above 0.5). This strategy was already proposed in Aone and Bennett (1995). Both closest-first and best-first clustering approaches are greedy in the sense that they do not take into account the relations between classification decisions. Another greedy approach, called *aggressive-merge* clustering, takes the transitive closure over all decisions (Denis and Baldridge 2009; Stoyanov et al. 2009), meaning that the active mention is clustered with all its preceding coreferent antecedents.

C. Nicolae and G. Nicolae (2006) built a graph from the output of the pairwise classifier where each edge is weighted with the classifier score. The optimal partitioning of the graph is obtained by cutting repeatedly the graph until a satisfactory stopping point is reached. Klenner (2007) used integer linear programming to enforce correct transitivity based on the output of a pairwise classifier.

The clustering step can also be made during learning. McCallum and Wellner (2005) presented a graph partitioning model where all mentions of a given cluster must be close to each other.

There are a few comparison of clustering approaches in the literature. While comparing closest-first and best-first strategies, Ng and Cardie (2002) found that the latter performs best on MUC data while Rahman and Ng (2009) observed the opposite situation. Denis and Baldridge (2008) compared closest-first and aggressive-merge clustering approaches and observed that both strategies impact the metric differently. One major weakness of all three clustering approaches, closest-first, best-first and aggressive-merge is that they rely only on local decisions taken independently.

Training instance creation The method proposed by Soon et al. (2001) where a negative instance is build for each antecedent located between the mention and the first true antecedent has been widely used in the literature (Ng and Cardie 2002c; Ponzetto and Strube 2006; Rahman and Ng 2009). In some cases, the heuristic is slightly modified. For instance, Ng and Cardie (2002) selected the closest non-pronominal antecedent for a pronominal anaphor. McCarthy and Lehnert (1995) and Stoyanov et al. (2009) used all mention pairs in their respective systems.

5.8.2 Mention-Ranking Model

The mention-pair model has two major issues. The first one is that the features extracted from the contexts of the two mentions considered at each timestep may not be sufficient for correctly making a coreference decision. The second one is that each pair of mentions is considered independently from the others and this can lead to classification contradictions. The mention-ranking model aims at addressing the second problem by ranking all possible antecedents and choosing the best one. Thus it removes the burden of selecting a clustering approach.

Yang et al. (2003) made an early attempt at designing a mention-ranking model. Their model, the *twin-candidate* model, takes each pair of mentions and predicts which one is better antecedent for the anaphor. The mention that win the most comparisons is selected as the correct antecedent for the anaphor. Although the model shows promising results, it does not take into account all antecedents at once.

The model presented in Denis and Baldridge (2008) is one of the first modern mention-ranking model. The authors consider all candidate antecedents and computed all the pairwise scores between these candidates and the anaphor. The candidate that obtain the best score is selected as antecedent. During learning, the model learns a parameter vector such that the closest correct antecedent gets a higher score compared to the others.

Denis and Baldridge (2008) used a combination of features including semantic compatibility, string similarity and morphosyntactic agreement features.

In the model, every mention must have an antecedent. However, the vast majority of these mentions are non anaphoric. Hence, Denis and Baldridge (2008) performed a classification step to determine if the active mention is anaphoric or not. One major drawback of this approach is that error made at this stage of the pipeline would propagate at the coreference resolution level.

Finally, as Denis and Baldridge (2008) chose only one correct antecedent during learning. They chose the closest correct antecedent as the reference for pronouns and the closest non-pronominal antecedent for non-prononominal mentions.

Learning algorithm In their initial model, Denis and Baldridge (2008) used a maximum entropy learner (Bergert et al. 1996) but other algorithms have been used in the literature. Rahman and Ng (2011) used support vector machines (Cortes and Vapnik 1995), K.-W. Chang et al. (2012) used an average perceptron while Yang et al. (2003) used a decision tree learner (Quinlan 1993).

Features Most papers use features similar to those presented in Ng and Cardie (2002) and Bengtson and Roth (2008). Durrett and Klein (2013) used mainly lexical features and heuristics for feature combination. Wiseman et al. (2015) used a neural network model to learn feature combinations and thus alleviated the need for engineering them. They showed performance improvement on the CoNLL 2011 shared task corpus (Pradhan et al. 2011).

Clustering In Yang et al. (2003), clustering at test time is done via tournament ranking. The antecedent that is classified as antecedent the largest number of times is selected as antecedent for the active mention.

Alternatives to the anaphoricity classification step described by Denis and Baldridge (2008) have been proposed in the literature. Among them, K.-W. Chang et al. (2012) built a dummy antecedent for the non-anaphoric case and give it the score 0 when considering a pair with the dummy mention. Alternatively, Durrett and Klein (2013) used a special feature that is triggered when the considered mention is the dummy mention.

Papers using dummy mentions for the non-anaphoric case use a cost-sensitive loss. For instance, Durrett and Klein (2013) and Wiseman et al. (2015) distinguished between *wrong link*, *false new* and *false anaphoric* errors.

Finally, Denis and Baldridge (2008) selected the closest antecedent as the correct antecedent. This approach has been embraced in a large number of research efforts. However, there are research efforts trying to learn what the best antecedent is. In K.-W. Chang et al. (2012), Durrett and Klein (2013) and Wiseman et al. (2015), the models choose the best scoring correct antecedent for the active mention (with the current model parameters) and use it as the referent antecedent.

5.8.3 Entity-Based Models

Entity-based models were introduced to overcome the limitation of mention-pair and mention-ranking models which rely only on local features from the mention contexts to make classification decisions.

One of the first entity-based approach was introduced by Luo et al. (2004). Similarly to the mention-pair model, Luo et al. (2004) processed the text in order and decide to which partially constructed cluster, or chain, the active mention has to be linked. By doing so, later coreference decisions depend on earlier ones and these decisions depend not only on the active mention and the antecedent but also on all the mentions of the entity in which the antecedent is in. In that respect, the model has been coined entity-mention model in the literature (Poesio et al. 2016a). Ng (2016) highlighted that the way this type of models incrementally processes a discourse makes them similar to earlier discourse models.

Luo et al. (2004) used several features at the mention and entity levels. It is worth noting that most of the traditional features used in the literature with the mention-pair model (Cardie and Wagstaf 1999; Fisher et al. 1995; Soon et al. 2001) can be easily transposed at the entity level via the use of logical predicates. Consider the head match feature. Luo et al. (2004) devised a entity-level feature that is triggered if any mention in the partially constructed entity matches the active mention.

Considering that best-first and closest-first clustering approaches are too greedy, Luo et al. (2004) modeled the search space as a Bell tree. They kept track of the k-best partial clustering and performed the search via beam search.

Luo et al. (2004) implemented a method similar to the one described in Soon et al. (2001) for training instance creation to reduce class imbalance. For each anaphor that does have an antecedent, a negative instance will be created between the considered anaphor and every preceding cluster only if one mention of these clusters lives between the anaphor and the closest mention of the antecedent cluster.

Similarly to the mention-pair model, the model of Luo et al. (2004) suffers from the fact that each decision is taken independently from the others and therefore fails to capture the competition between antecedents.

Learning algorithm Entity-based models have been used widely in the literature (Daumé III 2006; Daumé III and Marcu 2005a; Luo et al. 2004; C. Ma et al. 2014; Rahman and Ng 2011b; Webster and Curran 2014; Yang et al. 2008). Rahman and Ng (2009) proposed a cluster-ranking approach, where clusters are ranked in the same fashion as the mention-ranking approaches. This approach inspired Wiseman et al. (2016) who used neural networks to build partial cluster representations in an mention-ranking approach. Webster and Curran (2014) built a model inspired by shift-reduce parsing.

Several research efforts model the task as a clustering problem (K. Clark and Manning 2015; K. Clark and Manning 2016a; Culotta et al. 2007; Stoyanov and Eisner 2012). These models are also called *agglomerative-clustering* models in the literature. Entities are constructed via merge operation between partially constructed entities. As for its entity-mention counter-part, it allows to take into account information about the considered entities. In this line of approaches, the problem is modeled as a search problem where the objective is to learn the best sequence of merging operations. Similarly to Luo et al. (2004), Daumé III and Marcu (2005) modeled the coreference resolution task as a search over a Bell tree. They made use of the Learning as Search Optimization framework (Daumé III and Marcu 2005b). K. Clark and Manning (2015) and Stoyanov and Eisner (2012) implemented easy-first clustering by making easiest coreference decisions before the hardest ones. They relied on the output of a pairwise classifier for the ranking.

The task can also be modeled as a partition induction problem. Finley and Joachims (2005) trained a max-margin ranking model for learning to rank candidate coreference partitions.

Features Although the entity-mention model has not yielded any encouraging results compared to regular mention-pair or mention-ranking models, Luo et al. (2004) showed that it allows to use 20x less features and avoid coreference mistakes such as the one clustering a male and a female pronoun.

Most of the entity-mention models follows the initial work of Luo et al. (2004) by extending the usual feature set to the entity level. For instance distance features takes into account the average distance between the anaphor and the entity. However, different logical predicates can be used. For instance, a agreement feature can apply between *all*, *most*, *any* or *none* of the entity mentions and the active mention. K. Clark and Manning (2015) added the average probability of coreference between mentions in two clusters as a feature. They learned feature combination via the use of neural networks.

Yang et al. (2008) argued that the reason why the entity-mention model did not give good results is because the cluster-level features are not built adequately. Instead of devising features at the cluster level, we should build features at the mention level, by taking into

account each mention in the cluster. However, the implementation of such a model is limited by the fact that machine learning algorithm usually takes a fixed size input vector. Hence, the author used Inductive Linear Programming (ILP) to address this issue. They showed a performance improvement when comparing an entity-mention model learned via ILP versus a mention-pair model learned via ILP.

Clustering Best-first and closest-first clustering strategies that were previously used for mention-pair models can be implemented for entity-mention models in a similar fashion. Agglomerative clustering approaches learn from scratch and compare current merging decisions with good merging decisions. For Stoyanov and Eisner (2012), good decisions are those who lead to an increase of the evaluation metric.

Training instances Most of the literature working on entity-mention models follow the method presented in Luo et al. (2004) for the creation of training instances. Rahman and Ng (2011) did not follow this schema and created one instance between the active mention and all the partially constructed clusters.

Culotta et al. (2007) employed first order predicates to construct cluster-level features. The naive way of creating one positive instance for each subset of mentions is untractable. Hence, the authors set up error driven sampling where the instances are created from the errors the model makes on the training instances. First they initialize randomly, then perform agglomerative clustering until a mistake is made. They update the weights and repeat for the next document. When the model makes an error, a negative instance is created based on the resulting clustering. A positive instance is created by merging two coreferent clusters. Update is performed via ranking so that the positive instance gets a higher score than the negative one.

5.8.4 Tree-Based Models

Tree-based modeling of the coreference resolution task was initially introduced by Yu and Joachims (2009) but gained popularity after Fernandes et al. (2012) won the CoNLL-2012 coreference shared task (Pradhan et al. 2012). The *antecedent tree* encodes all pairwise decisions at the document level. The objective in this model becomes to predict a coreference tree. One advantage of this structure is that its enforces that there is only one antecedent per mention. The non-anaphoric case is modeled by a dummy mention, located at the root of the tree. All subtrees are thus entities.

Learning algorithm The antecedent-tree model was used in several research efforts with light modifications (Björkelund and Kuhn 2014; K.-W. Chang et al. 2013; Lassalle and Denis 2015). Yu and Joachims (2009) did not use a dummy mention to model the non-anaphoric case and therefore the output of their model is a set of trees instead of a single tree. They used latent structural support vector machines as learning algorithm.

Fernandes et al. (2014), Björkelund and Kuhn (2014) and Lassalle and Denis (2015) used Latent structured perceptron (Collins 2002; X. Sun et al. 2009) while K.-W. Chang et al. (2013) used stochastic gradient descent.

Features Most models make use of standard features including lexical and combination features. They do not devise any feature for the dummy antecedent with the exception of Lassalle and Denis (2015) who devised features for anaphoricity detection that they use for the dummy antecedent.

K.-W. Chang et al. (2013) and Lassalle and Denis (2015) used must-link and cannot-link constraints and applied them either during training (Lassalle and Denis 2015) or inference (K.-W. Chang et al. 2013). The latter did observe a performance increase.

Björkelund and Kuhn (2014) allowed for non-local features such as mention types in an entity. They applied left-to-right decoding via beam-search.

Training instances Most models do not use any resampling but implement cost-sensitive loss functions. Yu and Joachims (2009) rewarded coreferent linking by 1 and penalized linking two non-coreferent mentions by -1[3]. Fernandes et al. (2014) and Björkelund and Kuhn (2014) used a simpler cost function: edges between two non-coreferent mentions have cost 1 while erroneous links with dummy mention is 1.5. Lassalle and Denis (2015) and K.-W. Chang et al. (2013) penalized all wrong links with cost 1.

5.9 Coreference Resolution in the Clinical Domain

Literature on coreference resolution in the clinical domain is rather sparse. It was initiated with the i2b2 challenge on coreference resolution (Uzuner et al. 2012). Participating teams submitted rule-based, hybrid and supervised machine learning systems. Several systems tried to use world knowledge derived from the corpus itself or from external resources such as the UMLS® Metathesaurus, Wikipedia or Wordnet (Fellbaum 1998).

Grouin et al. (2011) modeled the task as a three-step process and devised an mention-pair model. First, they discarded singletons by using a SVM with some handcrafted features. Next, they classify each candidate mention pairs with a combination of string matching patterns and rules. They created a knowledge base from the training corpus which contains all pairs of words that are coreferent. Finally, mention partition is obtained via closure on the coreference graph.

Xu et al. (2012) split the task into three subtasks: one related to people, one related to pronouns and the other related to clinical events. They built a binary classifier to determine whether a mention related to people is a mention of the patient. They selected attributes ranging from mention position to string matching and obtained a high performance (f1-score: 0.996). Concerning the clinical event mentions, they extracted world-knowledge features from Wikipedia, WordNet and other knowledge databases. They combined these features with semantic, grammatical and lexical features. Final partitioning is done via best-first clustering using the SVM confidence scores.

Outside the i2b2 shared task, Jindal and Roth (2013) built a semantic representation of the medical mentions by extracting their UMLS® concept IDs via MetaMap (Aronson 2001). They also derived mention types via Wikipedia. The authors added a normalization step and used a deterministic rule-based algorithm to select the best antecedent. For coreference resolution of mentions referring to people, they devised a two-layer algorithm which separates the mention set in three parts: mentions corresponding to patients, mentions corresponding to any of the doctors and the rest of the mentions.

To the best of our knowledge, there was no attempt do devise an entity-based approach for coreference resolution in the clinical domain. More information about the i2b2 shared task can be found in the original paper (Uzuner et al. 2012).

5.10 Evaluation Metrics

Measuring coreference resolution approach performance is hard and is an active research area. The traditional precision, recall and f1-score metrics used in the NLP literature do not fit well for this task. A metric should be able to capture how well a given approach has clustered mention together compared to the gold standard. In this section, we will present the different metrics that are used to measure performance.

As we present the different metrics used in the literature for coreference resolution evaluation, we will use an example borrowed from Pradhan et al. (2014) to illustrate the computation of these metrics:

- the key (K) is composed of two entities with mentions $\{a, b, c\}$ and $\{d, e, f, g\}$;
- the response (R) contains three entities with mentions $\{a, b\}$, $\{c, d\}$ and $\{f, g, h, i\}$;
- the mention e is missing from the response and mentions h et i are spurious.

5.10.1 The MUC Score

The MUC score was proposed by Vilain et al. (1995) for performance evaluation in the sixth and seventh MUCs (Hirschman and Chinchor 1998; Sundheim 1995). It is a link-based metric that looks at how the system outputs partitions in comparison to the reference. The larger the number of components in these partitions is, the worse the performance should be.

The first step in computing the MUC score is to create partitions with respect to the key and response respectively as shown in Figure 5.1.

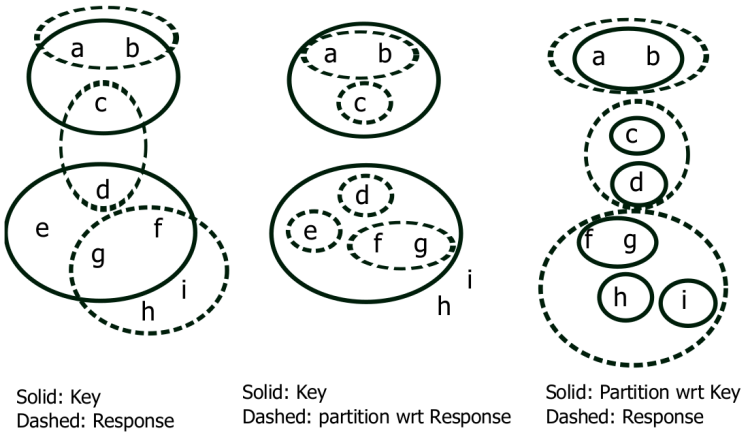

Figure 5.1: Reproduction of Figure 1 from Pradhan et al. (2014). Original caption: "Example key and response entities along with the partitions for computing the MUC score."

Applied to our example, MUC precision, recall and f1-score are computed as following:

$$\text{Recall} = \frac{\sum_{i=1}^{N_k}(|K_i| - |p(K_i)|)}{\sum_{i=1}^{N_k}(|K_i| - 1)} = \frac{(3-2)+(4-3)}{(3-1)+(4-1)} = 0.40 \tag{5.1}$$

$$\text{Precision} = \frac{\sum_{i=1}^{N_r}(|R_i| - |p'(R_i)|)}{\sum_{i=1}^{N_r}(|R_i| - 1)} = \frac{(2-1)+(2-2)+(4-3)}{(2-1)+(2-1)+(4-1)} = 0.40 \tag{5.2}$$

$$\text{F1-Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.40 \times 0.40}{0.40 + 0.40} = 0.40 \tag{5.3}$$

where:

- K_i is the i^{th} key entity and $p(K_i)$ is the set of partitions created by intersecting K_i with response entities (cf. the middle sub-figure in Figure 5.1);
- R_i is the i^{th} response entity and $p'(R_i)$ is the set of partitions created by intersecting R_i with key entities (cf. the right-most sub-figure in Figure 5.1);
- N_k and N_r are the number of key and response entities, respectively.

The MUC metric suffers two major shortcomings (Bagga and B. Baldwin 1998). First, it is unable to reward successful identification of singleton clusters as the metric was devised for the MUC datasets which do not annotate singleton clusters. Second, the metric does not consider the impact of coreference errors. Intuitively, a mistake involving a large cluster is more damaging than a mistake involving a smaller one and the MUC metric fails to capture that aspect.

5.10.2 The B^3 Algorithm

The B^3 (Bagga and B. Baldwin 1998) metric was devised to address the shortcomings of the MUC score. It is a mention-based metric which allows to take into account mention cluster sizes. It was originally build to evaluate cross-document coreference resolution (Ng 2017).

The main idea behind recall computation is that each mention is assigned a credit computed as the ratio between the number of correct mentions in the predicted entity (which contains the key mention) to the size of the key entity. The final recall score is the sum of all credits normalized by the sum of key mentions. The precision is computed by switching key and response in the computation.

The original paper from Bagga and B. Baldwin (1998) fails to specify how to score predicted mentions. The examples in the paper describe the case where predicted mentions are the same as the key mentions. Moreover, the authors did not provide any reference implementation, making it difficult to compute the score in this situation. This led to various implementations in the literature. Stoyanov et al. (2009) introduce the notion of *twinless mention* which denotes mentions which are either spurious or missing from the predicted mention set. The authors describe two variations of the B^3 metric, B^3_{all} and B^3_0. In the former, all predicted twinless mentions are retained whereas the latter discards them. Rahman and Ng (2009) present one other variation of the metric where predicted twinless mentions

which are singletons are discarded before computing the metric. Finally, Cai and Strube (2010) propose a method where key and predicted mentions are manipulated depending on whether we are computing precision or recall.

Pradhan et al. (2014) argue that the B^3 metric was intended to work without any modification in the computation and propose an implementation of the metric. This is the one we are describing in the equations below. The computation of the metric with our running example is as follows:

$$\text{Recall} = \frac{\sum_{i=1}^{M_k}\sum_{j=1}^{R_r}\frac{|K_i \cap R_j|^2}{|K_i|}}{\sum_{i=1}^{N_k}|K_i|} = \frac{1}{7}\times(\frac{2^2}{3}+\frac{1^2}{3}+\frac{1^2}{4}+\frac{2^2}{4}) = \frac{1}{7}\times\frac{35}{12} \approx 0.42 \tag{5.4}$$

$$\text{Precision} = \frac{\sum_{i=1}^{M_k}\sum_{j=1}^{R_r}\frac{|K_i \cap R_j|^2}{|R_j|}}{\sum_{i=1}^{N_k}|R_j|} = \frac{1}{8}\times(\frac{2^2}{2}+\frac{1^2}{2}+\frac{1^2}{2}+\frac{2^2}{4}) = \frac{1}{8}\times\frac{4}{1} = 0.50 \tag{5.5}$$

$$\text{F1-Score} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} = \frac{2\times 0.42\times 0.50}{0.42+0.50} = 0.46 \tag{5.6}$$

The B^3 metric suffer from several shortcomings. For instance, the same gold cluster is allowed to be aligned to multiple different system clusters and vice versa (Luo 2005). Therefore, the metric does not correctly penalize systems that output an incorrect number of clusters.

Recasens and Hovy (2011) highlight that the metric is highly sensitive to the number of singleton clusters. The metric tends to be less sensitive to how well the systems extract coreference links when in presence of a high number of singletons.

5.10.3 Constrained Entity-Aligned F-Measure

Constrained Entity Aligned F-Measure (CEAF) was originally proposed by Luo (2005) to address the shortcomings of the B^3 metric. While MUC was a link-based metric, B^3 was mention-based, CEAF is devised as an entity based metric. The main idea of the metric is that an entity should only be used once in the evaluation. MUC and B^3 rely on intersections of entities and therefore entities can be used more than once during the computation.

The main idea behind CEAF metrics is to find the optimal alignment between gold and predicted clusters. Each gold cluster is aligned to at most one predicted cluster and vice versa. Luo (2005) describes two similarity scores in their paper: a mention-based ($CEAF_m$) and an entity-based version ($CEAF_e$).

Similarly to the B^3 metric, CEAF was considered to be underspecified for twinless mentions. This led to the creation of several variations of the metric in the literature that are similar to the variations implemented for the B^3 metric (cf. methods of Rahman and Ng (2009) and Cai and Strube (2010) in previous section). As for the B^3 metric, Pradhan et al. (2014) argue that CEAF was intended to work in the case where predicted mentions are different from key mentions and propose an implementation of the metric.

$CEAF_m$ recall is the number of aligned mentions divided by the number of key mentions while precision is the number of aligned mentions divided by the number of response mentions (Pradhan et al. 2014). The computation with our running example is as follows:

$$\text{Recall} = \frac{|K_1 \cap R_1| + |K_2 \cap R_3|}{|K_1| + |K_2|} = \frac{(2+2)}{(3+4)} \approx 0.57 \tag{5.7}$$

$$\text{Precision} = \frac{|K_1 \cap R_1| + |K_2 \cap R_3|}{|R_1| + |R_2| + |R_3|} = \frac{(2+2)}{(2+2+4)} = 0.50 \tag{5.8}$$

$$\text{F1-Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.57 \times 0.50}{0.67 + 0.50} = 0.53 \tag{5.9}$$

CEAF_e recall and precision are computed:

$$\text{Recall} = \frac{\phi_4(K_1, R_1) + \phi_4(K_2, R_3)}{N_k} = \frac{\frac{(2\times2)}{(3+2)} + \frac{(2\times2)}{(4+4)}}{2} = 0.65 \tag{5.10}$$

$$\text{Precision} = \frac{\phi_4(K_1, R_1) + \phi_4(K_2, R_3)}{N_r} = \frac{\frac{(2\times2)}{(3+2)} + \frac{(2\times2)}{(4+4)}}{3} \approx 0.43 \tag{5.11}$$

$$\text{F1-Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.65 \times 0.43}{0.65 + 0.43} = 0.52 \tag{5.12}$$

where $\phi_4(K_i, R_j)$ denotes the similarity between a key entity K_i and a response entity R_j and is computed as follows (Luo 2005):

$$\phi_4(K_i, R_j) = \frac{2 \times |K_i \cap R_j|}{|K_i| + |R_j|} \tag{5.13}$$

Although addressing the shortcomings of the previous metrics, Recasens and Hovy (2011) point out that the CEAF metric is as sensitive to the number of singletons as the B^3 score.

5.10.4 BLANC

The main objective of the BLANC score (Recasens and Hovy 2011) is to provide a metric that is more suitable for situations where singletons are annotated (e.g. in the ACE corpora). It also addresses the shortcoming of the MUC metric which does not take into account the impact of errors.

Two F-scores are computed, one which captures how accurately a system predicts coreference links (F_coref) and another to capture how accurately a system classifies mentions as singletons ($\text{F}_\text{non-coref}$). The final BLANC score is the arithmetic mean between the two F-scores.

The original computation of the metric only addressed the situation where gold mentions were provided. Pradhan et al. (2014) adapted the metric to the end-to-end situation where gold and predicted mentions could mismatch.

Let C_k be the set of coreference links in the key and C_r be the set of coreference links in the response. Let N_k be the set of non-coreference links in the key and N_r be the set of

non-coreference links in the response. A link between a mention pair m and n is denoted as mn. The set states for our example are as follows:

$$C_k = \{ab, ac, bc, de, df, dg, ef, eg, fg\}$$
$$N_k = \{ad, ae, af, ag, bd, be, bf, bg, cd, ce, cf, cg\}$$
$$C_r = \{ab, cd, fg, fh, fi, gh, gi, hi\}$$
$$N_r = \{ac, ad, af, ag, ah, ai, bc, bd, bf, bg, bh, bi, cf, cg, ch, ci, df, dg, dh, di\}$$

F_{coref} is computed as follows:

$$\text{R}_c = \frac{|C_k \cap C_r|}{|C_k|} = \frac{2}{9} \approx 0.22 \tag{5.14}$$

$$\text{P}_c = \frac{|C_k \cap C_r|}{|C_r|} = \frac{2}{8} = 0.25 \tag{5.15}$$

$$\text{F}_c = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.22 \times 0.25}{0.22 + 0.25} = 0.23 \tag{5.16}$$

$\text{F}_{\text{non-coref}}$ is computed as follows:

$$\text{R}_n = \frac{|N_k \cap N_r|}{|N_k|} = \frac{8}{12} \approx 0.67 \tag{5.17}$$

$$\text{P}_n = \frac{|N_k \cap N_r|}{|N_r|} = \frac{8}{20} = 0.40 \tag{5.18}$$

$$\text{F}_n = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.67 \times 0.40}{0.67 + 0.40} = 0.50 \tag{5.19}$$

The final BLANC score is computed as follows:

$$\text{BLANC} = \frac{F_c + F_n}{2} = \frac{0.23 + 0.50}{2} = 0.36 \tag{5.20}$$

5.10.5 The CoNLL Score

Each metric described above measures how well a system performs for coreference resolution with respect to a certain aspect. The consequence of this diversity is that it is difficult to compare system performances.

The CoNLL score (Pradhan et al. 2012; Pradhan et al. 2011) has been devised as a way of combining the strength of these metrics. It is computed as the unweighted arithmetic mean of MUC, B^3 and CEAF_e scores. While each metric provides a view of systems strengths and weaknesses, the CoNLL score allows to compare objectively system outputs according to three dimensions.

An implementation of all the metrics described in the section is available in the CoNLL scorer[1] (Pradhan et al. 2014).

5.11 Conclusion

Coreference resolution is a complex NLP topic that brings together several research domains. Carefully annotated corpora are needed to foster new approaches for coreference resolution. As we showed in this chapter, the linguistic phenomenon called coreference has been difficult to grasp in the NLP community and led to several annotation mistakes that have since been corrected.

Beyond the annotation process, devising approaches for coreference resolution is a difficult task. The literature has evolved from local pairwise classifiers to entity-based models and succeeded in improving the state-of-the-art on the topic. However, research on coreference resolution in the clinical domain remains sparse.

In the next chapter, we present an neural entity-based approach for coreference resolution in the clinical domain. It is based on an entity-aware mention-ranking model inspired by recent research efforts on the topic (Wiseman et al. 2016).

1. https://github.com/conll/reference-coreference-scorers

Chapter 6

Neural Entity-Based Approach for Coreference Resolution in the Clinical Domain

6.1 Introduction

This chapter addresses coreference resolution in the clinical domain. Recent neural approaches have allowed for better performance in the general domain. Specifically, entity-based approaches seem to be a promising area of research (K. Clark and Manning 2016a; Wiseman et al. 2016) although simple neural-based mention-ranking models present state-of-the-art results (K. Lee et al. 2017). To the best of our knowledge, these types of models have not been transposed to the clinical domain. Hence, we present a neural entity-based approach for coreference resolution in the clinical domain. Specifically, we experiment with

various model configurations that include character-level representations, attention mechanisms and network pretraining and measure their impact on the final coreference score.

As we mentioned in the introduction, temporal information extraction and coreference resolution are interlinked. Temporal clues can be valuable for a coreference resolution system. In this context, we devise a temporal feature based on the relationship that exists between medical events and DCTs and embed that feature into our model.

The remainder of the chapter is organized as follows. We present the corpus that has been used in our experiments in Section 6.2. We discuss how we split the task into two subtasks in Section 6.3. We present our approach for mention extraction in Section 6.4 and for building a temporal feature in Section 6.5. We describe our model components in Section 6.6 and our experimental setup in Section 6.7. Results of our experiments are presented in Section 6.8. Finally, we discuss these results in Section 6.9 and we conclude the chapter in Section 6.10.

6.2 Data

As we saw in Chapter 5, there are only two clinical corpora annotated with coreference chains available to the community. We chose to use the i2b2 corpus task1c in our experiments (Uzuner et al. 2012). Specifically, we selected the i2b2/VA part that does not contain documents from the University of Pittsburgh Medical Center (UPMC). This situation mimics one of the experimental setup that was proposed to the participants during the shared task. The dataset contains 194 clinical documents from the Beth Israel Deaconess Medical Center (BETH) and 230 clinical documents from Partners Healthcare (PARTNERS) (Table 6.1). The number of coreference chains, and their average and maximum lengths (i.e. the number of mentions per chain) are presented in Table 6.2.

Institution	Train	Test	Total
BETH	115	79	194
PARTNERS	136	94	230
Combined	251	173	424

Table 6.1: i2b2/VA task1c corpus train and test file counts.

Institution	Chain count	Chain avg. len.	Chain max. len.
BETH	1,816	4.155	122
PARTNERS	1,395	4.352	105

Table 6.2: i2b2/VA task1c corpus chain count, chain average length (avg. len.) and chain maximum length (max. len.).

The corpus is annotated with five *concept* (i.e. *entity*) types: *person*, *pronoun*, *test*, *treatment* and *problem*. Singletons are annotated (i.e. entities that do not take part in any coreference chain). Uzuner et al. (2012) do not mention explicitly that the medical entities (tests,

treatments and problems) are event mentions. However, as we saw in Chapter 5, these entity mentions can be considered as event mentions when found in a clinical document. Pronouns participate in either people-related chains or event-related chains. The relation annotated in the corpus is the one of identity (cf. Chapter 5, Section 5.2). Other anaphoric relation types (e.g. set–subset) are not annotated.

Corpus Format Conversion

The documents are shipped as i2b2-formatted files. Each clinical text has three associated documents: a *text* file that contains the raw text, a *concept* file that holds entity offsets and types, and a *chain* file that tracks coreference chains. The text file holds the text content of the clinical document. It has been segmented in sentences and tokenized. There is one sentence per line and tokens are separated by spaces.

The concept file encodes entity positions following a custom bi-offset format. Each offset boundary (begin or end) is a 2-dimension coordinate where the first dimension points to the line, and the second points to the token (Figure 6.1). This offset information is completed with the surface form of the concept and its type. There is one entity per line.

```
...
c="thrombocytopenic" 43:3 43:3||t="problem"
c="back surgery" 25:1 25:2||t="treatment"
c="platelets" 42:13 42:13||t="treatment"
...
```

Figure 6.1: Concept annotations extracted from the concept file associated to the document clinical-13.

The chain file contains coreference chains. A chain is composed of several entities which are encoded similarly to the format described above except that their types are omitted. The chain type is encoded via a custom chain attribute (Figure 6.2). There is one chain per line.

```
...
c="hepatitis c cirrhosis" 12:16 12:18||c="hepatitis c cirrhosis" 20:1
20:3||c="which" 20:5 20:5||t="coref problem"
c="transplant" 48:8 48:8||
c="an orthotopic liver transplant" 49:10 49:13||
c="the procedure" 50:22 50:23||c="the procedure" 53:0 53:1||
c="the procedure" 53:20 53:21||c="surgery" 59:13 59:13||
c="the liver transplant" 59:32 59:34||c="liver transplant" 88:0 88:1||
t="coref treatment"
...
```

Figure 6.2: Chain annotations extracted from the chain file associated to the document clinical-13.

The corpus was released during the i2b2 evaluation campaign on coreference (Uzuner et al. 2012). Participants were asked to format their system outputs to match the i2b2 format. Evaluation was performed with a custom Python script coded for the shared task[1].

Preliminary experiments performed with the corpus and the evaluation script revealed several difficulties and problems. First, the i2b2 annotation format is cumbersome to use. The vast majority of text processing tools are using character-based offsets to encode annotations within documents. This has become the usual way of devising NLP pipelines. Furthermore, by enforcing a line/token offset system, the i2b2 format can lead to annotation mismatch as the way of segmenting lines may vary (e.g. the way of splitting a double space).

Second, the evaluation script implements custom versions of the B^3 and CEAF metrics. The script has been devised before Pradhan's implementation of the coreference metrics that was used during the CoNLL challenges (Pradhan et al. 2014). Metric computation is done by modifying key and response partitions. Preliminary comparisons between the CoNLL scorer and the i2b2 evaluation script outputs revealed discrepancies between scores. Furthermore, the script suffers from an important computation time (several minutes for the CEAF metric).

These observations motivated us to implement the transformation of the corpus to the CoNLL format. This conversion has two main advantages. It will allow the community to work on coreference in the clinical domain more easily by using a well-known and well-tested format. Furthermore, it allows to use the CoNLL scorer for evaluation which has become the reference evaluation script for researchers working on coreference resolution.

One drawback of using the CoNLL format is that the entity type information is lost. This information is important when dealing with coreference in the clinical domain as two mentions that are not of the same type are not likely to be coreferent. Thus, in order to keep this valuable information, we also convert the corpus to the brat format. Mapping between brat and CoNLL versions of the corpus is performed via offset tracking in the CoNLL version. The entire transformation process is documented and the scripts will be available online.

Conversion quality was tested by converting the resulting CoNLL files back to the i2b2 format. We asserted the quality by using the official evaluation script on the converted files and obtained an optimal score. Brat version statistics are presented in Table 6.3. We report the number of entities for each type (person, problem, pronoun, test and treatment) as well as the number of pairwise coreference links (coref_person, coref_pronoun, coref_test and coref_treatment). Chain statistics are discussed in Section 6.3. Overall, corpus size is much smaller than the OntoNotes corpus in the general domain which includes 2,083 documents annotated with 131,846 mentions and 97,556 coreference links.

An example of a sentence transformed to both formats (brat and CoNLL) is presented in Appendix C in Figures C.1 and C.2. The description of the CoNLL file columns is presented in Table C.1.

6.3 Task Division

The annotated mentions can be divided in two groups. The first regroups event mentions (tests, treatments and problems). The second one comprises people mentions (person). Both groups differ in several ways. First, event and people entities differ from a semantic view-

1. https://github.com/jtourille/i2b2-coreference-evaluation

	Train	Test	Total
entities			
person	11,026	7,242	18,268
problem	11,924	7,596	19,520
pronoun	2,187	1,274	3,461
test	8,071	5,672	13,743
treatment	8,328	5,722	14,050
relations			
coref_person	9,226	5,938	15,164
coref_problem	3,292	2,236	5,528
coref_test	748	526	1,274
coref_treatment	2,067	1,706	3,773

Table 6.3: i2b2 task1c corpus: brat version statistics.

point. As we saw in Chapter 5, despite the fact that clinical events are often realized by noun phrases, they hold an implicit argument structure while person-related entities do not.

Second, coreference chain statistics presented in Table 6.4 show that coreference chains related to people tend to be longer (more than ten mentions per chain) and the maximum length is much larger than the one of other chain types.

Finally, the vast majority of mentions in coreference chains related to people takes the form of personal pronouns. The diversity of surface forms is therefore limited compared to the one of clinical events.

Considering these three major differences, it seems legitimate to consider both coreference resolution tasks as different subtasks. It seems that the features used for supervised learning in both cases will be different enough to train two separate models.

Pronouns can be found in either the chains related to clinical events or those related to people. A distribution of pronouns according to the chain type is presented in Table 6.5. Pronouns are mostly singletons (60.76%)[2]. The remaining are distributed across all chain types. We note that coreference chains related to people contain less pronouns than the others (3.79%).

Following these observations, we decided to include all pronouns in both models (people and clinical events). Our hypothesis is that the models will be able to learn which pronoun has to be included in the chains.

To summarize, our approach consists in dividing the coreference task into two subtasks, one related to people mentions, the other to clinical event mentions. In both subtasks, all pronoun entities are included.

Gold vs. End-to-End Coreference Resolution. Up to this point, we considered the situation where gold mentions are given and the task is to extract coreference links. This setting is often considered as a simplified version of the task. To measure the effectiveness of our

2. Although pronouns always refers to something in the text, this something may not fall in the annotated entity categories.

		TRAIN			TEST			ALL		
		Chain count	Chain avg. len.	Chain max. len.	Chain count	Chain avg. len.	Chain max. len.	Chain count	Chain avg. len.	Chain max. len.
BETH	person	379	14.20	149	268	12.51	122	647	13.48	149
	problem	1,019	2.97	17	703	2.86	12	1,722	2.92	17
	test	322	2.31	10	219	2.50	9	541	2.39	10
	treatment	779	2.67	20	626	2.61	14	1,402	2.64	20
PARTNERS	person	375	12.29	123	304	10.39	105	679	11.44	123
	problem	685	2.88	11	500	2.85	14	1,185	2.87	14
	test	246	2.32	6	143	2.38	6	389	2.34	6
	treatment	486	2.60	17	448	2.56	9	934	2.58	17
ALL	person	754	13.24	149	572	11.38	122	1326	12.44	149
	problem	1,704	2,93	17	1,203	2.86	14	2,907	2.90	17
	test	568	2.32	10	362	2.45	9	930	2.37	10
	treatment	1,262	2.64	20	1,074	2.59	14	2,336	2.62	20

Table 6.4: Coreference chain statistics. For each corpus part (BETH and PARTNERS) we report the number of coreference chains, and their average and maximum lengths. We also report aggregated numbers.

	TRAIN	TEST	ALL
person	81	50	131 (3.79%)
problem	357	201	558 (16.12%)
test	247	149	396 (11.44%)
treatment	188	85	273 (7.89%)
singletons	1,314	789	2,103 (60.76%)

Table 6.5: Pronoun distribution per chain type.

approach on system-predicted mentions, we devise a NER module for mention extraction and apply our coreference resolution models on these mentions.

6.4 Mention Extraction

Before presenting how we performed mention extraction, we would like to clarify some terminology that we will use in the rest of this chapter. Coreference resolution objective is to cluster event or entity mentions that are referring to the same real-world event or entity. However, to smooth the description of our approach, we will refer to event mentions as *events* and people mentions as *people*. We will use the word real-word (event or person) to differentiate between the two concepts.

Mention extraction is performed via a hybrid approach. Events (tests, pronouns and problems) are extracted via a fully supervised machine learning approach using YASET (Tourille et al. 2018). One model is learned for the three event types. A detailed description of the neural model is presented in Chapter 4, Section 4.3.1.

People are extracted via a combination of regular expressions and supervised machine learning. We noticed that several mentions overlap with each other, (e.g. the mentions *his* and *his primary physician*). In all overlapping situations, we removed the mentions with the smallest span length and devised a regular expression for each of them. The final list contains six mention forms: *his*, *her*, *your*, *the patient*, *the pt* and *our*. For the main group of mentions, we learn a NER model with YASET.

Pronouns are extracted via the use of regular expressions. Pronoun list includes: *it*, *its*, *that*, *their*, *them*, *these*, *they*, *this*, *those*, *which*. Originally, the pronoun "there" was included in the list but it seems that its annotation within the corpus is not coherent. There is a large number of them which are not annotated and among those which are annotated, only one participates to a coreference chain. Therefore, we decided to remove the pronoun from the list. Extracted pronouns are added to the list of mentions in both event and people related coreference models.

Hyperparameters. For both models, we used word embeddings computed[3] with the gensim implementation of word2vec on the MIMIC III corpus. Mini-batch size is set to 8. Main Bi-LSTM hidden layer size is set to 100. We initialized character embeddings randomly and set their size to 25. The character-level Bi-LSTM hidden layer has a size of 25. We set the

3. algorithm=skip-gram, vector size=100, window=8

Out Of Vocabulary (OOV) token replacement rate to 0.5 for the NER model related to events and 0.25 for the one related to people.

NER Model Performance. Model performance for event and people extraction are presented in Tables 6.6 and 6.7. The NER model for events gives similar performance across all event types. The f1-score is ranging from 86.99 for treatments to 88.82 for tests. The performance gap across event types is largely imputable to their respective recall scores which range from 85.55 for treatments to 89.03 for tests. The precision is stable for all categories (≈88.50).

Category	P	R	F1
problem	88.55	87.18	87.86
test	88.61	89.03	88.82
treatment	88.49	85.55	86.99
OVERALL	88.55	87.27	87.90

Table 6.6: NER model performance for event extraction on the development corpus part (20% of the training corpus). We report precision (P), recall (R) and f1-score (F1) for each category and micro-average on all categories.

Category	P	R	F1
person	93.69	91.16	92.41

Table 6.7: NER model performance for people extraction on the development corpus part (20% of the training corpus). We report precision (P), recall (R) and f1-score (F1).

The NER model for people extraction obtains a high performance with a f1-score of 92.41. Model performance is evenly distributed between precision (93.69) and recall (91.16).

We report the final performance on the test part of the corpus in Table 6.8. The scores are computed with *brateval* and after that all steps described above have been applied (supervised learning and application of regular expressions for people and pronouns). As expected, pronoun extraction obtains a very high f1-score (0.9647) with very little false negatives (16). Scores for events are similar to the ones obtained on the development corpus (from 0.8443 for problems to 0.8771 for tests). The f1-score for people is very high (0.9403). Overall, our mention extraction method reaches a balanced f1-score of 0.8803.

Coreference Resolution Evaluation. To measure the impact of mention extraction errors on the coreference score, we project the gold coreference chains on the predicted mentions and perform coreference resolution evaluation with the official CoNLL scorer (Table 6.9). The ceiling CoNLL f1-score that can be obtained with the predicted mentions is 91.84. Most of the performance decrease is imputable to a drop in recall for all three metrics that compose the CoNLL score.

	TP	FP	FN	P	R	F1
person	6,632	233	609	0.9661	0.9159	0.9403
problem	6,327	1,064	1,269	0.8560	0.8329	0.8443
pronoun	2,516	168	16	0.9374	0.9937	0.9647
test	4,980	703	692	0.8763	0.8780	0.8771
treatment	4,812	720	910	0.8698	0.8410	0.8552
OVERALL	24,023	3,051	3,482	0.8873	0.8734	0.8803

Table 6.8: Mention extraction performance on the test part of the corpus. Metrics are computed using brateval. We report true positives (TP), false positives (FP), false negatives (FN), precision (P), recall (R) and f1-score (F1).

Metric	P	R	F1
B^3	100.00	85.24	92.03
MUC	100.00	89.14	94.25
$CEAFE_e$	97.59	82.21	89.24
CONLL	99.20	85.53	91.84

Table 6.9: Coreference resolution scores obtained with the predicted mentions on the test corpus. Gold coreference chains are projected on the predicted mentions and the score is computed with the official CoNLL scorer. We report precision (P), recall (R) and f1-score (F1) for all coreference metrics.

6.5 Building a Temporal Feature

Event temporal location seems to play a major role in event coreference resolution (cf. Chapter 5, Section 5.4). Being able to take into account temporal clues in a coreference resolution system could improve its performance. However, since the i2b2 corpus does not include temporal information, it has therefore to come from another source.

Usually, temporally annotated datasets contain two types of temporal relations: relations between event and/or temporal expressions and relations between event and the DCT (cf. Chapter 2). Intuitively, the DCT relation could bring a valuable information to a coreference resolution system. Two events which have similar or non-contradictory DCT relations (e.g. *before* and *before-overlap*) have a higher chance to be coreferent than two events happening respectively *before* and *after* the DCT.

Among available corpora, the THYME corpus seems to be the more suitable to learn a DCT relation extraction model. However, due to the fact that events from the i2b2 and THYME corpora have different definitions, it would be dubious to try to assign a DCT relation to i2b2 events with a model learned on the THYME corpus. Event definition in the THYME corpus is larger and regroups clinical concepts beyond the three event categories annotated in the i2b2 corpus.

The solution we adopted consists in learning a NER model for event extraction on the THYME corpus by incorporating the DCT relation into the IOB tagging scheme. The model extract events with their associated DCT relations. Then, this model is used to extract THYME events in the i2b2 corpus. This approach has several drawbacks. The model learned on the THYME corpus will extract much more events than there are in the i2b2 corpus as the definition of an event in the THYME corpus is more loose than the one of the i2b2 corpus.

Furthermore, it is not sure that every i2b2 events will have a corresponding THYME event extracted within its span. Following this observation, we decided to build a temporal feature based on the sentence context of the events. Each i2b2 event is assigned a DCT relation based on the DCT relations of THYME events extracted in the sentence. For instance, if an i2b2 event is located in a sentence comprising two THYME-predicted events which have two DCT relations *before* and *before-overlap*, the final DCT relation associated with the i2b2 event will be *before/before-overlap*. If there is no THYME event within the sentence, the DCT relation is declared as *undefined*.

There is one condition for this feature to be useful: sentences must be temporally coherent, meaning that they must not contain contradictory DCT relations. From our perspective, temporally coherent sentences relate clinical events that are temporally located in a semi interval started or closed by the DCT. For instance, a sentence containing *before*, *before/overlap* and *overlap* DCT relations would be considered as coherent. The same goes for sentences which combine *overlap* and *after* DCT relations. However, this would not be the case for sentences having both *before* and *after* DCT relations.

We verify sentence temporal coherence in the training data (i.e. the THYME corpus). Figure 6.3 shows the number of sentences according to the number of different DCT relations they contain. The vast majority of sentences which contain events have only one type of DCT relation (78.86%). For those sentences, there is therefore no contradictory temporal information.

Approximately 18% of sentences contain at least two events with different DCT relations. Figure 6.4 presents the number of sentences per pair of DCT relations. Among those sen-

tences, only a small proportion lacks temporal coherence. Around 5% of the sentences contain two events with *before* and *after* DCT relations. Less than 1% of sentences have two events with *after* and *before-overlap* relations.

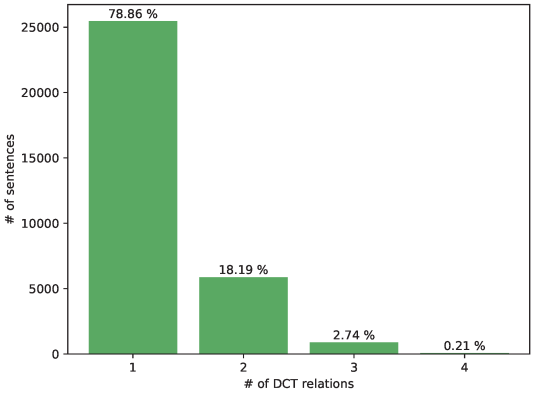

Figure 6.3: Number of sentences per number of DCT relations in the THYME corpus.

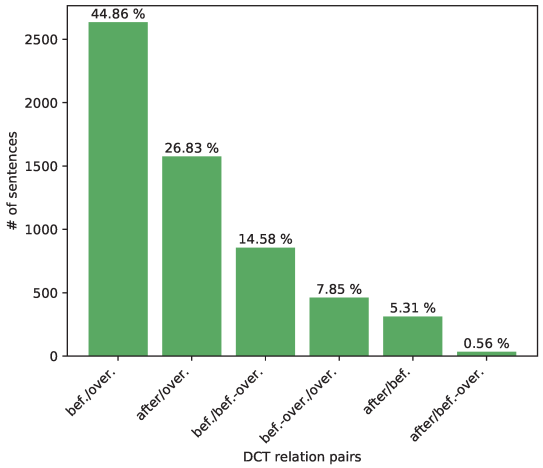

Figure 6.4: Number of sentences per pair of DCT relations in the THYME corpus.

Sentences that contains more than two different DCT relations represent less than 3% of the corpus and therefore will marginally impact the learned model. The coherence of the training corpus seems sufficient enough to learn a NER model. This model will be used to extract events in the i2b2 corpus. These events will in turn be used to assign a DCT relation to mentions.

NER Model. We train a NER model for event extraction on the THYME corpus using YASET. We used word embeddings computed[4] with the gensim implementation of word2vec on the MIMIC III corpus. Mini-batch size is set to 8. Main Bi-LSTM hidden layer size is set to 256. We initialized character embeddings randomly and set their size to 25. The character-level Bi-LSTM hidden layer has a size of 25. We set the OOV token replacement rate to 0.5.

Model performance is presented in Table 6.10. Performance is not well distributed across DCT relations. The relations *after*, *before* and *overlap* present similar results (f1-scores between 76.71 and 80.16). The relation *before/overlap* has a much lower performance with a f1-score of 60.13. The overall score is similar to the best result on the task during the 2016 edition of Clinical TempEval. H.-J. Lee et al. (2016) obtained a global f1-score of 0.756 with a precision of 0.766 and a recall of 0.746.

The NER model is then used to extract THYME clinical events in the i2b2 corpus. We perform the same analysis that we did on the THYME corpus. Figure 6.5 shows that the number of DCT relations per sentence follows the same distribution as the one observed for the THYME corpus with a large majority of sentences having only one type of DCT relation. Among the sentences which contain two DCT relations, a small fraction of them are not

4. algorithm=skip-gram, vector size=200, window=2

Category	P	R	F1
AFTER	75.00	78.49	76.71
BEFORE	81.24	79.11	80.16
BEFORE/OVERLAP	64.67	56.18	60.13
OVERLAP	77.88	78.44	78.16
OVERALL	78.12	77.44	77.78

Table 6.10: NER model performance for THYME clinical events on the development corpus part (20% of the training corpus). We report precision (P), recall (R) and f1-score (F1) for each category and micro-average on all categories.

coherent (8.17% have a least one *after* and *before* relations, 0.27% of them have a least one *after* and *before/overlap* relations).

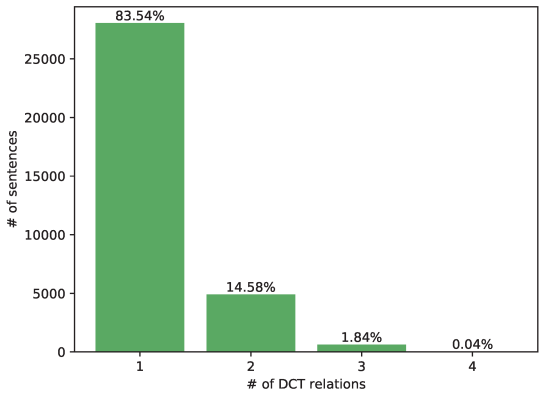

Figure 6.5: Number of sentences per number of DCT relations in the i2b2 corpus.

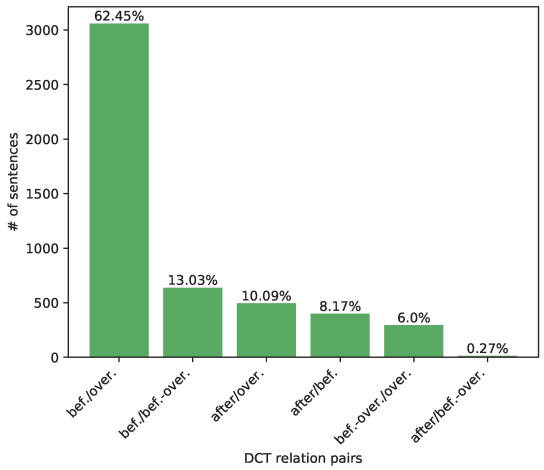

Figure 6.6: Number of sentences per pair of DCT relations in the i2b2 corpus.

We also analyze the resulting temporal feature. Figure 6.7 shows the number of clinical event per computed DCT relation type. There are 13 relation types. The majority of i2b2 clinical events are assigned the relation *before* (50.3%). The relations *overlap* and *before + overlap* are assigned to 14.47% and 13.27% of the clinical events respectively. The rest of the DCT relation types are below 5%.

Concerning the coreference chains, 43.2% of them include clinical events that have the same computed DCT relation and 43.49% include at most two different relations (Figure 6.8). Among the latter, 83.48% of the chains are coherent (i.e. the DCT relations are not incompatible). This suggests that most coreference chains are temporally coherent and that our devised temporal feature could bring useful information to the coreference resolution system.

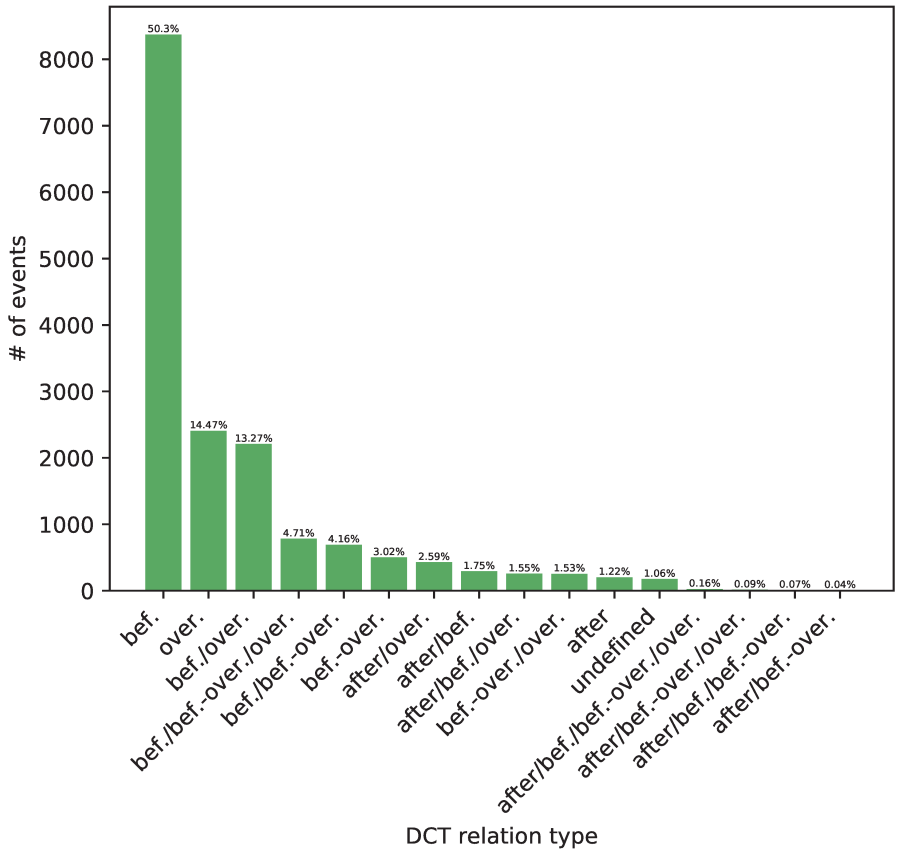

Figure 6.7: Number of events according to the DCT relation type (corpus i2b2).

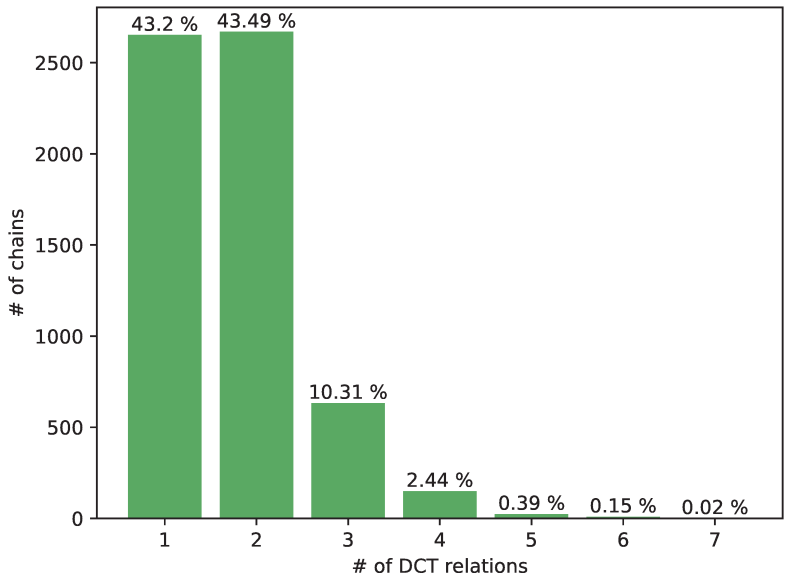

Figure 6.8: Number of chains according to the number of DCT relations they contain (corpus i2b2).

6.6 Neural Entity-Based Approach for Coreference Resolution

We devised an entity-aware model inspired by recent neural approaches in the general domain (K. Lee et al. 2017; Wiseman et al. 2016). The main component of our approach is a mention-ranking model (cf. Chapter 5, Section 5.8.2) which predicts the correct antecedent for any given active mention. Non-anaphoric cases are handled via the use of a dummy antecedent (Durrett and Klein 2013; Wiseman et al. 2016).

While local coreference models produce state-of-the art results (K. Lee et al. 2017; Wiseman et al. 2015), devising global approaches seems to be a promising research direction. Incorporating entity-level information in the coreference model has been shown to be efficient in recent work (K. Clark and Manning 2015; K. Clark and Manning 2016b; Wiseman et al. 2016; Wiseman et al. 2015).

Our approach is similar to the one described in Wiseman et al. (2016). Our mention-ranking component is enhanced by including cluster-level information of partially constructed clusters. This cluster-level information is computed using a LSTM that processes cluster mentions by order of apparition in the text. The main advantage of this approach is that it simplifies inference by requiring only a left-to-right pass on the text's mentions.

In the remainder of this section, we present a detailed overview of the model. First, we describe how input embeddings are computed (Section 6.6.1). Second, we present how mention representations are built (Section 6.6.2). Third, we show how our model takes into account cluster-level information (Section 6.6.3). Finally, we describe the pairwise scorer used to score a antecedent–mention pair (Section 6.6.4) and how training is performed (Section 6.6.5).

6.6.1 Input Embeddings

Similarly to the method presented in Chapter 4 Section 4.3.4, input embeddings are built by concatenating a character-based embedding and a pre-computed word embedding.

An overview of the character-based embedding computation is presented in Figure 4.4. Following Lample et al. (2016), the character-based representation is constructed with a Bi-LSTM. First, a random embedding is generated for every character present in the training corpus. Token characters are then processed with a Bi-LSTM (forward and backward LSTM in Figure 6.9). The final character-based representation is the result of the concatenation of the forward and backward representations.

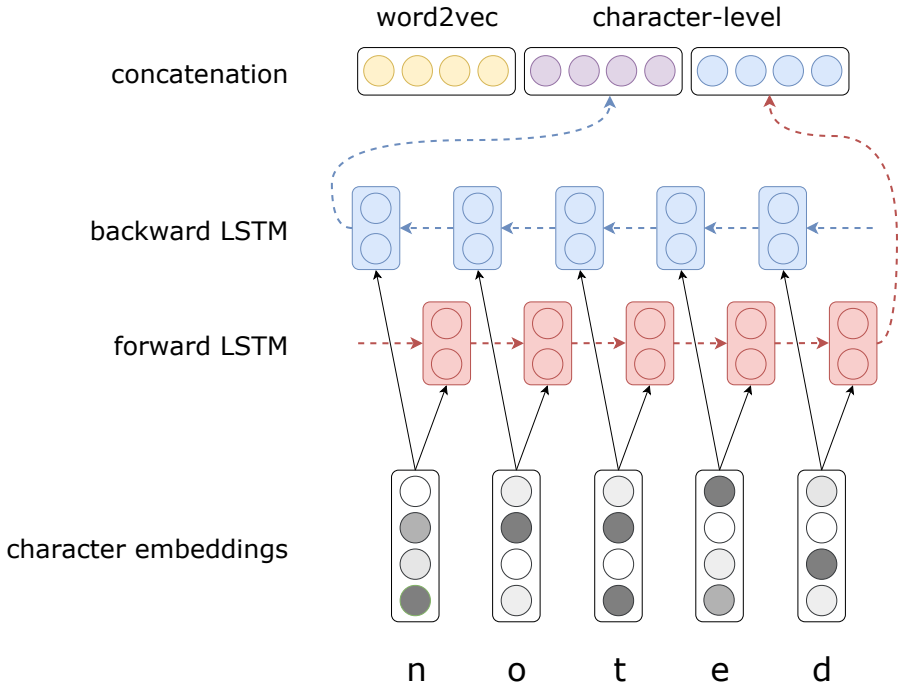

Figure 6.9: Example of token representation computation.

6.6.2 Mention Representation

An overview of mention representation computation is presented in Figure 6.10. For a given mention, embedded in its sentential context, we compute two contextual representations with two LSTMs. The forward LSTM processes the sentence from the first token of the sentence to the last token of the mention (forward LSTM in Figure 6.10). The backward LSTM processes the sentence from the last token of the sentence to the first token of the mention (backward LSTM in Figure 6.10). We implemented another approach during development where each contextual mention representation was composed of three parts: left and right contexts, and the mention itself. Each part was computed by processing the tokens with a bi-LSTM. However, this solution largely decreased the performance of the model.

Then, these two left and right contextual representations are concatenated. We add two feature embeddings, one to represent the temporal feature (DCT relation) and one to represent the mention type (person, pronoun, test, problem or treatment).

Finally, we add an attention mechanism. Attention on both left and right contexts is computed via a weighted sum of the LSTM hidden states. Weights are computed using a feed forward network with one hidden layer. Attention has proven to be useful for coreference resolution (K. Lee et al. 2017). It allows the network to focus on the most informative parts of the input and can further increase the performance of a given approach.

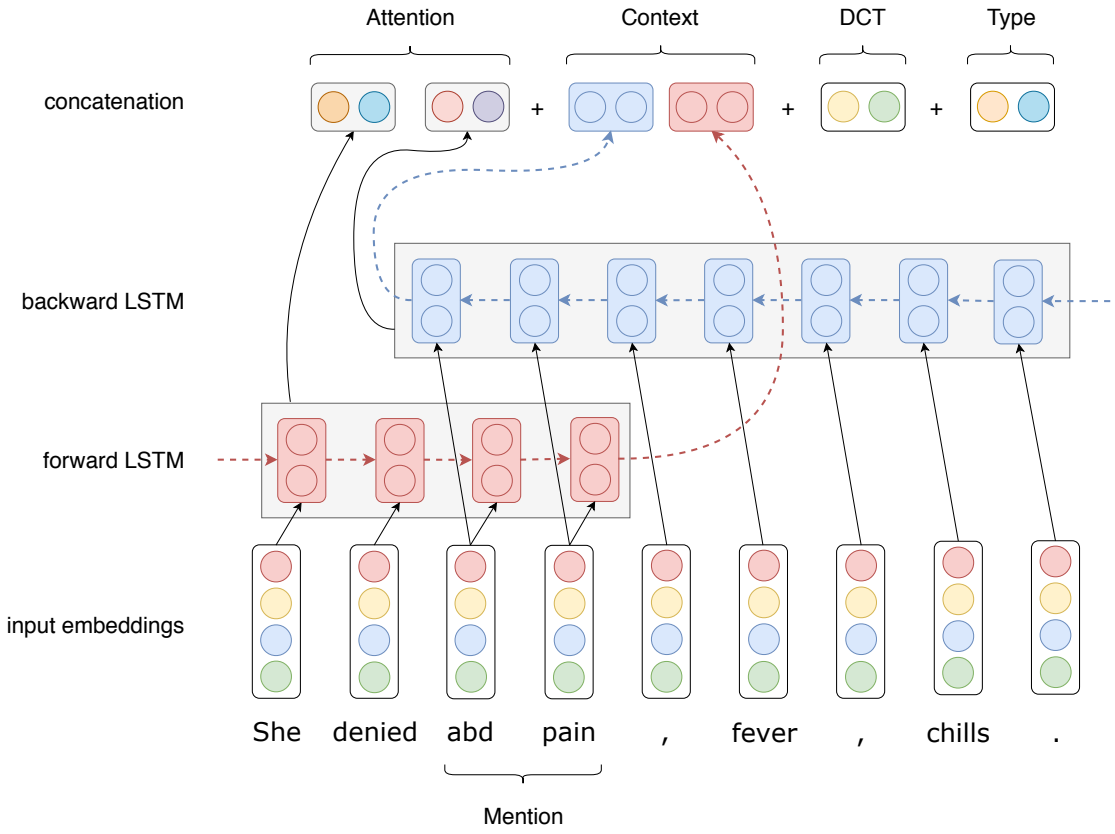

Figure 6.10: Example of mention representation computation. Mention sentential context is processed with a Bi-LSTM. The resulting dense representation is concatenated to one embedding for the DCT relation, one for the mention type and an attention representation of both left and right sentence contexts.

6.6.3 Cluster-Level Representation

Similarly to Wiseman et al. (2016), we used a LSTM to build cluster representations. To produce a cluster representation, we process its mentions with a LSTM by order of apparition in the text. We use mention representations that have been computed following the method described in the previous section. An overview of the cluster representation computation is presented in Figure 6.11. Our hypothesis is that this cluster-level representation will be

able to capture similarities within the cluster. These similarities may be found in the DCT relations, in the type of the mentions or the linguistic context of the mentions.

From an implementation viewpoint, as we process the text from left to right, we maintain n cluster representations. These representations are built with the same LSTM (i.e. the same set of parameters) and updated after each mention processing. Singleton clusters are composed of only one mention.

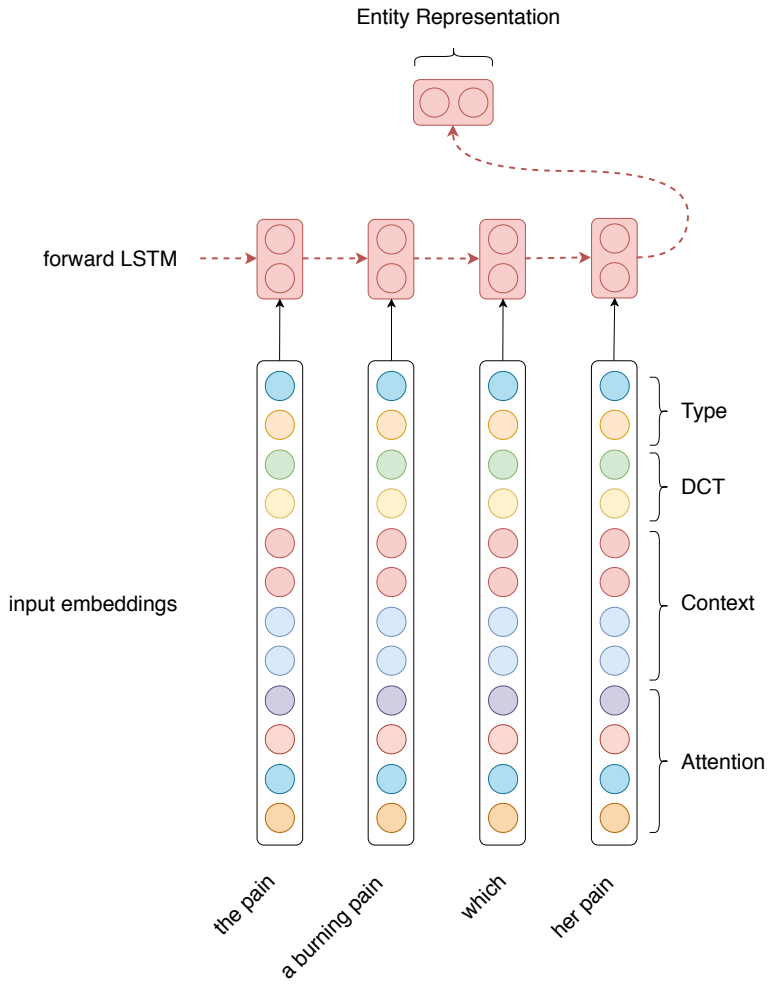

Figure 6.11: Example of cluster representation computation. The cluster is composed of four mentions: *the pain*, *a burning pain*, *which* and *her pain*.

6.6.4 Pairwise Scorer

An overview of the computation process is presented in Figure 6.12. The pairwise scorer assigns a score to every antecedent–mention pair. It takes as input the active mention and the antecedent representations. The latter is composed of the mention representation as described in Section 6.6.2 and the partial cluster representation in which the antecedent

belongs as described in Section 6.6.3. We add a distance feature embedding representing the number of sentences separating the active mention and its antecedent. Following K. Clark and Manning (2016), the distance is binned into the following buckets: [1, 2, 3, 4, 5–7, 8–15, 16–31, 32–63, 64+].

This input is then fed to a feed forward neural network with two hidden layers. Each hidden layer has a size equal to the half of the preceding layer size. The output of the network is the antecedent–mention pair score.

Pairwise Scorer Specialization. We implement pairwise scorer specialization according to the mention type. In this scenario, we build a different feed forward network for each mention type (person, pronoun, test, treatment, problem). During learning, depending on the active mention type, the corresponding feed forward network will be used to compute all pairwise scores. This approach can be seen as similar to multitask approaches which have been shown to be efficient in several domains, as for instance in sequence labeling (Crichton et al. 2017; Rei 2017).

6.6.5 Training

Individual antecedent–mention scores are then combined and a softmax layer is applied. Following K. Lee et al. (2017), we optimize the marginal log-likelihood of all correct antecedents implied by the gold clustering:

$$\log \prod_{i=1}^{N} \sum_{\hat{y} \in \mathcal{Y}(i)} P(\hat{y}) \tag{6.1}$$

where $\mathcal{Y}(i)$ is the set of possible assignments for each y_i ($\mathcal{Y}(i) = \{\epsilon, 1, ..., i-1\}$) with the dummy antecedent ϵ and all preceding mentions.

We fix the score of the dummy antecedent to zero. This strategy prevents antecedent filtering to introduce noise (i.e. the case where all gold antecedents have been pruned). In this situation, the learning objective will push the scores of non-antecedent mentions lower and not incorrectly push the score of the dummy antecedent higher (K. Lee et al. 2017). We experimented with the cost-sensitive slack-rescaled learning objective presented in Wiseman et al. (2016) and Durrett and Klein (2013) but the learning objective presented in K. Lee et al. (2017) proved to be more efficient.

6.6.6 Wrap-Up

This section aims to summarize how the network pieces work together. Let consider the example presented in Figure 6.13. The system is processing one document, composed of one sentence. Five mentions have already been processed in the document and the current active mention is the entity "your". The mention-ranking approach consider all mentions located before the active mention as potential true antecedents. We add the case where the active mention is consider as non anaphoric.

The system has already predicted that there are three partial entities up to this point in the document composed of the following mentions: (you, you), (further chest pain) and (other symptoms, which). The rest of the process is as follows:

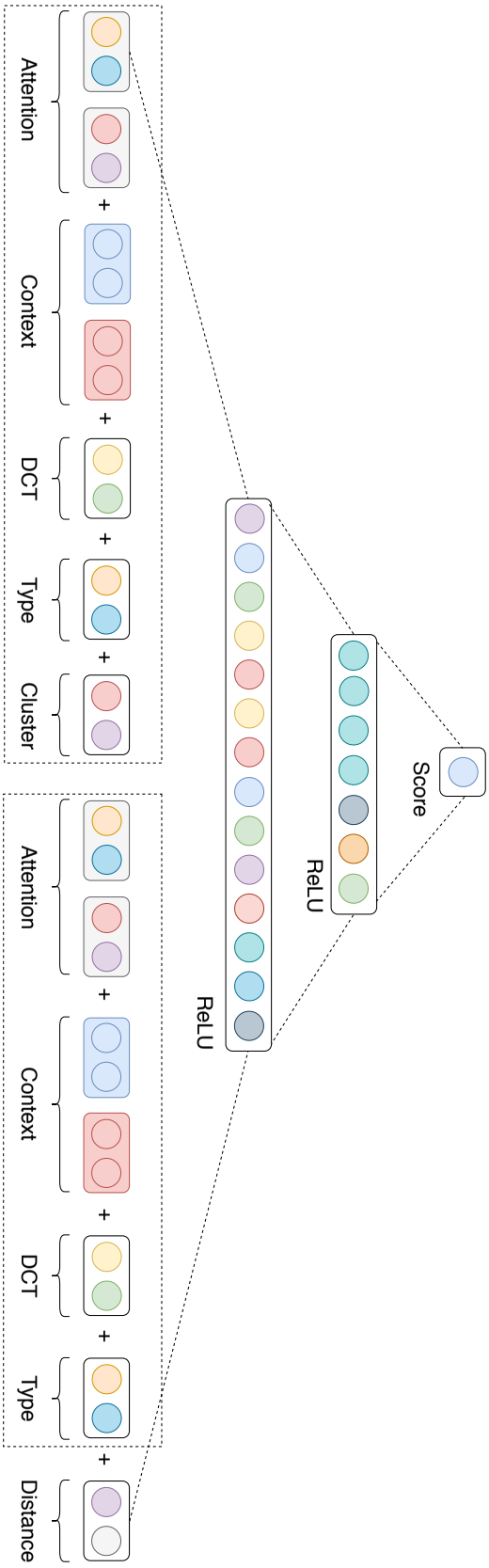

Figure 6.12: Pairwise scorer architecture overview.

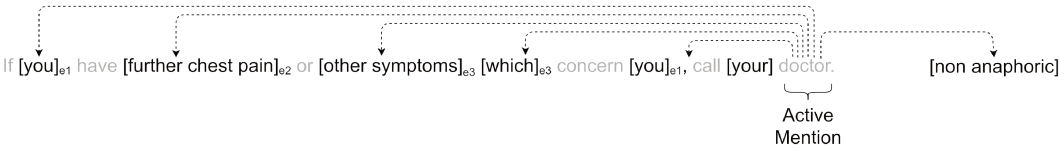

Figure 6.13: Example of a document being process by our system. Five mentions have already been processed in the document and the current active mention is the entity "your". There are three partial entities up to this point in the document composed of the following mentions: (you, you), (further chest pain) and (other symptoms, which).

1. We compute mention representations for the antecedents and the active mention by following the approach presented in Section 6.6.2.
2. We compute intermediate entity representations (also called intermediate clusters) with the approach presented in Section 6.6.3.
3. For each pair (antecedent, active mention), we compute a score following the approach presented in Section 6.6.3.
4. We regroup all the scores (+ the score 0 for the non-anaphoric case) and apply a softmax function to get a probability distribution. The antecedent that get the highest probability will be consider as the true antecedent.
5. We compute the loss according to the training objective presented in Section 6.6.5 where we maximize the probabilities of all true antecedents. In our example, they include the mentions *you* and *you*.

Since we are processing the last mention of the document, we back-propagate the accumulated loss and optimize the parameters.

6.7 Experimental Setup

6.7.1 Experiment Configurations

We perform several experiments using our approach. To measure the added value of the model components, we setup a baseline configuration for which we removed several parts of the architecture described in the previous section. Input embeddings are only composed of the pre-computed word embeddings. Mention representations are composed of the contextual representations without feature embeddings and the attention mechanism. We also remove the cluster-level representation.

Starting with this configuration, we activate each component: character-level mention representation (Section 6.6.1), mention type and DCT relation feature embeddings, attention mechanism (Section 6.6.2), cluster-level representation and pairwise scorer specialization (Section 6.6.4).

Following previous research efforts, we implement a pretraining strategy (K. Clark and Manning 2016a; Wiseman et al. 2016; Wiseman et al. 2015). We train our network on a

simplified version of the task by selecting active mentions that are known to be anaphoric. The task is then to assign the correct antecedent to the active mention.

Finally, we implement antecedent filtering. For people and events, depending on the active mention type, we consider antecedents that are of the same type or of the type *pronoun*. For pronouns, all antecedents are considered.

6.7.2 Hyperparameters

The model was built using Pytorch (Paszke et al. 2017). Contextual and cluster-level LSTM hidden layers sizes are fixed to 100. We fix the hidden layer size of the character Bi-LSTM and the one of the character embeddings to 25. Feature embedding size is set to 50.

Network is trained end-to-end using mini-batches of 1 document. We use Adam as optimization algorithm with an initial learning rate of 0.001. We apply learning rate decay of 1% at every iteration. We apply dropout on pairwise scorer hidden layers and on feature embeddings with a rate of 0.2. Dropout is also applied on input embeddings with a rate of 0.5.

We implement the learning of LSTM initial hidden states (Gers et al. 2002). Compared to initializing them randomly or keeping them at zero, this approach allowed us to increase significantly the performance of our model during development.

6.8 Results

Performance on Gold Mentions. Performance of all experiments on gold mentions are presented in Table 6.11. We report the score distributions over 10 runs in Figure 6.14.

Surprisingly, our cluster-level representation degrades the performance of our system. For both models (people and events), the average CoNLL f1-score is lower for the configuration that includes the cluster-level representation. Also, the scores indicate that using a cluster-level representation seems to increase precision and decrease recall. Intuitively, this additional information allows to better match mentions within clusters and therefore increase precision.

We observe that there is an overlap between the two distributions (baseline vs. cluster-rpr). It means that comparing two extreme values of these models could have resulted in an opposite conclusion. This confirms the need for distribution comparisons instead of single-shot comparisons which has already been mentioned in previous efforts (Reimers and Gurevych 2017).

Concerning the other experiments, only 3 out of 7 configurations did perform better than our baseline approach. Network pretraining is the component that brings the largest f1-score increase. The event model benefits the most from this pretraining step with an increase of +2.69 F1 while the people model performance is only improved by +0.43 F1.

The attention mechanism brings an increase of +0.39 F1, balanced between people and event models. Our filtering strategy does improve the performance by only +0.01 F1.

Concerning the components that degrade performance, the use of character embeddings decreases the overall f1-score of both people and event models. However, the final decrease when models are combined is limited at -0.06 F1.

Both feature embeddings (entity type and DCT relation) decrease the performance of both people and event models. The drop is the largest for the latter with a decrease of -1.7 F1 for the entity type and -3.31 F1 for the DCT relation. The performance gap is narrower for the people model with a decrease of -0.16 F1 for the entity type and -0.5 F1 for the DCT relation.

The strategy involving the specialization of the last feed-forward network decreases the overall performance (-1.85 F1). However this is not the case for the people model for which we note a +0.08 F1 increase while event model performance decreases by -3.62 F1.

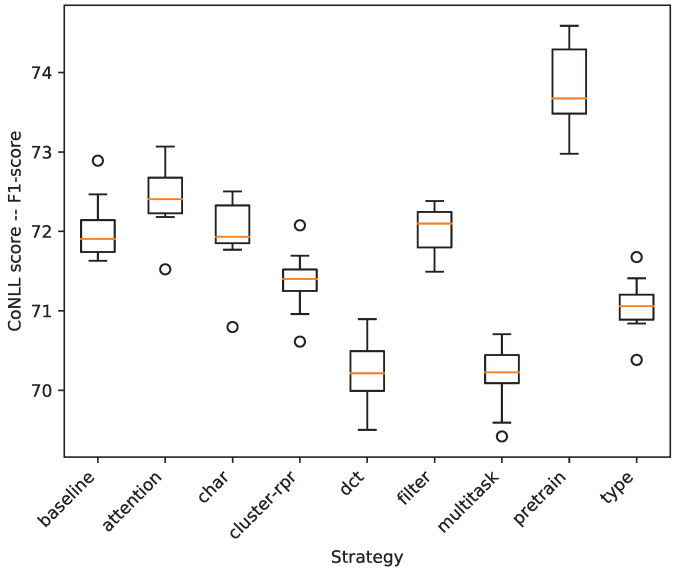

Figure 6.14: CoNLL f1-score distribution on gold mentions over 10 runs for all configurations. Small circles represent distribution outliers.

Performance on Predicted Mentions (end-to-end task). Performance of all experiments on predicted mentions are presented in Table 6.12. We report the score distributions in Figure 6.15

Overall, the scores for the combined version of our models drop on average by -5.00 F1 (between -4.7 and -5.38). People models suffer less from the error propagation phenomenon with a decrease ranging from -0.06 F1 to -0.51 F1. Event model decrease is more pronounced (between -0.79 F1 and -1.92 F1).

Other performance differences include the fact that the type embedding increases people model performance by +0.18 F1. Concerning the event model, the attention mechanism decreases the performance by -0.31 F1 and the filtering approach by -0.63 F1

6.9 Discussion

Cluster-Level Representation. Our experiment with the cluster-level representation seems to lead to an opposite conclusion to the one presented in Wiseman et al. (2016). However,

	MUC			B^3			$CEAF_e$			CoNLL			
experiment	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
People model													
baseline	95.78 (± 0.46)	93.33 (± 0.51)	94.54 (± 0.19)	92.16 (± 0.72)	88.97 (± 1.22)	90.53 (± 0.57)	75.35 (± 1.13)	66.16 (± 1.44)	70.44 (± 0.71)	87.77 (± 0.71)	82.82 (± 0.99)	85.17 (± 0.46)	
attention	95.98 (± 0.41)	93.52 (± 0.31)	94.73 (± 0.14)	92.46 (± 0.74)	89.59 (± 0.63)	91.00 (± 0.35)	75.79 (± 1.55)	67.15 (± 1.39)	71.19 (± 0.79)	88.08 (± 0.82)	83.42 (± 0.55)	85.64 (± 0.38)	↑
char	95.95 (± 0.49)	93.07 (± 0.36)	94.49 (± 0.15)	92.41 (± 0.67)	88.70 (± 0.59)	90.51 (± 0.16)	75.26 (± 1.83)	65.89 (± 1.30)	70.23 (± 0.40)	87.87 (± 0.98)	82.55 (± 0.69)	85.08 (± 0.21)	↓
cluster	95.11 (± 0.51)	93.51 (± 0.39)	94.30 (± 0.19)	90.65 (± 0.78)	89.94 (± 0.58)	90.29 (± 0.36)	76.67 (± 1.30)	63.64 (± 1.52)	69.53 (± 1.01)	87.48 (± 0.77)	82.36 (± 0.72)	84.71 (± 0.49)	↓
dct	95.87 (± 0.72)	92.76 (± 0.77)	94.28 (± 0.13)	92.33 (± 1.03)	88.66 (± 1.12)	90.44 (± 0.28)	75.64 (± 2.33)	64.08 (± 2.78)	69.29 (± 0.97)	87.95 (± 1.32)	81.83 (± 1.53)	84.67 (± 0.43)	↓
filter	95.73 (± 0.72)	93.29 (± 0.47)	94.49 (± 0.19)	92.02 (± 1.04)	89.04 (± 0.78)	90.50 (± 0.44)	75.07 (± 2.82)	66.12 (± 1.65)	70.25 (± 0.66)	87.60 (± 1.48)	82.82 (± 0.89)	85.08 (± 0.39)	↓
multitask	96.20 (± 0.46)	93.04 (± 0.39)	94.59 (± 0.09)	92.69 (± 0.54)	88.43 (± 0.84)	90.51 (± 0.40)	76.67 (± 1.49)	65.55 (± 1.74)	70.64 (± 0.83)	88.52 (± 0.76)	82.34 (± 0.86)	85.25 (± 0.34)	↑
pretrain	96.49 (± 0.23)	92.90 (± 0.34)	94.66 (± 0.12)	93.29 (± 0.39)	88.58 (± 0.80)	90.88 (± 0.32)	77.00 (± 0.72)	66.32 (± 1.46)	71.25 (± 0.85)	88.93 (± 0.30)	82.60 (± 0.79)	85.60 (± 0.41)	↑
type	96.16 (± 0.49)	92.96 (± 0.38)	94.53 (± 0.23)	92.66 (± 0.72)	88.59 (± 0.56)	90.58 (± 0.41)	76.11 (± 2.07)	64.74 (± 1.33)	69.94 (± 0.74)	88.31 (± 1.05)	82.09 (± 0.60)	85.01 (± 0.40)	↓
Event model													
baseline	65.83 (± 1.86)	57.82 (± 1.41)	61.53 (± 0.44)	62.02 (± 2.15)	54.93 (± 1.60)	58.20 (± 0.45)	67.61 (± 1.56)	51.10 (± 1.27)	58.18 (± 0.59)	65.15 (± 1.81)	54.62 (± 1.35)	59.30 (± 0.45)	
attention	67.32 (± 1.45)	57.28 (± 2.15)	61.84 (± 0.82)	63.73 (± 1.86)	54.08 (± 2.37)	58.44 (± 0.84)	68.10 (± 1.07)	51.17 (± 1.84)	58.40 (± 1.01)	66.39 (± 1.45)	54.18 (± 2.08)	59.56 (± 0.86)	↑
char	67.08 (± 1.63)	56.02 (± 1.85)	61.01 (± 0.84)	63.99 (± 2.07)	53.13 (± 2.01)	58.00 (± 0.91)	68.11 (± 1.12)	50.63 (± 1.55)	58.06 (± 1.04)	66.39 (± 1.57)	53.26 (± 1.76)	59.02 (± 0.92)	↓
cluster	66.57 (± 1.14)	55.19 (± 1.12)	60.33 (± 0.53)	62.89 (± 1.32)	52.24 (± 1.31)	57.04 (± 0.60)	68.12 (± 0.73)	48.95 (± 1.06)	56.96 (± 0.70)	65.86 (± 1.03)	52.12 (± 1.10)	58.11 (± 0.58)	↓
dct	64.55 (± 1.64)	52.26 (± 1.45)	57.73 (± 0.84)	61.59 (± 1.92)	49.58 (± 1.53)	54.89 (± 0.68)	66.37 (± 0.81)	47.49 (± 0.84)	55.36 (± 0.55)	64.17 (± 1.42)	49.78 (± 1.20)	55.99 (± 0.66)	↓
filter	66.20 (± 1.35)	57.29 (± 1.61)	61.39 (± 0.53)	62.78 (± 1.89)	54.43 (± 1.68)	58.26 (± 0.56)	67.27 (± 0.78)	51.38 (± 1.13)	58.25 (± 0.70)	65.42 (± 1.31)	54.37 (± 1.37)	59.30 (± 0.56)	=
multitask	64.19 (± 1.24)	52.25 (± 1.67)	57.58 (± 0.62)	61.08 (± 1.48)	49.55 (± 1.74)	54.67 (± 0.69)	65.70 (± 0.78)	47.05 (± 1.51)	54.81 (± 0.90)	63.66 (± 1.12)	49.62 (± 1.60)	55.68 (± 0.72)	↓
pretrain	69.57 (± 2.11)	58.93 (± 2.97)	63.71 (± 1.10)	66.99 (± 2.63)	55.90 (± 3.16)	60.81 (± 1.11)	69.31 (± 1.65)	55.26 (± 2.22)	61.44 (± 1.17)	68.63 (± 2.09)	56.70 (± 2.77)	61.99 (± 1.11)	↑
type	64.74 (± 1.82)	55.44 (± 2.11)	59.67 (± 0.83)	61.24 (± 2.18)	52.60 (± 2.26)	56.51 (± 0.74)	66.11 (± 1.37)	49.55 (± 1.39)	56.62 (± 0.72)	64.03 (± 1.76)	52.53 (± 1.88)	57.60 (± 0.75)	↓
Combined													
baseline	83.77 (± 1.19)	78.19 (± 0.69)	80.88 (± 0.30)	77.70 (± 1.45)	71.39 (± 0.98)	74.40 (± 0.47)	69.20 (± 1.40)	54.23 (± 1.15)	60.78 (± 0.53)	76.89 (± 1.31)	67.94 (± 0.89)	72.02 (± 0.38)	
attention	84.74 (± 0.94)	78.07 (± 1.01)	81.25 (± 0.18)	78.89 (± 1.29)	71.25 (± 1.46)	74.85 (± 0.31)	69.68 (± 0.81)	54.46 (± 1.46)	61.12 (± 0.76)	77.77 (± 1.01)	67.93 (± 1.27)	72.41 (± 0.40)	↑
char	84.71 (± 1.05)	77.26 (± 0.85)	80.81 (± 0.28)	79.05 (± 1.38)	70.32 (± 1.10)	74.41 (± 0.35)	69.59 (± 1.09)	53.80 (± 1.24)	60.67 (± 0.81)	77.78 (± 1.14)	67.13 (± 1.01)	71.96 (± 0.46)	↓
cluster	84.17 (± 0.67)	77.18 (± 0.48)	80.52 (± 0.27)	77.80 (± 0.84)	70.46 (± 0.69)	73.94 (± 0.32)	69.90 (± 0.64)	52.01 (± 0.90)	59.64 (± 0.63)	77.29 (± 0.69)	66.55 (± 0.64)	71.37 (± 0.37)	↓
dct	83.88 (± 1.08)	75.48 (± 0.80)	79.45 (± 0.34)	78.08 (± 1.38)	68.47 (± 1.02)	72.94 (± 0.36)	68.29 (± 0.77)	50.88 (± 0.99)	58.30 (± 0.54)	76.75 (± 1.04)	64.94 (± 0.88)	70.23 (± 0.38)	↓
filter	84.02 (± 1.00)	77.94 (± 0.83)	80.86 (± 0.17)	78.07 (± 1.27)	71.16 (± 0.96)	74.44 (± 0.27)	68.84 (± 0.76)	54.45 (± 1.12)	60.79 (± 0.54)	76.98 (± 1.00)	67.85 (± 0.90)	72.03 (± 0.29)	↑
multitask	83.91 (± 0.88)	75.65 (± 0.74)	79.56 (± 0.15)	78.02 (± 1.03)	68.34 (± 0.98)	72.84 (± 0.36)	67.98 (± 0.66)	50.76 (± 1.35)	58.10 (± 0.75)	76.64 (± 0.84)	64.92 (± 1.00)	70.17 (± 0.38)	↓
pretrain	85.85 (± 1.27)	78.42 (± 1.24)	81.95 (± 0.28)	80.63 (± 1.73)	71.72 (± 1.67)	75.88 (± 0.43)	70.79 (± 1.43)	57.68 (± 1.70)	63.53 (± 0.89)	79.09 (± 1.43)	69.27 (± 1.53)	73.79 (± 0.53)	↑
type	83.70 (± 1.27)	76.96 (± 0.98)	80.18 (± 0.31)	77.68 (± 1.65)	70.00 (± 1.18)	73.62 (± 0.28)	68.11 (± 1.17)	52.69 (± 1.19)	59.40 (± 0.52)	76.50 (± 1.35)	66.55 (± 1.08)	71.06 (± 0.33)	↓

Table 6.11: Detailed coreference scores on gold mentions for all configurations. We report precision (P), recall (R) and f1-score (F1) for all four metrics averaged over 10 runs. We report the standard deviation in brackets.

	MUC			B^3			$CEAF_e$			CoNLL			
experiment	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
People model													
baseline	94.97 (± 0.38)	94.83 (± 0.44)	94.90 (± 0.18)	90.60 (± 0.55)	91.50 (± 1.01)	91.04 (± 0.42)	71.24 (± 1.06)	65.75 (± 1.38)	68.37 (± 0.80)	85.60 (± 0.63)	84.03 (± 0.90)	84.77 (± 0.44)	
attention	95.17 (± 0.43)	94.96 (± 0.25)	95.06 (± 0.16)	90.96 (± 0.66)	91.73 (± 0.64)	91.34 (± 0.38)	72.53 (± 1.66)	66.41 (± 1.28)	69.31 (± 0.65)	86.22 (± 0.90)	84.36 (± 0.60)	85.24 (± 0.35)	↑
char	95.10 (± 0.53)	94.61 (± 0.35)	94.85 (± 0.15)	90.82 (± 0.72)	91.36 (± 0.51)	91.09 (± 0.22)	71.31 (± 2.57)	65.28 (± 1.63)	68.10 (± 0.76)	85.74 (± 1.27)	83.75 (± 0.78)	84.68 (± 0.35)	↓
cluster	95.47 (± 0.32)	94.17 (± 0.39)	94.82 (± 0.15)	91.22 (± 0.58)	90.93 (± 0.56)	91.07 (± 0.24)	74.30 (± 1.63)	60.55 (± 1.42)	66.70 (± 0.90)	87.00 (± 0.82)	81.88 (± 0.76)	84.20 (± 0.40)	↓
dct	95.25 (± 0.69)	94.30 (± 0.68)	94.76 (± 0.11)	91.05 (± 0.93)	90.95 (± 1.13)	90.99 (± 0.30)	72.83 (± 2.90)	63.59 (± 3.63)	67.75 (± 1.19)	86.38 (± 1.47)	82.95 (± 1.79)	84.50 (± 0.48)	↓
filter	94.91 (± 0.64)	94.83 (± 0.45)	94.87 (± 0.16)	90.51 (± 0.84)	91.78 (± 0.63)	91.13 (± 0.27)	71.27 (± 2.61)	65.52 (± 2.19)	68.19 (± 0.72)	85.56 (± 1.34)	84.04 (± 1.07)	84.73 (± 0.34)	↓
multitask	95.32 (± 0.46)	94.66 (± 0.33)	94.99 (± 0.08)	91.11 (± 0.59)	91.43 (± 0.68)	91.27 (± 0.28)	73.13 (± 1.63)	64.93 (± 1.27)	68.76 (± 0.63)	86.52 (± 0.83)	83.67 (± 0.70)	85.00 (± 0.27)	↑
pretrain	95.65 (± 0.30)	94.71 (± 0.27)	95.18 (± 0.11)	91.67 (± 0.45)	91.20 (± 0.77)	91.43 (± 0.26)	73.36 (± 1.60)	65.47 (± 1.36)	69.16 (± 0.71)	86.90 (± 0.71)	83.79 (± 0.69)	85.26 (± 0.32)	↑
type	95.50 (± 0.44)	94.59 (± 0.31)	95.05 (± 0.16)	91.37 (± 0.66)	91.25 (± 0.58)	91.31 (± 0.35)	73.18 (± 1.72)	64.42 (± 1.38)	68.50 (± 0.73)	86.68 (± 0.91)	83.42 (± 0.64)	84.95 (± 0.34)	↑
Event model													
baseline	59.65 (± 2.01)	60.04 (± 1.48)	59.80 (± 0.65)	55.95 (± 2.31)	57.59 (± 1.58)	56.70 (± 0.67)	62.51 (± 1.63)	53.35 (± 1.17)	57.54 (± 0.64)	59.37 (± 1.95)	56.99 (± 1.32)	58.01 (± 0.62)	
attention	63.43 (± 1.15)	56.00 (± 2.07)	59.44 (± 0.93)	60.53 (± 1.40)	53.08 (± 2.22)	56.51 (± 0.97)	64.49 (± 1.02)	51.37 (± 1.87)	57.16 (± 1.04)	62.82 (± 1.18)	53.48 (± 2.03)	57.70 (± 0.97)	↓
char	60.99 (± 1.59)	58.19 (± 2.07)	59.51 (± 0.87)	57.86 (± 1.89)	55.76 (± 2.20)	56.73 (± 0.92)	63.11 (± 1.01)	52.97 (± 1.69)	57.57 (± 1.01)	60.65 (± 1.47)	55.64 (± 1.94)	57.94 (± 0.93)	↓
cluster	62.46 (± 1.08)	55.31 (± 1.41)	58.65 (± 0.67)	59.53 (± 1.30)	52.44 (± 1.58)	55.73 (± 0.71)	63.50 (± 0.82)	50.64 (± 1.09)	56.33 (± 0.67)	61.83 (± 1.03)	52.80 (± 1.34)	56.90 (± 0.68)	↓
dct	58.91 (± 1.32)	54.36 (± 1.42)	56.52 (± 0.76)	56.03 (± 1.57)	52.04 (± 1.51)	53.92 (± 0.58)	61.65 (± 0.81)	49.91 (± 0.72)	55.16 (± 0.42)	58.86 (± 1.15)	52.10 (± 1.15)	55.20 (± 0.57)	↓
filter	59.63 (± 1.21)	58.33 (± 1.55)	58.94 (± 0.45)	56.48 (± 1.69)	55.91 (± 1.64)	56.15 (± 0.57)	61.85 (± 0.77)	52.96 (± 1.14)	57.05 (± 0.67)	59.32 (± 1.20)	55.74 (± 1.34)	57.38 (± 0.55)	↓
multitask	58.18 (± 1.27)	54.51 (± 1.56)	56.26 (± 0.76)	55.07 (± 1.33)	52.29 (± 1.66)	53.61 (± 0.78)	60.85 (± 0.98)	49.51 (± 1.53)	54.58 (± 0.88)	58.03 (± 1.16)	52.10 (± 1.55)	54.81 (± 0.79)	↓
pretrain	62.79 (± 2.41)	62.65 (± 3.09)	62.61 (± 0.76)	59.78 (± 2.92)	60.12 (± 3.34)	59.80 (± 0.78)	64.00 (± 1.86)	58.35 (± 2.19)	60.99 (± 0.97)	62.19 (± 2.37)	60.38 (± 2.86)	61.13 (± 0.83)	↑
type	58.28 (± 1.66)	56.53 (± 2.32)	57.34 (± 0.83)	55.00 (± 2.07)	54.13 (± 2.46)	54.48 (± 0.71)	60.77 (± 1.30)	51.17 (± 1.51)	55.53 (± 0.82)	58.02 (± 1.64)	53.95 (± 2.05)	55.78 (± 0.77)	↓
Combined													
baseline	80.88 (± 1.20)	72.56 (± 0.49)	76.49 (± 0.36)	74.35 (± 1.43)	65.01 (± 0.72)	69.36 (± 0.43)	64.75 (± 1.23)	47.90 (± 0.90)	55.05 (± 0.46)	73.33 (± 1.27)	61.82 (± 0.65)	66.96 (± 0.37)	
attention	83.50 (± 0.69)	71.19 (± 0.75)	76.85 (± 0.22)	77.57 (± 0.90)	63.14 (± 1.11)	69.60 (± 0.37)	66.47 (± 0.66)	46.43 (± 1.30)	54.65 (± 0.80)	75.85 (± 0.73)	60.25 (± 1.04)	67.03 (± 0.46)	↑
char	81.92 (± 1.15)	71.79 (± 0.82)	76.51 (± 0.31)	75.73 (± 1.39)	64.18 (± 1.06)	69.46 (± 0.41)	65.21 (± 1.03)	47.52 (± 1.24)	54.96 (± 0.85)	74.29 (± 1.16)	61.16 (± 1.01)	66.98 (± 0.51)	↑
cluster	83.23 (± 0.70)	70.53 (± 0.55)	76.35 (± 0.25)	77.10 (± 0.90)	62.55 (± 0.72)	69.05 (± 0.27)	65.92 (± 0.77)	45.09 (± 0.87)	53.55 (± 0.61)	75.42 (± 0.75)	59.39 (± 0.69)	66.32 (± 0.36)	↓
dct	81.45 (± 0.91)	70.26 (± 0.65)	75.44 (± 0.26)	75.18 (± 1.15)	62.41 (± 0.85)	68.19 (± 0.32)	64.25 (± 0.78)	45.08 (± 0.88)	52.98 (± 0.47)	73.63 (± 0.88)	59.25 (± 0.76)	65.53 (± 0.32)	↓
filter	81.12 (± 0.94)	71.96 (± 0.67)	76.26 (± 0.18)	74.75 (± 1.14)	64.44 (± 0.81)	69.20 (± 0.23)	64.15 (± 0.59)	47.55 (± 0.99)	54.61 (± 0.55)	73.34 (± 0.87)	61.32 (± 0.78)	66.69 (± 0.29)	↓
multitask	81.11 (± 0.74)	70.50 (± 0.59)	75.43 (± 0.27)	74.70 (± 0.83)	62.71 (± 0.81)	68.17 (± 0.39)	63.67 (± 0.68)	44.97 (± 1.18)	52.70 (± 0.78)	73.16 (± 0.71)	59.39 (± 0.85)	65.43 (± 0.44)	↓
pretrain	82.52 (± 1.55)	73.43 (± 1.10)	77.69 (± 0.28)	76.50 (± 2.00)	65.98 (± 1.47)	70.81 (± 0.40)	66.13 (± 1.64)	51.34 (± 1.50)	57.78 (± 0.79)	75.05 (± 1.69)	63.59 (± 1.35)	68.76 (± 0.46)	↑
type	80.93 (± 1.27)	71.19 (± 0.86)	75.74 (± 0.27)	74.46 (± 1.58)	63.43 (± 1.03)	68.48 (± 0.24)	63.63 (± 1.10)	46.14 (± 1.13)	53.47 (± 0.62)	73.01 (± 1.28)	60.25 (± 0.98)	65.90 (± 0.33)	↓

Table 6.12: Detailed coreference scores on predicted mentions for all configurations. We report precision (P), recall (R) and f1-score (F1) for all four metrics averaged over 10 runs. We report the standard deviation in brackets.

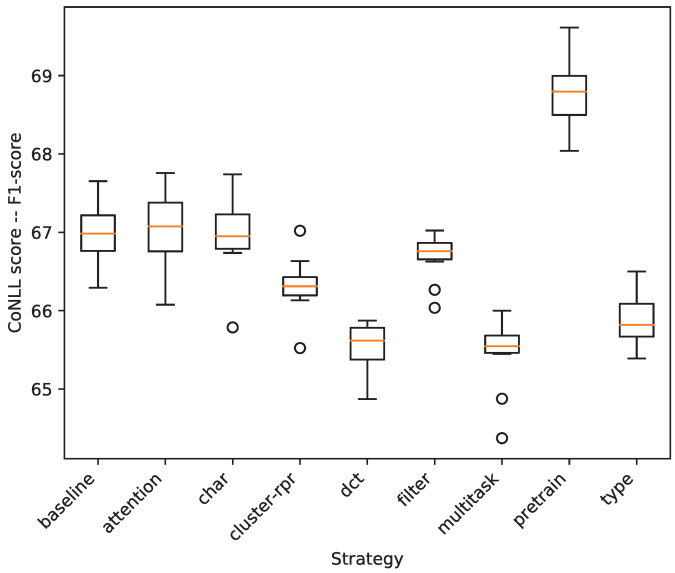

Figure 6.15: CoNLL f1-score distribution on predicted mentions over 10 runs for all configurations. Small circles represent distribution outliers.

there are several experimental disparities between the two research efforts. First, the domain is not the same, as Wiseman et al. (2016) use a corpus of news documents. Second, the authors did not run multiple experiments and perform a bootstrap resample test to evaluate the significance of their approach against a simpler version of their model. However, significance tests can be positive between two runs of the same model (Reimers and Gurevych 2017), suggesting that this method is not suitable for non-deterministic models such as the ones presented in Wiseman et al. (2016) and this chapter. Finally, the authors do not use dense representations as input but feature vectors extracted from the text.

Pretraining. Concerning the other components of our approach, the results confirm the general observation made in the literature that coreference resolution pretraining allows for a large increase in performance (K. Clark and Manning 2016b; Wiseman et al. 2016). We note that the increase is not as pronounced for the people model than it is for the event model. This can be explained by the number of singletons in the corpus. Table 6.13 presents the proportion of singletons across the different entity types. Only 10% of person mentions are singletons. Therefore the pretraining can only bring a marginal improvement for this entity type. We note also that test events have the highest percentage of singletons (86.86%).

Filtering. Interestingly, our filtering approach does not bring a large improvement. Intuitively, this could have allowed to eliminate some noise brought by other entity types. This result is coherent with the one obtained in the configuration where we used the entity type as feature embedding. In that particular case, the overall performance is decreased by -0.96 F1. These counter-intuitive results need further investigation. Specifically, the place where the embedding is used in the system should be questioned.

category	singletons		not singletons	
pronoun	2,103	60.76 %	1,358	39.24 %
test	11,937	86.86 %	1,806	13.14 %
person	1,911	10.46 %	16,357	89.54 %
treatment	8,225	58.54 %	5,825	41.46 %
problem	11,662	59.74 %	7,858	40.26 %

Table 6.13: Singleton distribution across categories.

Attention. Our attention mechanism allows for a small performance improvement, suggesting that the sentential context of the mentions brings only minor information for the classification. The event model is the one that benefits the most from this mechanism.

Multitasking. The multitasking approach, where the top-end feed forward layer is specialized according to the entity type brings a large decrease in performance (-1.8 F1). One possible reason for this performance drop is the limited size of our corpus. We have only 200 documents to train on while the main corpus for coreference resolution in the general domain proposes more than 2,000 documents (Pradhan et al. 2012; Pradhan et al. 2011).

Characters Embeddings. The use of character embeddings slightly decreases the performance of our approach (-0.06 F1). The character-level representation does not bring as much information as it is the case for other tasks such as NER (Lample et al. 2016; X. Ma and Hovy 2016). Table 6.14 presents the number of tokens that compose the corpus mentions. We report the number of unique tokens and the number of unknown tokens, i.e. those for which we do not have a pre-computed embedding. We observe that most tokens have a pre-computed embedding. Character embeddings are used to provide an alternative representation in the case where there is a high proportion of unknown tokens. They also allow to account for character casing, suffixes and prefixes These features have proved to be useful in previous NER efforts (Lample et al. 2016). In our case, the actual form of the tokens seem to play little role for coreference resolution.

category	# tok.	unique tok.	unk. tok.
problem	19,786	2,659	155
treatment	11,333	1,816	129
person	23,857	1,622	610
pronoun	1,360	13	1
test	4,402	682	32

Table 6.14: Number of tokens, unique tokens and unknown tokens per category.

Temporal Feature. The last component, the temporal feature embedding, decreases the overall performance of our approach (-1.79 F1) with a larger gap for the event model (-

3.31 F1) than for the people model (-0.5 F1). Looking at these results, it seems that our temporal feature did not capture useful temporal information for coreference resolution. In that case, we need to rethink what kind of temporal information could be appropriate for coreference resolution.

One other explanation could be that the quality of this feature is not sufficient. The THYME and i2b2 corpora are not from the same clinical subdomain. During the 2017 edition of the Clinical TempEval shared task (Bethard et al. 2017), we obtained a f1-score of 0.51 for DCT relation extraction. The train corpus was composed of documents related to colon cancer patients and the test corpus comprised documents related to brain cancer patients. This represents a drop of -0.246 F1 in comparison to the best result obtained during the 2016 edition (Bethard et al. 2016) where both train and test corpora were from the same subdomain (colon cancer patients). The difference between the THYME and i2b2 corpora seems to be more pronounced than it was during the Clinical TempEval shared task. This could lead to an even more pronounced drop in performance.

Also, the way our approach incorporates the temporal feature is maybe not the best solution. During development, we tried to use a feature embedding at token-level by concatenating it to the character-level representation and the pre-computed word embedding without success. Other possibilities should be explored such as antecedent filtering according to their DCT relation. However this needs a carefully made mapping to account for all DCT combinations.

Finally, the way of computing DCT relations may not be appropriate. Assigning only one relation instead of a concatenation of relations could improve the performance. However, this would result in a large proportion of undetermined relations.

Predicted Mentions. Concerning the decrease observed when extracting coreference links on predicted mentions, the difference between people and event model drops can be explained by the performance of our NER model (Table 6.8). People and pronouns obtained a very high f1-score, well balanced between precision and recall. The consequence is that error propagation will be not as prominent as it is for the event model.

Although we observe a performance drop when using the character embeddings on both the people and event models, the combined score shows an improvement. This happens due to a combination of several effects. First, we are dealing with averages. There is an outlier for the event model (Visible on Figure 6.15). If we were to remove this outlier, we would observe a performance increase for the event model. Second, event and people models are combined by pairs (one run implies with our approach implies one pair of models). It means that the way they are combined is completely arbitrary. Finally, chains sizes combined to model performances affects the metrics when both prediction sets are combined. All these parameters explains the counter-intuitive result showed in Table 6.12. We note that the improvement is anyway very small and cannot be considered as significant.

Optimal Configuration. Finally, we experiment with an *optimal* configuration where components that bring a performance improvement are used. For the people model, we use pretraining, the attention mechanism and the multitasking approach. For the event model, we use pretraining, the attention mechanism and the filtering of entities according to their type.

Results of this experiment are presented in Table 6.15. We report the performance obtained on gold and predicted mentions. Predicting on gold mentions, we obtain an overall f1-score of 73.82. The event model reaches a f1-score of 62.16, which is an improvement of +0.17 F1 in comparison to the best score obtained on gold mentions during development. As for the people model, the performance is lower than the best one obtained during development (i.e. attention mechanism) suggesting that the individual benefits of each component do no add up to each other when combined or that the information captured by these three components are contradictory (-0.02 F1).

Concerning the scores obtained on predicted mentions, the overall f1-score is 65.96, representing a drop of -7.86 F1 compared the scores obtained on gold mention with the same model. The event model is the one which decreases the most with a drop of -5.95 F1 while people model performance decrease only by -0.45 F1.

	MUC			B^3			$CEAF_e$			CoNLL		
model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Gold Entities												
people	96.28 (± 0.60)	93.29 (± 0.34)	94.76 (± 0.24)	92.91 (± 0.82)	88.90 (± 0.88)	90.86 (± 0.58)	76.76 (± 2.35)	66.54 (± 1.08)	71.26 (± 1.07)	88.65 (± 1.22)	82.91 (± 0.69)	85.62 (± 0.61)
events	68.88 (± 1.31)	59.89 (± 1.65)	64.04 (± 0.72)	66.01 (± 1.45)	56.84 (± 1.79)	61.04 (± 0.77)	68.28 (± 1.36)	55.81 (± 1.46)	61.40 (± 0.92)	67.72 (± 1.34)	57.51 (± 1.61)	62.16 (± 0.78)
combined	85.34 (± 0.93)	79.05 (± 0.83)	82.07 (± 0.27)	79.85 (± 1.14)	72.35 (± 1.21)	75.90 (± 0.45)	69.90 (± 1.16)	58.17 (± 1.33)	63.48 (± 0.76)	78.36 (± 1.04)	69.86 (± 1.10)	73.82 (± 0.47)
Predicted Entities												
people	95.59 (± 0.51)	94.78 (± 0.35)	95.18 (± 0.21)	91.56 (± 0.72)	91.18 (± 0.93)	91.37 (± 0.51)	72.90 (± 2.51)	65.52 (± 1.92)	68.96 (± 1.12)	86.68 (± 1.22)	83.83 (± 0.93)	85.17 (± 0.56)
events	53.48 (± 1.77)	62.61 (± 1.73)	57.64 (± 0.79)	49.43 (± 1.88)	61.05 (± 1.95)	54.57 (± 0.79)	57.08 (± 1.57)	55.79 (± 1.39)	56.40 (± 0.80)	53.33 (± 1.72)	59.82 (± 1.65)	56.21 (± 0.77)
combined	77.22 (± 1.31)	73.44 (± 0.73)	75.27 (± 0.37)	70.10 (± 1.49)	66.40 (± 1.14)	68.18 (± 0.38)	60.27 (± 1.28)	49.64 (± 1.16)	54.42 (± 0.56)	69.20 (± 1.34)	63.16 (± 0.98)	65.96 (± 0.38)

Table 6.15: Best configuration performance. We report precision (P), recall (R) and f1-score (F1) for all four coreference metrics (MUC, B^3, $CEAF_e$, CoNLL). We present the scores for the configurations where gold mentions and predicted mentions are used.

Comparison to the Literature. We converted several system outputs submitted to the i2b2 shared task (task1c). Although we had access to other submissions besides the ones presented in this section, we were not able to convert them due to file format problems. System performances are presented in Table 6.16. Our approach is competitive with the ones presented during the workshop which reached f1-scores ranging from 68.34 to 73.41.

Most approaches devised for coreference resolution in the clinical domain use a rich feature set to build mention representations. These features are extracted from the text itself and from external resources such as clinical knowledge bases. Our model achieves similar performance without the use of such features. Incorporating this information into our model as we did in our temporal information extraction approach (cf. Chapter 6) could improve the global performance.

	MUC			B^3			$CEAF_e$			CoNLL		
model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Jindal and Roth (2011)	72,34	**89,81**	80,14	65,01	**86,18**	74,11	59,25	**74,45**	**65,98**	65,53	**83,48**	73,41
Hinote et al. (2011)	85,32	74,72	79,67	78,23	65,90	71,53	64,68	56,34	60,23	76,08	65,65	70,48
Anick et al. (2011)	**87,80**	73,21	79,85	**82,83**	58,21	68,37	69,25	52,72	59,86	**79,96**	61,38	69,36
Grouin et al. (2011)	77,49	79,88	78,67	55,98	66,91	60,96	**72,62**	59,46	65,38	68,70	68,75	68,34
our model	85.34 (± 0.93)	79.05 (± 0.83)	**82.07** (± 0.27)	79.85 (± 1.14)	72.35 (± 1.21)	**75.90** (± 0.45)	69.90 (± 1.16)	58.17 (± 1.33)	63.48 (± 0.76)	78.36 (± 1.04)	69.86 (± 1.10)	**73.82** (± 0.47)

Table 6.16: Performance comparison with other system outputs submitted during the i2b2 shared task (Uzuner et al. 2012). We converted the runs to the CoNLL format and evaluated with the official CoNLL scorer. We report precision (P), recall (R) and f1-score (F1) for all four coreference metrics (MUC, B^3, $CEAF_e$, CoNLL).

6.10 Conclusion

In this chapter, we presented a neural entity-based approach for coreference resolution in the clinical domain. Our attempt to include a cluster-level representation in our system did not bring any performance improvement. It may suggest that our approach is not suitable for the clinical domain. Other possibilities for building an entity-aware approach will be investigated. For instance we will look at recent cluster-merging approaches such as the one presented in K. Clark and Manning (2016) or joint approaches where mentions and coreference links are learned together such as the one presented in K. Lee et al. (2017).

The temporal feature did not bring any performance improvement. This result could due to the quality of our feature, to the way of using it into our network or to the nature of the feature itself.

Our experiments suggest that pretraining, which has proven to be valuable for coreference resolution in the general domain, allows for better performance in our situation. Then, character embeddings seem to play little role for coreference resolution as they do no bring any performance improvement. The same observation can be made for our filtering and multi-tasking approaches, although their poor performance could be due to the limited size of our corpus.

Overall, our model reaches a f1-score of 65.96 (when using predicted mentions) and 73.82 (when using gold mentions) and is competitive with the scores obtained during the i2b2 shared task. With the transformation of the i2b2 corpus to CoNLL format and its future release, other research teams will have the possibility to develop coreference resolution models for the clinical domain and may compare objectively their results using a reference implementation of the coreference metrics.

Chapter 7

Conclusion

Extracting clinical timelines from electronic health records is a complex task that is related to several NLP research topics. This thesis presented several contributions to temporal information extraction and coreference resolution in clinical narratives. This chapter makes a brief summary of the thesis (Section 7.1), discusses future research directions (Section 7.2) and addresses some of the remaining challenges for clinical timeline extraction (Section 7.3).

7.1 Summary

Temporal Information Extraction

In the first part of this thesis, dedicated to temporal information extraction, we started by reviewing the literature on the topic. We showed that time is a complex linguistic phenomenon that has been researched for a long time by the community. We presented the resources annotated with temporal information. Then, we reviewed computational linguistics approaches that have been devised for automatic temporal information extraction.

Our first main contribution to the topic is a feature-based approach for narrative container extraction. We tested our approach on clinical documents written in English and showed competitive performance. As we mentioned in the introduction, research efforts in the NLP community tend to be biased toward the English language as most annotated corpora are written in that language. However, we showed that our approach can be adapted to other languages by replacing language sensitive resources along our preprocessing pipeline. We experimented on a corpus of clinical documents written in French and obtained comparable results to the ones obtained on the English dataset.

Our second main contribution is a neural-based approach for temporal information extraction. We devised an approach that makes use of the rich feature sets available in most corpora. We studied how the inclusion of such categorical features influence the performance of our approach. Then, we performed an evaluation of our approach in the context of the 2017 edition of the Clinical TempEval shared task that included a track on domain adaptation and obtained competitive performance.

Furthermore, noticing a lack of efficient neural sequence labeling tools in the community, we decided to pack our module for NER into an open source tool called YASET available online[1]. This tool reaches state-of-the-art performance on various NER tasks.

Coreference Resolution

The second part of our thesis was related to clinical event coreference resolution. We started by reviewing the literature on the topic. We presented the different aspects of this phenomenon from a linguistic perspective. Specifically, we addressed the terminology differences between linguistics and computational linguistics. We emphasized the differences between event coreference in the general and the clinical domains. Then, we reviewed the literature in computational linguistics. We presented the resources and the approaches that have been devised for automatic coreference resolution.

In our main contribution to the topic, we devised a neural entity-based coreference resolution approach. We performed an empirical study with several configurations ranging from the use of an attention mechanism to pretraining the network, to measure how the use of these components influences the overall performance. As we mentioned in the introduction, coreference resolution could benefit from temporal information. To verify this hypothesis, we devised a temporal feature based on the DCT relations that was included in our neural model. Unfortunately, this feature did not bring any performance improvement.

7.2 Future Research Directions

Starting from the work presented in this thesis, several research directions arise. First, our approaches for temporal information extraction and coreference resolution are done in a pipeline fashion where each step is independent from the other. However, learning jointly multiple tasks can be beneficial. Clues used in several subtasks add-up to each other and improve the overall performance. For instance, Leeuwenberg and Moens (2017) devised an approach for jointly learning containment and DCT relations. For coreference resolution, K. Lee et al. (2017) learned a joint model for mention extraction and coreference resolution and improved the state-of-the-art in the general domain.

In addition, both NLP topics addressed in this thesis involve structured predictions where classification decisions depend on previously made decisions. In this context, several approaches can be used to take into account this dependence. For instance, imitation learning is an active machine learning topic which allows to model such problems adequately. It has been successfully applied for coreference resolution (K. Clark and Manning 2016a) and could be applied to temporal relation extraction as well.

Then, the experiments involving the use of categorical features in our neural network approaches showed promising results. Their individual benefit need to be assessed. Furthermore, the way of incorporating them into a neural network need also further investigation. Positioning them at the adequate depth level in our network could bring further performance improvement.

Finally, within-sentence relation extraction can benefit from the syntactic structure of the sentence. It has been tested in the general domain and showed promising results (Miwa and

1. https://github.com/jtourille/yaset

Bansal 2016). The idea can also be implemented at the scale of the document (cross-sentence relation extraction). Peng et al. (2017) performed n-ary relation extraction over multiple sentences by modeling documents as graphs. They extracted drug–gene interactions from a large corpus of biomedical research papers.

7.3 Extracting Clinical Timelines: Are We There Yet?

Automatic timeline extraction from text is still an unresolved task. Among all remaining issues is the fact that most research efforts on the subject address within-document clinical timeline extraction. As we mentioned in the introductory chapter, the objective is to provide a temporal summary of a given electronic health record. This task requires corpora annotated with cross-document temporal information and coreference links. Moreover, the community needs an annotated resource that provides target timelines as examples.

This rises another issue related to data availability and corpora creation. Clinical data is sensitive by nature and must me handled appropriately. Most annotation efforts are unknown from the NLP community as they cannot be distributed freely. Moreover, the annotation process is highly costly as the knowledge needed to annotate clinical concepts can only be acquired with a thorough training.

Although unstructured data taking the form of raw text is the principal source of information in EHR, there is structured data that is not yet used in NLP research efforts. This structured data includes laboratory results, prescriptions and vital recording. Being able to leverage this structured information by combining it to information extracted from clinical documents could benefit to clinical timeline extraction.

Another issue concerns the diversity of clinical subdomains. The clinical domain is in fact a collection of subdomains where text genre, style and vocabulary may differ. As we have shown in this thesis, the performance drops significantly when train and test domains are not the same. As we cannot annotate all clinical subdomains, the NLP research community needs to work on clinical domain adaptation.

In this thesis, we worked with event mentions. For temporal information, we extracted relations between event mentions and/or temporal expressions, while for coreference resolution, we regrouped event mentions that are referring to the same real-world event. However, if we want to be able to query information extracted from EHRs, we need to normalize these events by associating them to a clinical concept (e.g. a clinical concept from UMLS®). The task is called entity normalization and is an active research topic.

Finally, another issue concerns the complexity of the involved NLP tasks. As we showed in this thesis, both temporal information extraction and coreference resolution tasks can not be considered as resolved. The linguistic phenomena involved in these tasks are difficult to capture within annotation schemes and automatic extraction approaches. Further research efforts are needed on these two complex topics.

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* (cit. on p. 61).

Ahn, David, Sisay F. Adafre, and Maarten de Rijke (Apr. 2005). "Towards Task-Based Temporal Extraction and Recognition". In: *Annotating, Extracting and Reasoning About Time and Events* (Schloss Dagstuhl, Germany, Apr. 10–15, 2005). Ed. by Frank Schilder, Graham Katz, and James Pustejovsky. Springer (cit. on p. 12).

Allan, James, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang (Feb. 1998a). "Topic Detection and Tracking Pilot Study: Final Report". In: *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop* (Herdnon, Virgina, USA, Feb. 8, 1998–Feb. 11, 1998). Defense Advanced Research Projects Agency, pp. 194–218 (cit. on p. 20).

Allan, James, Ron Papka, and Victor Lavrenko (Aug. 1998b). "On-line New Event Detection and Tracking". In: *Proceedings of the 21st ACM/SIGIR International Conference on Research and Developement in Information Retrieval* (Melbourne, Australia, Aug. 24, 1998–Aug. 28, 1998). Association for Computing Machinery, pp. 37–45 (cit. on p. 20).

Allen, James F. (1983). "Maintaining Knowledge About Temporal Intervals". In: *Communications of the Association for Computational Machinery* 26.11, pp. 832–843 (cit. on pp. 11, 16).

Anick, P., P. Hong, N. Xue, and al. (Oct. 2011). "Coreference Resolution for Electronic Medical Records". In: *Proceedings of the 2011 i2b2/VA/Cincinatti Workshop on Challenges in Natural Language Processing for Clinical Data* (Boston, Massachusetts, USA, Oct. 2011) (cit. on p. 121).

Aone, Chinatsu and Scott William Bennett (Aug. 1995). "Applying Machine Learning to Anaphora Resolution". In: *Proceedings of the 1995 International Joint Conference on Artificial Intelligence* (Montreal, Quebec, Aug. 20, 1995–Aug. 25, 1995), pp. 302–314 (cit. on pp. 80, 81).

Aronson, Alan R. (Nov. 2001). "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap Program". In: *Proceedings of the 2001 AMIA Annual Symposium*

(Washington, D.C., USA, Nov. 3, 2001–Nov. 7, 2001). American Medical Informatics Association, pp. 17–21 (cit. on pp. 36, 45, 86).

Asahara, Masayuki, Sachi Kato, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa (2014). "BCCWJ-TimeBank: Temporal and Event Information Annotation on Japanese Text". In: *International Journal of Computational Linguistics and Chinese Language Processing* 19.3, pp. 1–24 (cit. on p. 29).

Atefeh, Farzindar and Wael Khreich (2015). "A Survey of Techniques for Event Detection in Twitter". In: *Computational Intelligence* 31.1, pp. 132–162 (cit. on p. 21).

Bach, Emmon (1986). "The Algebra of Events". In: *Linguistics and Philosophy* 9, pp. 5–16 (cit. on pp. 12, 14).

Bagga, Amit and Breck Baldwin (May 1998). "Algorithms for Scoring Coreference Chains". In: *Proceedings of the 1st International Conference on Language Resources and Evaluation* (Granada, Spain, May 28, 1998–May 30, 1998). European Language Resources Association, pp. 563–566 (cit. on p. 88).

Barwise, Jon and John Perry (1981). "Situations and Attitudes". In: *The Journal of Philosophy* 78.11, pp. 668–691 (cit. on p. 71).

Bengtson, Eric and Dan Roth (Oct. 2008). "Understanding the Value of Features for Coreference Resolution". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Waikiki, Honolulu, Hawaii, USA, Oct. 25, 2008–Oct. 27, 2008). Association for Computational Linguistics, pp. 294–303 (cit. on pp. 81, 83).

Bergert, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra (1996). "A Maximum Entropy Approach to Natural Language Processing". In: *Computational Linguistics* 22.1, pp. 39–71 (cit. on pp. 81, 82).

Bergstra, James, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl (Dec. 2011). "Algorithms for Hyper-Parameter Optimization". In: *Proceedings of the 2011 Neural Information Processing Systems Conference* (Granada, Spain, Dec. 12, 2011–Dec. 14, 2011), pp. 2546–2554 (cit. on pp. 49, 50, 57).

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research* 12, pp. 281–305 (cit. on p. 49).

Bergstra, James, Daniel Yamins, and David Cox (June 2013). "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures". In: *Proceedings of the 30th International Conference on Machine Learning* (Atlanta, Georgia, US, June 16, 2013–June 21, 2013), pp. 115–123 (cit. on pp. 50, 57).

Bethard, Steven (2007). "Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach". PhD thesis. University of Colorado (cit. on pp. 13, 16).

Bethard, Steven (June 2013). "ClearTK-TimeML: A Minimalist Approach to TempEval 2013". In: *Proceedings of the 7th International Workshop on Semantic Evaluation* (Atlanta, Georgia, USA, June 14, 2013–June 15, 2013). Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 10–14 (cit. on pp. 34, 35).

Bethard, Steven, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen (June 2015). "SemEval-2015 Task 6: Clinical TempEval". In: *Proceedings of the 9th International Workshop on Semantic Evaluation* (Denver, Colorado, USA, June 4, 2015–June 15, 2015). Association for Computational Linguistics, pp. 806–814 (cit. on pp. 3, 23, 32, 42, 45, 50, 55, 57, 160).

Bethard, Steven and James H. Martin (July 2006). "Identification of Event Mentions and their Semantic Class". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, Australia, July 22, 2006–July 23, 2006). Association for Computational Linguistics, pp. 146–154 (cit. on p. 35).

Bethard, Steven, Philip Ogren, and Lee Becker (May 2014). "ClearTK 2.0: Design Patterns for Machine Learning in UIMA". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation* (Reykjavik, Iceland, May 26, 2014–May 31, 2014). European Language Resources Association, pp. 3289–3293 (cit. on p. 34).

Bethard, Steven, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen (June 2016). "SemEval-2016 Task 12: Clinical TempEval". In: *Proceedings of the 10th International Workshop on Semantic Evaluation* (San Diego, California, USA, June 16, 2016–June 17, 2016). Association for Computational Linguistics, pp. 1052–1062 (cit. on pp. 3, 5, 23, 32, 40, 42, 44, 47, 49, 55, 57, 68, 119, 152, 160, 162).

Bethard, Steven, Guergana Savova, Martha Palmer, and James Pustejovsky (Aug. 2017). "SemEval-2017 Task 12: Clinical TempEval". In: *Proceedings of the 11th International Workshop on Semantic Evaluation* (Vancouver, Canada, Aug. 3, 2017–Aug. 4, 2017). Association for Computational Linguistics, pp. 565–572 (cit. on pp. 3, 5, 23, 32, 40, 42, 55, 64, 119, 153–155, 160, 163).

Biber, Douglas, Stig Johansson, Geofrey Leech, Susan Conrad, and Edward Finegan (1999). *Longman Grammar of Spoken and Written English*. Longman (cit. on p. 11).

Bird, Steven and Edward Loper (July 2004). "NLTK: The Natural Language Toolkit". In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (Barcelona, Spain, July 21, 2004–July 26, 2004). Association for Computational Linguistics (cit. on p. 45).

Bittar, Andrè, Pascal Amsili, Pascal Denis, and Laurence Danlos (June 2011). "French TimeBank: An ISO-TimeML Annotated Reference Corpus". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 19, 2011–June 24, 2011). Association for Computational Linguistics, pp. 130–134 (cit. on pp. 11, 29).

Björkelund, Anders and Jonas Kuhn (June 2014). "Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, Maryland, USA, June 22, 2014–June 27, 2014). Association for Computational Linguistics, pp. 47–57 (cit. on pp. 85, 86).

Björkelund, Anders and Pierre Nugues (June 2011). "Exploring Lexicalized Features for Coreference Resolution". In: *Proceedings of the 15th Conference on Computational Natural*

Language Learning (Portland, Oregon, USA, June 23, 2011–June 24, 2011). Association for Computational Linguistics, pp. 45–50 (cit. on pp. 78, 81).

Bodenreider, Olivier (2004). "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology". In: *Nucleic Acids Research* 32, pp. 267–270 (cit. on p. 45).

Boguraev, Branimir and Rie Kubota Ando (July 2005). "TimeML-compliant Text Analysis for Temporal Reasoning". In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (Edinburgh, Scotland, July 30, 2005–Aug. 5, 2007), pp. 997–1003 (cit. on p. 36).

Bramsen, Philip, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay (Nov. 2006a). "Finding Temporal Order in Discharge Summaries". In: *Proceedings of the 2006 AMIA Annual Symposium* (Washington, D.C., USA, Nov. 11, 2006–Nov. 15, 2006). American Medical Informatics Association, pp. 81–85 (cit. on p. 22).

Bramsen, Philip, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay (July 2006b). "Inducing Temporal Graphs". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, Australia, July 22, 2006–July 23, 2006). Association for Computational Linguistics, pp. 189–198 (cit. on pp. 38, 64).

Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard (July 1987). "A Centering Approach to Pronouns". In: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics* (Stanford, California, USA, July 6, 1987–July 9, 1987). Association for Computational Linguistics, pp. 155–162 (cit. on p. 79).

Brown, Keith (2005). *Encyclopedia of Language and Linguistics*. Elsevier (cit. on p. 14).

Brown, Peter F., Peter V. deSouza, Robert L. Mercer, T. J. Watson, Vincent J. Della Pietra, and Jenifer C. Lai (1992). "Class-Based n-gram Models of Natural Language". In: *Computational Linguistics* 18.4, pp. 467–479 (cit. on p. 36).

Cai, Jie and Michael Strube (Sept. 2010). "Evaluation Metrics For End-to-End Coreference Resolution Systems". In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Tokyo, Japan, Sept. 24–25, 2010). Association for Computational Linguistics, pp. 28–36 (cit. on p. 89).

Campillos, Leonardo, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol (2018). "A French Clinical Corpus with Comprehensive Semantic Annotations: Development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)". In: *Language Resources and Evaluation* 52 (2), pp. 571–601 (cit. on pp. 3, 32, 33, 40, 160).

Cardie, Claire and Kiri Wagstaf (June 1999). "Noun Phrase Coreference as Clustering". In: *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (College Park, Maryland, USA, June 21–22, 1999). Association for Computational Linguistics, pp. 82–89 (cit. on pp. 80, 83).

Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof (June 2011). "Annotating Events, Temporal Expressions and Relations in Italian: the It-Timeml Experience for the Ita-TimeBank". In: *Proceedings of the 5th Linguistic An-*

notation Workshop (Portland, Oregon, USA, June 23–24, 2011). Portland, Oregon, USA: Association for Computational Linguistics, pp. 143–151 (cit. on p. 29).

Chambers, Nate (June 2013). "NavyTime: Event and Time Ordering from Raw Text". In: *Proceedings of the 7th International Workshop on Semantic Evaluation* (Atlanta, Georgia, USA, June 14, 2013–June 15, 2013). Association for Computational Linguistics, pp. 73–77 (cit. on p. 35).

Chambers, Nathanael and Dan Jurafsky (Oct. 2008). "Jointly Combining Implicit Constraints Improves Temporal Ordering". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Waikiki, Honolulu, Hawaii, USA, Oct. 25, 2008–Oct. 27, 2008). Association for Computational Linguistics, pp. 698–706 (cit. on p. 38).

Chang, Angel X. and Christopher D. Manning (May 2012). "SUTime: A Library for Recognizing and Normalizing Time Expressions". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (Istanbul, Turkey, May 21, 2012–May 27, 2012). European Language Resources Association (cit. on pp. 34, 35).

Chang, Kai-Wei, Rajhans Samdani, and Dan Roth (Oct. 2013). "A Constrained Latent Variable Model for Coreference Resolution". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, Washington, USA, Oct. 18, 2013–Oct. 21, 2013). Association for Computational Linguistics, pp. 601–612 (cit. on pp. 85, 86).

Chang, Kai-Wei, Rajhans Samdani, Alla Rozovskaya, Mark Sammons, and Dan Roth (July 2012). "Illinois-Coref: The UI System in the CoNLL-2012 Shared Task". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island, Korea, July 12, 2012–July 14, 2012). Association for Computational Linguistics, pp. 113–117 (cit. on pp. 82, 83).

Chang, Yung-Chun, Hong-Jie Dai, Johnny Chi-Yang Wu, Jian-Ming Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu (2013). "TEMPTING System: A Hybrid Method of Rule and Machine Learning for Temporal Relation Extraction in Patient Discharge Summaries". In: *Journal of Biomedical Informatics* 46 (2012 i2b2 NLP Challenge on Temporal Relations in Clinical Data), pp. 54–62 (cit. on p. 37).

Charniak, Eugene (1972). "Toward A Model of Children's Story Comprehension". PhD thesis. Massachusetts Institute of Technology (cit. on p. 79).

Charniak, Eugene and Mark Johnson (June 2005). "Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (Ann Arbor, Michigan, USA, June 25, 2005–June 30, 2005). Association for Computational Linguistics, pp. 173–180 (cit. on p. 45).

Cherry, Colin, Xiaodan Zhu, Joel Martin, and Berry de Bruijn (2013). "À la Recherche du Temps Perdu: Extracting Temporal Relations from Medical Text in the 2012 i2b2 NLP Challenge". In: *Journal of the American Medical Informatics Association* 20.5, pp. 843–848 (cit. on p. 36).

Chinchor, Nancy A. (Apr. 1998). "Overview of MUC-7/MET-2". In: *Proceedings of the 7th Message Understanding Conference* (Fairfax, Virginia, USA) (cit. on pp. 17, 22).

Clark, Herbert H. and Catherine R. Marshall (1981). "Definite Knowledge and Mutual Knowledge". In: *Elements of Discourse Understanding*, pp. 10–63 (cit. on p. 71).

Clark, Kevin and Christopher D. Manning (July 2015). "Entity-Centric Coreference Resolution with Model Stacking". In: *Proceedings of the Joint Conference of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (Beijing, China, July 26, 2015–July 31, 2015). Association for Computational Linguistics, pp. 1405–1415 (cit. on pp. 84, 106).

Clark, Kevin and Christopher D. Manning (Nov. 2016a). "Deep Reinforcement Learning for Mention-Ranking Coreference Models". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas, USA, Nov. 1, 2016–Nov. 5, 2016). Association for Computational Linguistics, pp. 2256–2262 (cit. on pp. 78, 84, 93, 112, 124).

Clark, Kevin and Christopher D. Manning (Aug. 2016b). "Improving Coreference Resolution by Learning Entity-Level Distributed Representations". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Berlin, Germany, Aug. 7, 2016–Aug. 12, 2016). Association for Computational Linguistics, pp. 643–653 (cit. on pp. 106, 110, 117, 122).

Cohan, Arman, Kevin Meurer, and Nazli Goharian (June 2016). "GUIR at SemEval-2016 task 12: Temporal Information Processing for Clinical Narratives". In: *Proceedings of the 10th International Workshop on Semantic Evaluation* (San Diego, California, USA, June 16, 2016–June 17, 2016). Association for Computational Linguistics, pp. 1248–1255 (cit. on pp. 35, 37).

Cohen, Kevin B., Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A. Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E. Hunter (2017). "Coreference Annotation and Resolution in the Colorado Richly Annotated Full Text (CRAFT) Corpus of Biomedical Journal Articles". In: *BMC Bioinformatics* 18.1 (cit. on p. 77).

Cohen, William W. (1995). "Fast Effective Rule Induction". In: *Proceedings of the 12th International Conference on Machine Learning* (Tahoe City, California, USA, July 1995), pp. 115–123 (cit. on p. 81).

Collins, Michael (July 2002). "Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (Philadelphia, Pennsylvania, USA, July 6, 2002–July 7, 2002). Association for Computational Lingusitics, pp. 1–8 (cit. on p. 85).

Comrie, Bernard (1976). *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems* (cit. on p. 12).

Cortes, Corinna and Vladimir Vapnik (1995). "Support-Vector Networks". In: *Machine Learning* 20.3, pp. 273–297 (cit. on pp. 81, 82).

Costa, Francisco and António Branco (May 2012). "TimeBankPT: A TimeML Annotated Corpus of Portuguese". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (Istanbul, Turkey, May 21, 2012–May 27, 2012). European Language Resources Association, pp. 3727–3734 (cit. on p. 29).

Crichton, Gamal, Sampo Pyysalo, Billy Chiu, and Anna Korhonen (2017). "A Neural Network Multi-task Learning Approach to Biomedical Named Entity Recognition". In: *BMC Bioinformatics* 18.1 (cit. on p. 110).

Culotta, Aron, Michael Wick, and Andrew McCallum (2007). "First-Order Probabilistic Models for Coreference Resolution". In: *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Los Angeles, California, USA, June 2, 2010–June 4, 2010). Association for Computational Linguistics, pp. 81–88 (cit. on pp. 84, 85).

Daelemans, Walter and Antal van den Bosch (2005). *Memory-Based Language Processing*. Cambridge University Press (cit. on p. 81).

Dalrymple, Mary, Stuart M. Shieber, and Fernando C. N. Pereira (1991). "Ellipsis and Higher-Order Unification". In: *Linguistics and Philosophy* 14, pp. 399–452 (cit. on p. 71).

Daumé III, Hal (2006). "Practical Structured Learning Techniques for Natural Language Processing". PhD thesis. University of Southern California (cit. on p. 84).

Daumé III, Hal and Daniel Marcu (Oct. 2005a). "A Large-Scale Exploration of Effective Global Features for a Joint Entity Detection and Tracking Model". In: *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing* (Oct. 6, 2005–Oct. 8, 2005). Association for Computational Linguistics (cit. on p. 84).

Daumé III, Hal and Daniel Marcu (Aug. 2005b). "Learning as Search Optimization: Approximate Large Margin Methods for Structured Prediction". In: *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany, Aug. 7, 2005–Aug. 11, 2005), pp. 169–176 (cit. on p. 84).

Deemter, Kees van and Rodger Kibble (2000). "On Coreferring: Coreference in MUC and Related Annotation Schemes". In: *Computational Linguistics* 26.4, pp. 629–637 (cit. on pp. 74, 75).

Denis, Pascal and Jason Baldridge (Oct. 2008). "Specialized Models and Ranking for Coreference Resolution". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Waikiki, Honolulu, Hawaii, USA, Oct. 25, 2008–Oct. 27, 2008). Association for Computational Linguistics, pp. 660–669 (cit. on pp. 81–83).

Denis, Pascal and Jason Baldridge (2009). "Global Joint Models for Coreference Resolution and Named Entity Classification". In: *Procesamiento del Lenguaje Natural* 42, pp. 87–96 (cit. on p. 81).

Denis, Pascal and Philippe Muller (July 2011). "Predicting Globally-Coherent Temporal Structures from Texts via Endpoint Inference and Graph Decomposition". In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (Barcelona, Spain, July 16, 2011–July 22, 2011) (cit. on pp. 38, 64).

Derczynski, Leon R. A. (2017). *Automatically Ordering Events and Times in Text*. Springer (cit. on pp. 11, 12, 16, 17).

Dligach, Dmitriy, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova (Apr. 2017). "Neural Temporal Relation Extraction". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Valencia, Spain, Apr. 3, 2017–Apr. 7, 2017). Association for Computational Linguistics, pp. 746–751 (cit. on pp. 3, 37, 62, 160).

Dowty, David (1979). *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Springer (cit. on p. 12).

Durrett, Greg and Dan Klein (Oct. 2013). "Easy Victories and Uphill Battles in Coreference Resolution". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, Washington, USA, Oct. 18, 2013–Oct. 21, 2013). Association for Computational Linguistics, pp. 1971–1982 (cit. on pp. 78, 83, 106, 110).

Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database* (cit. on p. 86).

Fernandes, Eraldo Rezende, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú (July 2012). "Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island, Korea, July 12, 2012–July 14, 2012). Association for Computational Linguistics, pp. 41–48 (cit. on p. 85).

Fernandes, Eraldo Rezende, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú (2014). "Latent Trees for Coreference Resolution". In: *Computational Linguistics* 40.4, pp. 801–835 (cit. on pp. 85, 86).

Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson (2005). *TIDES–2005 Standard for the Annotation of Temporal Expressions*. Tech. rep. MITRE (cit. on pp. 11, 18).

Finley, Thomas and Thorsten Joachims (Aug. 2005). "Supervised Clustering with Support Vector Machines". In: *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany, Aug. 7, 2005–Aug. 11, 2005), pp. 217–224 (cit. on p. 84).

Fisher, David, Stephen Soderland, Joseph McCarthy, Fangfang Feng, and Wendy Lehnert (1995). "Description of the UMass System as used for MUC-6". In: *Proceedings of the 6th Message Understanding Conference*. Morgan Kaufmann, pp. 127–140 (cit. on pp. 80, 83).

Forascu, Corina and Dan Tufiș (May 2012). "Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (Istanbul, Turkey, May 21, 2012–May 27, 2012). European Language Resources Association (cit. on p. 29).

Freksa, Christian (1992). "Temporal reasoning based on semi-intervals". In: *Artificial Intelligence* 54.1, pp. 199–227 (cit. on p. 17).

Galescu, Lucian and Nate Blaylock (Jan. 2012). "A Corpus of Clinical Narratives Annotated with Temporal Information". In: *Proceedings of the 2nd ACM/SIGHIT International Health*

Informatics Symposium (Miami, Florida, USA, Jan. 28, 2012–Jan. 30, 2012). Association for Computing Machinery, pp. 715–720 (cit. on pp. 19, 22, 31).

Galton, Antony (1990). "A Critical Examination of Allen's Theory of Action and Time". In: *Artificial Intelligence*, pp. 159–188 (cit. on p. 16).

Garnham, Alan (2001). *Mental Models and the Interpretation of Anaphora*. Psychology Press (cit. on pp. 73, 74).

Gers, Felix A., Nicol N. Schraudolph, and Jürgen Schmidhuber (2002). "Learning Precise Timing with LSTM Recurrent Networks". In: *Journal of Machine Learning Research* 3, pp. 115–143 (cit. on p. 113).

Grishman, Ralph and Beth M. Sundheim (1996). "Message Understanding Conference–6: A Brief History". In: *Proceedings of the 17st International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics* (Montreal, Quebec, Canada, Aug. 10, 1998–Aug. 14, 1998). Association for Computational Linguistics, pp. 466–471 (cit. on p. 20).

Grosz, Barbara J. and Candace L. Sidner (1986). "Attention, Intentions and the Structure of Discourse". In: *Computational Linguistics* 12.3, pp. 175–204 (cit. on p. 79).

Grosz, Barbara J., Scott Weinstein, and Aravind K. Joshi (1995). "Centering: A Framework for Modeling the Local Coherence of Discourse". In: *Computational Linguistics* 21.2, pp. 203–225 (cit. on p. 79).

Grouin, Cyril, Marco Dinarelli, Sophie Rosset, Guillaume Wisniewski, and Pierre Zweigenbaum (Oct. 2011). "Coreference Resolution in Clinical Reports. The LIMSI Participation in the i2b2/VA 2011 Challenge". In: *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* (Washington, DC, USA) (cit. on pp. 78, 86, 121).

Grouin, Cyril, Natalia Grabar, Thierry Hamon, Sophie Rosset, Xavier Tannier, and Pierre Zweigenbaum (2013). "Eventual Situations for Timeline Extraction from Clinical Reports". In: *Journal of the American Medical Informatics Association* 20.5, pp. 820–827 (cit. on p. 37).

Grouin, Cyril and Aurélie Névéol (2014). "De-Identification of Clinical Notes in French: Towards a Protocol for Reference Corpus Developpement". In: *Journal of Biomedical Informatics* 50, pp. 151–61 (cit. on pp. 2, 32, 159).

Hagège, Caroline and Xavier Tannier (June 2007). "XRCE-T: XIP Temporal Module for TempEval campaign". In: *Proceedings of the 4th International Workshop on Semantic Evaluation* (Prague, Czech Republic, June 23, 2007–June 24, 2007). Association for Computational Linguistics, pp. 492–495 (cit. on p. 36).

Harkema, Henk, Andrea Setzer, Rob Gaizauskas, and Mark Hepple (Sept. 2005). "Mining and Modelling Temporal Clinical Data". In: *Proceedings of the 4th UK e-Science All Hands Meeting* (Notthingham, England, Sept. 19, 2005–Sept. 22, 2005) (cit. on p. 33).

Hepple, Mark, Andrea Setzer, and Robert Gaizauskas (June 2007). "USFD: Preliminary Exploration of Features and Classifiers for the TempEval-2007 Task". In: *Proceedings of the 4th*

International Workshop on Semantic Evaluation (Prague, Czech Republic, June 23, 2007–June 24, 2007). Association for Computational Linguistics, pp. 438–441 (cit. on p. 36).

Hinote, David, Carlos Ramirez, and Ping Chen (Oct. 2011). "A Comparative Study of Coreference Resolution in Clinical Text". In: *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* (Washington, DC, USA) (cit. on p. 121).

Hirschman, Lynette and Nancy A. Chinchor (1998). "MUC-7 Coreference Task Definition". In: *Proceedings of the 7th Message Understanding Conference*. Morgan Kaufmann (cit. on pp. 3, 74, 79, 87, 161).

Hobbs, Jerry R. (1976). *Pronoun Resolution*. Research rep. 76-1. Department of Computer Sciences, City College, City University of New York (cit. on p. 79).

Hobbs, Jerry R. (1978). "Resolving Pronoun References". In: *Lingua* 44, pp. 311–338 (cit. on p. 79).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780 (cit. on p. 56).

Hoste, Veronique (2005). "Optimization Issues in Machine Learning of Coreference Resolution". PhD thesis. Universiteit Antwerpen (cit. on p. 81).

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel (June 2006). "OntoNotes: The 90% Solution". In: *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (New York, New York, USA, June 4, 2006–June 9, 2006). Association for Computational Linguistics, pp. 57–60 (cit. on p. 72).

Hripcsak, George, Nicholas D. Soulakis, Li Li, Frances P. Morrison, Albert M. Lai, Carol Friedman, Neil S. Calman, and Farzad Mostashari (2009). "Syndromic Surveillance Using Ambulatory Electronic Health Records". In: *Journal of the American Medical Informatics Association* 16.3, pp. 354–361 (cit. on p. 22).

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). "Bidirectional LSTM-CRF Models for Sequence Tagging". In: *Computing Research Repository* (cit. on p. 56).

Jindal, Prateek and Dan Roth (Oct. 2011). "Using Domain Knowledge and Domain-Inspired Discourse Model for Coreference Resolution in Clinical Narratives". In: *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* (Washington, DC, USA) (cit. on p. 121).

Jindal, Prateek and Dan Roth (2013). "Using Domain Knowledge and Domain-inspired Discourse Model for Coreference Resolution for Clinical Narratives". In: *Journal of the American Medical Association*, pp. 256–362 (cit. on p. 86).

Jixin, Ma and Brian Knight (1994). "A General Temporal Theory". In: *The Computer Journal* (cit. on p. 16).

Joachims, Thorsten, Thomas Finley, and Chun-Nam John Yu (2009). "Cutting-Plane Training of Structural SVMs". In: *Machine Learning* 77.1, pp. 27–59 (cit. on p. 35).

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark (2016). "MIMIC-III, a Freely Accessible Critical Care Database". In: *Scientific Data* 3 (cit. on pp. 44, 65).

Jung, Hyuckchul and Amanda Stent (June 2013). "ATT1: Temporal Annotation Using Big Windows and Rich Syntactic and Semantic Features". In: *Proceedings of the 7th International Workshop on Semantic Evaluation* (Atlanta, Georgia, USA, June 14, 2013–June 15, 2013). Association for Computational Linguistics, pp. 20–24 (cit. on pp. 34, 35).

Kamp, Hans and Uwe Reyle (1993). *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Springer (cit. on p. 71).

Karttunen, Lauri (1976). "Discourse Referents". In: *Syntax and Semantics*. Ed. by James D. McCawley. Vol. 7 (cit. on pp. 73, 74).

Katz, Graham and Fabrizio Arosio (2001). "The Annotation of Temporal Information in Natural Language Sentences". In: *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing* (Toulouse, France). Association for Computational Linguistics, pp. 104–111 (cit. on p. 22).

Kehler, Andrew, Douglas Appelt, Lara Taylor, and Aleksandr Simma (May 2004). "The (Non) Utility of Predicate-Argument Frequencies for Pronoun Interpretation". In: *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Boston, Massachusetts, USA, May 2, 2004–May 7, 2004). Association for Computational Linguistics (cit. on p. 81).

Khalifa, Abdulrahman, Sumithra Velupillai, and Stephane Meystre (June 2016). "UtahBMI at SemEval-2016 Task 12: Extracting Temporal Information from Clinical Text". In: *Proceedings of the 10th International Workshop on Semantic Evaluation* (San Diego, California, USA, June 16, 2016–June 17, 2016). Association for Computational Linguistics, pp. 1256–1262 (cit. on pp. 35–37, 49).

Kim, Jin-Dong, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii (June 2009). "Overview of BioNLP'09 Shared Task on Event Extraction". In: *Proceedings of BioNLP 2009 Workshop* (Boulder, Colorado, USA, June 4, 2009–June 5, 2009). Association for Computational Linguistics, pp. 1–9 (cit. on p. 20).

Kim, Jin-Dong, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii (June 2011). "Overview of BioNLP Shared Task 2011". In: *Proceedings of BioNLP 2011 Workshop* (Portland, Oregon, USA, June 23, 2011–June 24, 2011). Association for Computational Linguistics, pp. 1–6 (cit. on p. 20).

Kim, Jin-Dong, Yue Wang, Nicola Colic, Seung Han Beak, Yong Hwan Kim, and Min Song (Aug. 2016). "Refactoring the Genia Event Extraction Shared Task Toward a General Framework for IE-Driven KB Development". In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* (Berlin, Germany, Aug. 12, 2016). Association for Computational Linguistics, pp. 23–31 (cit. on p. 20).

Kim, Jin-Dong, Yue Wang, and Yamamoto Yasunori (Aug. 2013). "The Genia Event Extraction Shared Task, 2013 Edition: Overview". In: *Proceedings of the 2013 BioNLP Workshop* (Sofia, Bulgaria, Aug. 8, 2013–Aug. 9, 2013). Association for Computational Linguistics, pp. 8–15 (cit. on p. 20).

Kingma, Diederik P. and Jimmy Ba (May 2015). "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference on Learning Representations* (San Diego, California, USA, May 7–9, 2015) (cit. on p. 66).

Klenner, Manfred (Sept. 2007). "Enforcing Consistency on Coreference Sets". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (Hissar, Bulgaria, Sept. 7–9, 2015). Association for Computational Linguistics, pp. 323–328 (cit. on p. 81).

Kolomiyets, Oleksandr and Marie-Francine Moens (June 2013). "KUL: Data-driven Approach to Temporal Parsing of Newswire Articles". In: *Proceedings of the 7th International Workshop on Semantic Evaluation* (Atlanta, Georgia, USA, June 14, 2013–June 15, 2013). Association for Computational Linguistics, pp. 83–87 (cit. on p. 35).

Kolya, Anup Kumar, Amitava Kundu, Rajdeep Gupta, Asif Ekbal, and Sivaji Bandyopadhyay (June 2013). "JU_CSE: A CRF Based Approach to Annotation of Temporal Expression, Event and Temporal Relations". In: *Proceedings of the 7th International Workshop on Semantic Evaluation* (Atlanta, Georgia, USA, June 14, 2013–June 15, 2013). Association for Computational Linguistics, pp. 64–72 (cit. on p. 35).

Kovačević, Aleksandar, Azad Dehghan, Michele Filannino, John A. Keane, and Goran Nenadic (2013). "Combining Rules and Machine Learning for Extraction of Temporal Expressions and Events from Clinical Narratives". In: *Journal of the American Medical Informatics Association* 20.5, pp. 859–866 (cit. on p. 36).

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (June 2016). "Neural Architectures for Named Entity Recognition". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California, USA, June 12, 2016–June 17, 2016). Association for Computational Linguistics, pp. 260–270 (cit. on pp. 56, 59, 107, 118).

Landragin, Frédéric, Antonella De Angeli, Frédéric Wolff, Patrice Lopez, and Laurent Romary (2002). "Relevance and Perceptual Constraints in Multimodal Referring Actions". In: *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. Ed. by Kees van Deemter and Rodger Kibble, pp. 395–413 (cit. on p. 71).

Lappin, Shalom and Herbert J. Leass (1994). "An Algorithm for Pronominal Anaphora Resolution". In: *Computational Linguistics* 20.4, pp. 535–561 (cit. on p. 79).

Lassalle, Emmanuel and Pascal Denis (Jan. 2015). "Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures". In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (Austin, Texas, USA, Jan. 25, 2015–Jan. 30, 2015). Association for the Advancement of Artificial Intelligence, pp. 2274–2280 (cit. on pp. 85, 86).

Lee, Hee-Jin, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu (June 2016). "UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes". In: *Proceedings of the 10th International Workshop on Semantic Evaluation* (San Diego, California, USA, June 16, 2016–June 17, 2016). Association for Computational Linguistics, pp. 1292–1297 (cit. on pp. 35–37, 49, 61, 68, 103).

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (June 2011). "Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task". In: *Proceedings of the 15th Conference on Computational Natural Language Learning* (Portland, Oregon, USA, June 23, 2011–June 24, 2011). Association for Computational Linguistics, pp. 28–34 (cit. on p. 78).

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (Sept. 2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark, Sept. 7, 2017–Sept. 11, 2017). Association for Computational Linguistics, pp. 188–197 (cit. on pp. 4, 93, 106, 108, 110, 122, 124, 161).

Leeuwenberg, Artuur and Marie-Francine Moens (Aug. 2017a). "KULeuven-LIIR at SemEval-2017 Task 12: Cross-Domain Temporal Information Extraction from Clinical Records". In: *Proceedings of the 11th International Workshop on Semantic Evaluation* (Vancouver, Canada, Aug. 3, 2017–Aug. 4, 2017). Association for Computational Linguistics, pp. 1030–1034 (cit. on p. 36).

Leeuwenberg, Artuur and Marie-Francine Moens (Apr. 2017b). "Structured Learning for Temporal Relation Extraction from Clinical Records". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Valencia, Spain, Apr. 3, 2017–Apr. 7, 2017). Association for Computational Linguistics, pp. 1150–1158 (cit. on pp. 61, 124).

Lin, Chen, Dmitriy Dligach, Timothy A. Miller, Steven Bethard, and Guergana K. Savova (2016a). "Multilayered Temporal Modeling for the Clinical Domain". In: *Journal of the American Medical Informatics Association* 23.2, pp. 387–395 (cit. on pp. 2, 159).

Lin, Chen, Timothy Miller, Dimitry Dligach, Steven Bethard, and Guergana Savova (Aug. 2016b). "Improving Temporal Relation Extraction with Training Instance Augmentation". In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* (Berlin, Germany, Aug. 12, 2016). Association for Computational Linguistics, pp. 108–113 (cit. on p. 61).

Lin, Yu-Kai, Hsinchun Chen, and Randall A. Brown (2013). "MedTime: A Temporal Information Extraction System for Clinical Narratives". In: *Journal of Biomedical Informatics* 46, pp. 20–28 (cit. on pp. 34, 36).

Linguistic Data Consortium (2005). "The ACE 2005 (ACE05) Evaluation Plan". In: (cit. on p. 20).

Liu, Haibin, Tom Christiansen, William A. Baumgartner, and Karin Verspoor (2012). "BioLemmatizer: A Lemmatization Tool for Morphological Processing of Biomedical Text". In: *Journal of Biomedical Semantics* 3.1 (cit. on p. 45).

Llorens, Hector, Leon Derczynski, Robert Gaizauskas, and Estela Saquete (May 2012). "TIMEN: An Open Temporal Expression Normalisation Resource". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (Istanbul, Turkey, May 21, 2012–May 27, 2012). European Language Resources Association, pp. 3044–3051 (cit. on p. 34).

Lu, Jing and Vincent Ng (July 2018). "Event Coreference Resolution: A Survey of Two Decades of Research". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm, Sweden, July 13, 2018–July 19, 2018), pp. 5479–5486 (cit. on pp. 4, 76, 77, 80, 161).

Luo, Xiaoqiang (Oct. 2005). "On Coreference Resolution Performance Metrics". In: *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing* (Oct. 6, 2005–Oct. 8, 2005). Association for Computational Linguistics, pp. 25–32 (cit. on pp. 89, 90).

Luo, Xiaoqiang, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos (July 2004). "A Mention-Synchronous Coreference Resolution Algorithm Based On the Bell Tree". In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (Barcelona, Spain, July 21, 2004–July 26, 2004). Association for Computational Linguistics, pp. 136–143 (cit. on pp. 83–85).

Ma, Chao, Janardhan Rao Doppa, J. Walker Orr, Prashanth Mannem, Xiaoli Fern, Tom Dietterich, and Prasad Tadepalli (Oct. 2014). "Prune-and-Score: Learning for Greedy Coreference Resolution". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Doha, Qatar, Oct. 25, 2014–Oct. 29, 2014). Association for Computational Linguistics, pp. 2115–2126 (cit. on p. 84).

Ma, Xuezhe and Eduard Hovy (Aug. 2016). "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Berlin, Germany, Aug. 7, 2016–Aug. 12, 2016). Association for Computational Linguistics, pp. 1064–1074 (cit. on pp. 56, 118).

MacAvaney, Sean, Arman Cohan, and Nazli Goharian (Aug. 2017). "GUIR at SemEval-2017 Task 12: A Framework for Cross-Domain Clinical Temporal Information Extraction". In: *Proceedings of the 11th International Workshop on Semantic Evaluation* (Vancouver, Canada, Aug. 3, 2017–Aug. 4, 2017). Association for Computational Linguistics, pp. 1024–1029 (cit. on p. 68).

Mani, Inderjeet, Barry Schiffman, and Jianping Zhang (May 2003). "Inferring Temporal Ordering of Events in News". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Edmonton, Canada, May 27, 2003–June 1, 2003). Association for Computational Linguistics (cit. on p. 36).

Mani, Inderjeet, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky (July 2006). "Machine Learning of Temporal Relations". In: *Proceedings of the 21st International*

Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (Sydney, Australia, July 17, 2006–July 21, 2006). Association for Computational Linguistics, pp. 753–760 (cit. on p. 36).

Mani, Inderjeet and George Wilson (Oct. 2000). "Robust Temporal Processing of News". In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (Hong Kong, Oct. 3, 2000–Oct. 6, 2000). Association for Computational Linguistics (cit. on pp. 11, 34).

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (June 2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, Maryland, USA, June 22, 2014–June 27, 2014). Association for Computational Linguistics, pp. 55–60 (cit. on pp. 34, 50).

March, Olivia and Timothy Baldwin (Dec. 2008). "Automatic Event Reference Identification". In: *Proceedings of the 2008 Australasian Language Technology Association Workshop* (Hobart, Australia, Dec. 8, 2018–Dec. 10, 2018). Australasian Language Technology Association, pp. 79–87 (cit. on p. 35).

Marşic, Georgiana (2011). "Temporal Processing of News: Annotation of Temporal Expressions, Verbal Events and Temporal Relations". PhD thesis. University of Wolverhampton (cit. on pp. 11–15).

Martschat, Sebastian (2017). "Structured Representations for Coreference Resolution". PhD thesis. Heidelberg University (cit. on p. 80).

McCallum, Andrew and Ben Wellner (Dec. 2005). "Conditional Models of Identity Uncertainty with Application to Noun Coreference". In: *Proceedings of the 2005 Neural Information Processing Systems Conference* (Vancouver, British Columbia, Canada, Dec. 5, 2005–Dec. 8, 2005), pp. 905–912 (cit. on p. 81).

McCarthy, Joseph F. and Wendy G. Lehnert (Aug. 1995). "Using Decision Trees for Coreference Resolution". In: *Proceedings of the 1995 International Joint Conference on Artificial Intelligence* (Montreal, Quebec, Aug. 20, 1995–Aug. 25, 1995), pp. 1050–1055 (cit. on pp. 80–82).

McClosky, David (2010). "Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing". PhD thesis. Brown University (cit. on p. 45).

McCray, Alexa T., Anita Burgun, and Olivier Bodenreider (2001). "Aggregating UMLS Semantic Types for Reducing Conceptual Complexity". In: *Studies in Health Technology and Informatics* 1, pp. 216–220 (cit. on p. 40).

Meystre, Stephane M., Guergana K. Savova, Karin C. Kipper-Schuler, and John F. Hurdle (2008). "Extracting Information from Textual Documents in the Electronic Health Record: a Review of Recent Research". In: *Yearbook Of Medical Informatics*, pp. 128–144 (cit. on pp. 1, 2, 158, 159).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *Computing Research Repository (arXiv)* (cit. on p. 44).

Miller, Timothy, Steven Bethard, Hadi Amiri, and Guergana Savova (Aug. 2017a). "Unsupervised Domain Adaptation for Clinical Negation Detection". In: *Proceedings of the 2017 BioNLP Workshop* (Vancouver, Canada, Aug. 4, 2017). Association for Computational Linguistics, pp. 165–170 (cit. on p. 64).

Miller, Timothy, Dmitriy Dligach, Steven Bethard, Chen Lin, and Guergana Savova (2017b). "Towards Generalizable Entity-Centric Clinical Coreference Resolution". In: *Journal of Biomedical Informatics* 69, pp. 251–258 (cit. on p. 78).

Mitamura, Teruko, Zhengzhong Liu, and Eduard Hovy (Nov. 2015). "Overview of TAC-KBP 2015 Event Nugget Track". In: *Proceedings of the 2015 Text Analysis Conference* (Gaithersburg, Maryland, USA, Nov. 16–17, 2015) (cit. on p. 20).

Miwa, Makoto and Sophia Ananiadou (2015). "Adaptable, High Recall, Event Extraction System with Minimal Configuration". In: *BMC Bioinformatics* 16.10 (cit. on p. 64).

Miwa, Makoto and Mohit Bansal (Aug. 2016). "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Berlin, Germany, Aug. 7, 2016–Aug. 12, 2016). Association for Computational Linguistics, pp. 1105–1116 (cit. on p. 124).

Moriceau, Véronique and Xavier Tannier (May 2014). "French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation* (Reykjavik, Iceland, May 26, 2014–May 31, 2014). European Language Resources Association, pp. 3239–3243 (cit. on p. 34).

Mowery, Danielle, Henk Harkema, and Wendy Chapman (June 2008). "Temporal Annotation of Clinical Text". In: *Proceedings of BioNLP 2008 Workshop* (Columbus, Ohio, USA, June 19, 2008). Association for Computational Linguistics, pp. 106–107 (cit. on p. 33).

Neamatullah, Ishna, Margaret M. Douglass, H. Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford (2008). "Automated De-identification of Free-text Medical Records". In: *BMC Medical Informatics and Decision Making* 8.1 (cit. on pp. 2, 159).

Negri, Matteo and Luca Marseglia (2005). *Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004*. Research rep. Information Society Technologies (cit. on p. 34).

Névéol, Aurélie, Kevin B. Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Grégoire Rey, Aude Robert, Xavier Tannier, and Pierre Zweigenbaum (2016). "Clinical Information Extraction at the CLEF eHealth Evaluation lab 2016". In: *CLEF eHealth Evaluation Lab* (cit. on pp. 1, 158).

Névéol, Aurélie, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum (2018). "Clinical Natural Language Processing in Languages other than English: Opportunities and Challenges". In: *Journal of Biomedical Semantics* 9.1 (cit. on p. 40).

Ng, Vincent (July 2010). "Supervised Noun Phrase Coreference Research: The First Fifteen Years". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Uppsala, Sweden, July 11, 2010–July 16, 2010). Association for Computational Linguistics, pp. 1396–1411 (cit. on p. 80).

Ng, Vincent (2016). "Advanced Machine Learning Models for Coreference Resolution". In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuckardt, and Yannick Versley. Springer, pp. 283–313 (cit. on pp. 80, 83).

Ng, Vincent (Feb. 2017). "Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research". In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (San Francisco, California, USA, Feb. 4, 2017–Feb. 9, 2017). Association for the Advancement of Artificial Intelligence, pp. 4877–4884 (cit. on pp. 80, 88).

Ng, Vincent and Claire Cardie (July 2002a). "Combining Sample Selection and Error-Driven Pruning for Machine Learning of Coreference Rules". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (Philadelphia, Pennsylvania, USA, July 6, 2002–July 7, 2002). Association for Computational Lingusitics, pp. 55–62 (cit. on p. 81).

Ng, Vincent and Claire Cardie (Aug. 2002b). "Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution". In: *Proceedings of the 19th International Conference on Computational Linguistics* (Taipei, Taiwan, Aug. 24, 2002–Sept. 1, 2002). Association for Computational Linguistics (cit. on p. 81).

Ng, Vincent and Claire Cardie (June 2002c). "Improving Machine Learning Approaches to Coreference Resolution". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, Pennsylvania, USA, June 6, 2002–June 12, 2002). Association for Computational Linguistics, pp. 104–111 (cit. on pp. 81–83).

Nicolae, Cristina and Gabriel Nicolae (July 2006). "BestCut: A Graph Algorithm for Coreference Resolution". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, Australia, July 22, 2006–July 23, 2006). Association for Computational Linguistics, pp. 275–283 (cit. on p. 81).

Nikfarjam, Azadeh, Ehsan Emadzadeh, and Graciela Gonzalez (2013). "Towards Generating a Patient's Timeline: Extracting Temporal Relationships from Clinical Notes". In: *Journal of Biomedical Informatics* 46, pp. 54–62 (cit. on p. 37).

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer (Dec. 2017). "Automatic Differentiation in PyTorch". In: *Proceedings of the NIPS 2017 Autodiff Workshop* (Long Beach, California, USA, Dec. 9, 2017) (cit. on p. 113).

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 45).

Peng, Nanyun, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih (2017). "Cross-Sentence N-ary Relation Extraction with Graph LSTMs". In: *Transactions of the Association for Computational Linguistics* 5, pp. 101–115 (cit. on p. 125).

Penz, Janet F. E., Adam B. Wilcox, and John F. Hurdle (2007). "Automated Identification of Adverse Events Related to Central Venous Catheters". In: *Journal of Biomedical Informatics* 40.2, pp. 174–182 (cit. on pp. 1, 158).

Pestian, John P., Christopher Brew, Paweł Matykiewicz, D. J. Hovermale, Neil Johnson, Kevin B. Cohen, and Włodzisław Duch (June 2007). "A Shared Task Involving Multi-label Classification of Clinical Free Text". In: *Proceedings of BioNLP 2007 Workshop* (Prague, Czech Republic, June 29, 2007). Association for Computational Linguistics, pp. 97–104 (cit. on pp. 1, 158).

Poesio, Massimo (2016). "Linguistic and Cognitive Evidence About Anaphora". In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuckardt, and Yannick Versley. Springer, pp. 55–94 (cit. on pp. 71–74).

Poesio, Massimo, Roland Stuckardt, and Yannick Versley, eds. (2016a). *Anaphora Resolution: Algorithms, Resources, and Applications*. Springer (cit. on pp. 80, 83).

Poesio, Massimo, Roland Stuckardt, Yannick Versley, and Renata Vieira (2016b). "Early Approaches to Anaphora Resolution: Theoretically Inspired and Heuristic based". In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuckardt, and Yannick Versley. Springer, pp. 23–54 (cit. on p. 79).

Poesio, Massimo and Renata Vieira (1998). "A Corpus-based Investigation of Definite Description Use". In: *Computational Linguistics* 24.2 (cit. on p. 74).

Ponzetto, Simone Paolo and Michael Strube (June 2006). "Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution". In: *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (New York, New York, USA, June 4, 2006–June 9, 2006). Association for Computational Linguistics, pp. 192–199 (cit. on pp. 81, 82).

Pradhan, Sameer, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube (June 2014). "Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, Maryland, USA, June 22, 2014–June 27, 2014). Association for Computational Linguistics, pp. 30–35 (cit. on pp. 87, 89, 90, 92, 96).

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (July 2012). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island, Korea, July 12, 2012–July 14, 2012). Association for Computational Linguistics, pp. 1–40 (cit. on pp. 72, 85, 91, 118).

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue (June 2011). "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes". In: *Proceedings of the 15th Conference on Computational Natural Language*

Learning (Portland, Oregon, USA, June 23, 2011–June 24, 2011). Association for Computational Linguistics, pp. 1–27 (cit. on pp. 72, 75, 83, 91, 118).

Pradhan, Sameer, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla (Sept. 2007). "Unrestricted Coreference: Identifying Entities and Events in OntoNotes". In: *Proceedings of the First IEEE International Conference on Semantic Computing* (Irvine, California, USA). IEEE, pp. 446–453 (cit. on pp. 72, 75).

Pustejovsky, James (1995). *The Generative Lexicon*. MIT Press (cit. on p. 12).

Pustejovsky, James, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo (2003). "The TIMEBANK Corpus". In: *Corpus Linguistics* (cit. on pp. 21, 22, 29).

Pustejovsky, James, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani (2005). "The specification language TimeML". In: *The Language of Time: A Reader*, pp. 545–557 (cit. on pp. 3, 18, 21, 22, 24, 160).

Pustejovsky, James, Kiyong Lee, Harry Bunt, and Laurent Romary (May 2010). "ISO-TimeML: An International Standard for Semantic Annotation". In: *Proceedings of the 7th International Conference on Language Resources and Evaluation* (Valletta, Malta, May 17, 2010–May 23, 2010). European Language Resources Association, pp. 394–397 (cit. on pp. 18, 21–24, 26, 33).

Pustejovsky, James and Amber Stubbs (June 2011). "Increasing Informativeness in Temporal Annotation". In: *Proceedings of the 5th Linguistic Annotation Workshop* (Portland, Oregon, USA), pp. 152–160 (cit. on p. 23).

Quinlan, J. Ross (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc. (cit. on pp. 80–82).

Rahman, Altaf and Vincent Ng (Aug. 2009). "Supervised Models for Coreference Resolution". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processin* (Singapore, Aug. 6, 2009–Aug. 7, 2009). Association for Computational Linguistics, pp. 968–977 (cit. on pp. 78, 81, 82, 84, 88, 89).

Rahman, Altaf and Vincent Ng (June 2011a). "Coreference Resolution with World Knowledge". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 19, 2011–June 24, 2011). Association for Computational Linguistics, pp. 814–824 (cit. on p. 81).

Rahman, Altaf and Vincent Ng (2011b). "Narrowing the Modeling Gap: A Cluster-Ranking Approach to Coreference Resolution". In: *Journal of Artificial Intelligence Research* 40, pp. 469–521 (cit. on pp. 81, 82, 84, 85).

Recasens, Marta and Eduard Hovy (Nov. 2009). "A Deeper Look into Features for Coreference Resolution". In: *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium* (Goa, India). Springer, pp. 29–42 (cit. on p. 81).

Recasens, Marta and Eduard Hovy (2011). "BLANC: Implementing the Rand Index for Coreference Evaluation". In: *Natural Language Engineering* 17.4, pp. 485–510 (cit. on pp. 89, 90).

Recasens, Marta, Eduard Hovy, and M. Antònia Martí (2011). "Identity, Non-Identity, and Near-Identity: Addressing the Complexity of Coreference". In: *Lingua* 121, pp. 1138–1152 (cit. on p. 73).

Řehůřek, Radim and Petr Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the 7th International Conference on Language Resources and Evaluation* (Valletta, Malta, May 17, 2010–May 23, 2010). European Language Resources Association, pp. 45–50 (cit. on p. 44).

Rei, Marek (July 2017). "Semi-supervised Multitask Learning for Sequence Labeling". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vancouver, Canada, July 30, 2017–Aug. 4, 2017). Association for Computational Linguistics, pp. 2121–2130 (cit. on p. 110).

Reichenbach, Hans (1947). *Elements of Symbolic Logic*. The Macmillan Company (cit. on p. 16).

Reimers, Nils and Iryna Gurevych (Sept. 2017). "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark, Sept. 7, 2017–Sept. 11, 2017). Association for Computational Linguistics, pp. 338–348 (cit. on pp. 113, 117).

Roberts, Angus, Robert Gaizauskas, George Demetriou, Yikun Guo, Andrea Setzer, and Ian Roberts (2008). "Semantic Annotation of Clinical Text: The CLEF Corpus". In: *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining* (Marrakech, Morocco, May 31, 2014). European Language Resources Association (cit. on p. 29).

Roberts, Angus, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer (2009). "Building a Semantically Annotated Corpus of Clinical Texts". In: *Journal of Biomedical Informatics* 42, pp. 950–966 (cit. on p. 26).

Roberts, Kirk, Bryan Rink, and Sanda M. Harabagiu (2013). "A Flexible Framework for Recognizing Events, Temporal Expressions, and Temporal Relations in Clinical Text". In: *Journal of the American Medical Informatics Association* 20, pp. 867–875 (cit. on pp. 34–36).

Saurí, Roser, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky (2006). *TimeML Annotation Guidelines: Version 1.2.1*. Tech. rep. (cit. on pp. 24, 25).

Savova, Guergana K., Steven Bethard, Will Styler, James Martin, Martha Palmer, James Masanz, and Wayne Ward (Nov. 2009). "Towards Temporal Relation Discovery from the Clinical Narrative". In: *Proceedings of the 2009 AMIA Annual Symposium* (San Francisco, California, USA, Nov. 14, 2009–Nov. 18, 2009). American Medical Informatics Association, pp. 568–572 (cit. on p. 33).

Savova, Guergana K., Wendy W. Chapman, Jiaping Zheng, and Rebecca S. Crowley (2011). "Anaphoric Relations in the Clinical Narrative: Corpus Creation". In: *Journal of the American Medical Informatics Association* 18, pp. 459–465 (cit. on p. 77).

Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute (2010). "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications". In: *Journal of the American Medical Informatics Association* 17, pp. 507–513 (cit. on pp. 36, 50, 78).

Setzer, Andrea (2001). "Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study". PhD thesis. University of Sheffield (cit. on p. 22).

Setzer, Andrea and Robert Gaizauskas (May 2002). "On the Importance of Annotating Event-Event Temporal Relations in Text". In: *Proceedings of the LREC 2002 Workshop on Annotation Standards for Temporal Information in Natural Language* (Las Palmas, Spain, May 27, 2002). European Language Resources Association, pp. 52–60 (cit. on p. 22).

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases". In: *Computational Linguistics* 27.4, pp. 521–544 (cit. on pp. 80–84).

Stoyanov, Veselin and Jason Eisner (Dec. 2012). "Easy-First Coreference Resolution". In: *Proceedings of the 24th International Conference on Computational Linguistics* (Mumbai, India, Dec. 8, 2012–Dec. 15, 2012). Association for Computational Linguistics, pp. 2519–2534 (cit. on pp. 84, 85).

Stoyanov, Veselin, Nathan Gilbert, Claire Cardie, and Ellen Riloff (Aug. 2009). "Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (Suntec, Singapore, Aug. 2, 2009–Aug. 7, 2009). Association for Computational Linguistics and Asian Federation of Natural Language Processing, pp. 656–664 (cit. on pp. 81, 82, 88).

Strötgen, Jannik and Michael Gertz (2013). "Multilingual and Cross-domain Temporal Tagging". In: *Language Resources and Evaluation* 47.2, pp. 269–298 (cit. on p. 34).

Strötgen, Jannik and Michael Gertz (Sept. 2015). "A Baseline Temporal Tagger for All Languages". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon, Portugal, Sept. 17, 2015–Sept. 21, 2015). Association for Computational Linguistics, pp. 541–547 (cit. on p. 17).

Strötgen, Jannik and Michael Gertz (2016). *Domain-Sensitive Temporal Tagging*. Synthesis Lectures on Human Language Technologies (cit. on pp. 17, 18).

Strube, Michael (Aug. 1998). "Never Look Back: An Alternative to Centering". In: *Proceedings of the 17st International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics* (Montreal, Quebec, Canada, Aug. 10, 1998–Aug. 14, 1998). Association for Computational Linguistics, pp. 1251–1257 (cit. on p. 79).

Strube, Michael and Udo Hahn (1999). "Functional Centering Grounding Referential Coherence in Information Structure". In: *Computational Linguistics* 25.3, pp. 309–344 (cit. on p. 79).

Stubbs, Amber and Özlem Uzuner (2015). "Annotating Longitudinal Clinical Narratives for De-Identification: The 2014 i2b2/UTHealth Corpus". In: *Journal of Biomedical Informatics* 58, pp. 20–29 (cit. on pp. 2, 159).

Styler IV, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky (2014a). "Temporal Annotation in the Clinical Domain". In: *Transactions of the Association for Computational Linguistics* 2, pp. 143–154 (cit. on pp. 19, 22–24, 26, 28, 32, 40, 64, 78).

Styler IV, William F., Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman, and Piet C. de Groen (2014b). *THYME Annotation Guidelines*. Tech. rep. (cit. on p. 26).

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum (May 2007). "Yago: A Core of Semantic Knowledge". In: *Proceedings of the 16th International World Wide Web Conference* (Banff, Alberta, Canada, May 8, 2007–May 12, 2007). International World Wide Web Conference Committee, pp. 697–706 (cit. on p. 81).

Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner (2013a). "Annotating Temporal Information in Clinical Narratives". In: *Journal of Biomedical Informatics* 46, pp. 5–12 (cit. on pp. 19, 31).

Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner (2013b). "Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge". In: *Journal of the American Medical Informatics Association* 20, pp. 806–813 (cit. on pp. 3, 160).

Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner (2013c). "Temporal Reasoning over Clinical Text: The State of the Art". In: *Journal of the American Medical Informatics Association* 20, pp. 814–819 (cit. on pp. 2, 159).

Sun, Xu, Takuya Matsuzaki, Daisuke Okanohara, and Jun'ichi Tsujii (July 2009). "Latent Variable Perceptron Algorithm for Structured Classification". In: *Proceedings of the 21st International Joint Conference on Artificial Intelligence* (Pasadena, California, USA, July 11, 2009–July 17, 2009), pp. 1236–1242 (cit. on p. 85).

Sundheim, Beth M. (Aug. 1993). "Tipster/MUC-5: Information Extraction System Evaluation". In: *Proceedings of the 5th Message Understanding Conference* (Baltimore, Maryland, USA, Aug. 25, 1993–Aug. 27, 1993). Morgan Kaufmann Publishers, pp. 147–163 (cit. on pp. 3, 17, 160).

Sundheim, Beth M. (1995). "Overview of Results of the MUC-6 Evaluation". In: *Proceedings of the 6th Message Understanding Conference*. Vol. 423–442. Morgan Kaufmann (cit. on pp. 3, 74, 79, 87, 161).

Tapi Nzali, Mike Donald, Xavier Tannier, and Aurelie Névéol (Sept. 2015). "Automatic Extraction of Time Expressions Accross Domains in French Narratives". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon, Portugal,

Sept. 17, 2015–Sept. 21, 2015). Association for Computational Linguistics, pp. 492–498 (cit. on pp. 19, 34).

Tourille, Julien, Matthieu Doutreligne, Olivier Ferret, Nicolas Paris, Aurélie Névéol, and Xavier Tannier (Oct. 2018). "Evaluation of a Sequence Tagging Tool for Biomedical Texts". In: *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis* (Brussels, Belgium, Oct. 31, 2018). Association for Computational Linguistics (cit. on pp. 7, 8, 54, 56, 99).

Tourille, Julien, Olivier Ferret, Aurélie Névéol, and Xavier Tannier (July 2016a). "Extraction de Relations Temporelles dans des Dossiers Électroniques Patient". In: *Actes de la Conférence Traitement Automatique des Langues Naturelles 2016* (Paris, France, July 4, 2016–July 8, 2016). Association pour le Traitement Automatique des Langues, pp. 459–466 (cit. on pp. 7, 39).

Tourille, Julien, Olivier Ferret, Aurélie Névéol, and Xavier Tannier (June 2016b). "LIMSI-COT at SemEval-2016 Task 12: Temporal Relation Identification Using a Pipeline of Classifiers". In: *Proceedings of the 10th International Workshop on Semantic Evaluation* (San Diego, California, USA, June 16, 2016–June 17, 2016). Association for Computational Linguistics, pp. 1136–1142 (cit. on pp. 7, 39).

Tourille, Julien, Olivier Ferret, Xavier Tannier, and Aurélie Névéol (Aug. 2017a). "LIMSI-COT at SemEval-2017 Task 12: Neural Architecture for Temporal Information Extraction from Clinical Narratives". In: *Proceedings of the 11th International Workshop on Semantic Evaluation* (Vancouver, Canada, Aug. 3, 2017–Aug. 4, 2017). Association for Computational Linguistics, pp. 597–602 (cit. on pp. 7, 8, 54).

Tourille, Julien, Olivier Ferret, Xavier Tannier, and Aurélie Névéol (July 2017b). "Neural Architecture for Temporal Relation Extraction: A Bi-LSTM Approach for Detecting Narrative Containers". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vancouver, Canada, July 30, 2017–Aug. 4, 2017). Association for Computational Linguistics, pp. 224–230 (cit. on pp. 7, 8, 54).

Tourille, Julien, Olivier Ferret, Xavier Tannier, and Aurélie Névéol (Apr. 2017c). "Temporal Information Extraction from Clinical Text". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Valencia, Spain, Apr. 3, 2017–Apr. 7, 2017). Association for Computational Linguistics, pp. 739–745 (cit. on pp. 7, 39).

Uzuner, Özlem (2009). "Recognizing Obesity and Comorbidities in Sparse Data". In: *Journal of the American Medical Informatics Association* 16.4, pp. 561–570 (cit. on pp. 2, 158).

Uzuner, Özlem, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R. South (2012). "Evaluating the State of the Art in Coreference Resolution for Electronic Medical Records". In: *Journal of the American Medical Informatics Association* 19.5, pp. 786–791 (cit. on pp. 4, 77, 86, 87, 94, 96, 121, 161).

Uzuner, Özlem, Ira Goldstein, Yuan Luo, and Isaac Kohane (2008). "Identifying Patient Smoking Status from Medical Discharge Records". In: *Journal of the American Medical Informatics Association* 15.1, pp. 14–24 (cit. on pp. 2, 158).

Uzuner, Özlem, Yuan Luo, and Peter Szolovits (2007). "Evaluating the State-of-the-Art in Automatic De-identification". In: *Journal of the American Medical Informatics Association* 14.5, pp. 550–563 (cit. on pp. 2, 159).

Uzuner, Özlem, Imre Solti, and Eithon Cadag (2010). "Extracting Medication Information from Clinical Text". In: *Journal of the American Medical Informatics Association* 17.5, pp. 514–518 (cit. on pp. 2, 158).

Uzuner, Özlem, Brett R. South, Shuying Shen, and Scott L. DuVall (2011). "2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text". In: *Journal of the American Medical Informatics Association* 18.5, pp. 552–556 (cit. on pp. 2, 31, 77, 158).

UzZaman, Naushad, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky (June 2013). "SemEval-2013 Task 1: TempEval-3 – Evaluating Time Expressions, Events, and Temporal Relations". In: *Proceedings of the 7th International Workshop on Semantic Evaluation* (Atlanta, Georgia, USA, June 14, 2013–June 15, 2013). Association for Computational Linguistics, pp. 1–9 (cit. on pp. 21, 23, 29, 45, 50).

Velupillai, Sumithra, D. Mowery, Brett R. South, Maria Kvist, and Hercules Dalianis (2015a). "Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis". In: *Yearbook of Medical Informatics* 10.1, pp. 183–192 (cit. on pp. 1, 2, 158, 159).

Velupillai, Sumithra, Danielle L. Mowery, Samir Abdelrahman, Lee Christensen, and Wendy Chapman (June 2015b). "BluLab: Temporal Information Extraction for the 2015 Clinical TempEval Challenge". In: *Proceedings of the 9th International Workshop on Semantic Evaluation* (Denver, Colorado, USA, June 4, 2015–June 15, 2015). Association for Computational Linguistics, pp. 815–819 (cit. on pp. 34, 36, 37).

Vendler, Zeno (1967). *Linguistics in Philosophy*. Cornell University Press (cit. on pp. 12–14).

Verhagen, Marc (2004). "Times Between the Lines". PhD thesis. Brandeis University (cit. on p. 17).

Verhagen, Marc, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky (June 2007). "SemEval-2007 Task 15: TempEval Temporal Relation Identification". In: *Proceedings of the 4th International Workshop on Semantic Evaluation* (Prague, Czech Republic, June 23, 2007–June 24, 2007). Association for Computational Linguistics, pp. 75–80 (cit. on pp. 21, 23, 29, 45, 50).

Verhagen, Marc, Roser Saurí, Tommaso Caselli, and James Pustejovsky (July 2010). "SemEval-2010 Task 13: TempEval-2". In: *Proceedings of the 5th International Workshop on Semantic Evaluation* (Uppsala, Sweden, July 15, 2010–July 16, 2010). Association for Computational Linguistics (cit. on pp. 21, 23, 29, 45, 50).

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman (1995). "A Model-Theoretic Coreference Scoring Scheme". In: *Proceedings of the 6th Message Understanding Conference*. Morgan Kaufmann, pp. 45–52 (cit. on p. 87).

Wang, Xiaoyan, George Hripcsak, Marianthi Markatou, and Carol Friedman (2009). "Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Elec-

tronic Health Records: A Feasibility Study". In: *Journal of the American Medical Informatics Association* 16.3, pp. 328–337 (cit. on pp. 1, 2, 158, 159).

Webster, Kellie and James R. Curran (Aug. 2014). "Limited Memory Incremental Coreference Resolution". In: *Proceedings of the 25th International Conference on Computational Linguistics* (Dublin, Ireland, Aug. 23, 2014–Aug. 29, 2014), pp. 2129–2139 (cit. on p. 84).

Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber (June 2016). "Learning Global Features for Coreference Resolution". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California, USA, June 12, 2016–June 17, 2016). Association for Computational Linguistics, pp. 994–1004 (cit. on pp. 84, 92, 93, 106, 108, 110, 112, 114, 117).

Wiseman, Sam, Alexander M. Rush, Stuart Shieber, and Jason Weston (July 2015). "Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution". In: *Proceedings of the Joint Conference of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (Beijing, China, July 26, 2015–July 31, 2015). Association for Computational Linguistics, pp. 1416–1426 (cit. on pp. 78, 83, 106, 112).

Xu, Yan, Jiahua Liu, Jiajun Wu, Yue Wang, Zhuowen Tu, JianTao Sun, Junichi Tsujii, and Eric IChao Chang (2012). "A Classification Approach to Coreference in Discharge Summaries: 2011 i2b2 Challenge". In: *Journal of the American Medical Informatics Association* 19.5, pp. 897–905 (cit. on pp. 81, 86).

Xu, Yan, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric IChao Chang (2013). "An End-to-End System to Identify Temporal Relation in Discharge Summaries: 2012 i2b2 Challenge". In: *Journal of the American Medical Informatics Association* 20.5, pp. 849–858 (cit. on p. 37).

Yang, Xiaofeng, Jian Su, Jun Lang, Chew Lim Tan, Ting Liu, and Sheng Li (June 2008). "An Entity-Mention Model for Coreference Resolution with Inductive Logic Programming". In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Columbus, Ohio, USA, June 15, 2008–June 20, 2008). Association for Computational Linguistics, pp. 843–851 (cit. on p. 84).

Yang, Xiaofeng, Guodong Zhou, Jian Su, and Chew Lim Tan (July 2003). "Coreference Resolution Using Competition Learning Approach". In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (Sapporo, Japan, July 7, 2003–July 12, 2003). Association for Computational Linguistics, pp. 176–183 (cit. on pp. 82, 83).

Yoshikawa, Katsumasa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto (Aug. 2009). "Jointly Identifying Temporal Relations with Markov Logic". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (Suntec, Singapore, Aug. 2, 2009–Aug. 7, 2009). Association for Computational Linguistics and Asian Federation of Natural Language Processing, pp. 405–413 (cit. on p. 38).

Yu, Chun-Nam John and Thorsten Joachims (June 2009). "Learning Structural SVMs with Latent Variables". In: *Proceedings of the 26th International Conference on Machine Learning* (Montreal, Quebec, June 14, 2009–June 18, 2009), pp. 1169–1176 (cit. on pp. 85, 86).

Zhang, Yaoyun, Buzhou Tang, Min Jiang, Jingqi Wang, and Hua Xu (2015). "Domain Adaptation for Semantic Role Labeling of Clinical Text". In: *Journal of the American Medical Informatics Association* 22.5, pp. 967–979 (cit. on p. 64).

Zheng, Jiaping, Wendy W. Chapman, Rebecca S. Crowley, and Guergana K. Savova (2011). "Coreference Resolution: A Review of General Methodologies and Applications in the Clinical Domain". In: *Journal of Biomedical Informatics* 44.6, pp. 1113–1122 (cit. on p. 80).

Appendix A

Clinical TempEval 2016 Results

	To document time			Narrative containers		
	P	R	F1	P	R	F1
Phase 2: systems are provided manually annotated EVENTs and TIMEX3s						
UTHealth-1	-	0.835	-	0.588	**0.559**	**0.573**
UTHealth-2	-	0.833	-	0.568	0.564	0.566
LIMSI-COT-lexical	-	0.769	-	0.704	0.436	0.538
GUIR-1	-	0.813	-	0.546	0.471	0.506
LIMSI-COT-embedding	-	0.807	-	**0.751**	0.320	0.449
KULeuven-LIIR-1	-	-	-	0.714	0.428	0.536
KULeuven-LIIR-2	-	-	-	0.715	0.429	0.536
VUACLTL-2	-	0.701	-	0.589	0.368	0.453
VUACLTL-1	-	0.701	-	0.642	0.345	0.449
UtahBMI-crf+svm	-	**0.843**	-	0.562	0.254	0.350
CDE-IIITH-dl	-	0.705	-	0.348	0.284	0.313
UtahBMI-svm	-	0.571	-	0.605	0.230	0.333
ULISBOA-1	-	-	-	0.273	0.255	0.264
brundlefly	-	0.742	-	-	-	-
uta-5	-	0.788	-	-	-	-
uta-6	-	0.786	-	-	-	-
LIMSI-1	-	0.687	-	-	-	-
CDE-IIITH-crf	-	0.588	-	0.493	0.185	0.269
LIMSI-2	-	0.679	-	-	-	-
ULISBOA-2	-	-	-	0.823	0.056	0.105
Baseline: memorize/closest	-	0.675	-	0.459	0.154	0.231
UtahBMI-crf+svm*	-	0.843	-	0.693	0.425	0.527
UtahBMI-svm*	-	0.571	-	0.711	0.372	0.489
Agreement: ann-ann	-	-	0.721	-	-	0.651
Agreement: adj-ann	-	-	0.844	-	-	0.817

Table A.1: Reproduction of the score table presented in Bethard et al. (2016). System performance and annotator agreement on temporal relation tasks: identifying relations between events and the document creation time (DOCTIMEREL), and identifying narrative container relations (CONTAINS). The best system score from each column is in bold. Systems marked with * were submitted after the competition deadline and are not considered official.

Appendix B

Clinical TempEval 2017 Results

Team	time span			time span + class		
	F1	P	R	F1	P	R
Unsupervised domain adaptation						
GUIR	**0.57**	0.61	0.53	0.51	0.55	0.47
KULeuven-LIIR	0.56	**0.72**	0.46	**0.53**	**0.68**	0.43
LIMSI-COT	0.51	0.42	**0.66**	0.49	0.40	**0.63**
ULISBOA	0.48	0.44	0.54	0.43	0.39	0.48
Hitachi	0.43	0.63	0.33	-	-	-
baseline	0.36	**0.72**	0.24	0.32	0.63	0.21
WuHanNLP	0.31	0.65	0.20	0.27	0.57	0.18
Supervised domain adaptation						
GUIR	**0.59**	0.57	0.62	**0.56**	0.54	0.59
LIMSI-COT	0.58	0.51	**0.67**	0.55	0.49	**0.64**
NTU-1	0.58	**0.58**	0.58	0.54	0.54	0.54
KULeuven-LIIR	0.56	0.57	0.55	0.54	**0.55**	0.53
ULISBOA	0.55	0.52	0.60	0.52	0.48	0.56
UTD	0.54	0.56	0.52	0.44	0.46	0.43
Hitachi	0.51	0.53	0.48	-	-	-
WuHanNLP	0.43	0.45	0.41	0.40	0.42	0.38
XJNLP*	0.41	0.33	0.52	0.35	0.29	0.45
UNICA	0.37	0.31	0.45	0.31	0.26	0.38
baseline	0.35	0.53	0.26	0.32	0.49	0.24
IIIT	0.31	0.39	0.25	0.19	0.24	0.16
Annotator agreement						
ann-ann	0.81	-	-	0.79	-	-
adj-ann	0.86	-	-	0.85	-	-

Table B.1: Reproduction of Table 2 from Bethard et al. (2017). Original caption: "System performance and annotator agreement on TIMEX3 tasks: identifying the time expression's span (character offsets) and class (DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP or SET)".

Team	event span			event span + modality			event span + degree			event span + polarity			event span + type		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Unsupervised domain adaptation															
LIMSI-COT	**0.72**	0.62	**0.84**	**0.64**	0.55	**0.75**	**0.71**	0.62	**0.83**	**0.69**	0.60	**0.82**	**0.70**	0.61	**0.82**
GUIR	0.71	0.64	0.80	0.56	0.50	0.64	0.68	0.61	0.77	0.65	0.59	0.74	0.68	0.61	0.76
KULeuven-LIIR	0.68	**0.70**	0.67	0.62	**0.63**	0.61	0.67	**0.69**	0.66	0.67	**0.68**	0.65	0.66	**0.67**	0.65
ULISBOA	0.68	0.62	0.77	0.61	0.55	0.68	0.68	0.61	0.76	0.66	0.60	0.74	0.66	0.60	0.74
Hitachi	0.68	0.67	0.69	-	-	-	-	-	-	-	-	-	-	-	-
baseline	0.63	0.65	0.61	0.55	0.57	0.54	0.62	0.64	0.60	0.58	0.60	0.56	0.60	0.62	0.59
WuHanNLP	0.62	0.59	0.66	0.55	0.52	0.58	0.61	0.58	0.65	0.6	0.57	0.63	0.60	0.57	0.63
Supervised domain adaptation															
LIMSI-COT	**0.76**	**0.69**	0.85	**0.69**	**0.63**	**0.78**	**0.75**	**0.68**	0.84	**0.75**	**0.68**	0.83	**0.75**	**0.68**	0.83
GUIR	0.74	0.68	0.82	0.66	0.60	0.72	0.73	0.67	0.80	0.58	0.54	0.64	0.72	0.66	0.79
NTU-1	0.73	0.62	**0.87**	0.63	0.54	0.75	0.72	0.62	**0.86**	0.70	0.60	**0.84**	0.70	0.60	**0.85**
ULISBOA	0.73	0.65	0.83	0.64	0.57	0.73	0.72	0.64	0.82	0.71	0.63	0.81	0.71	0.63	0.80
KULeuven-LIIR	0.72	0.67	0.78	0.66	0.61	0.71	0.71	0.66	0.77	0.71	0.66	0.76	0.70	0.65	0.76
Hitachi	0.71	0.67	0.76	-	-	-	-	-	-	-	-	-	-	-	-
baseline	0.70	0.67	0.74	0.62	0.59	0.65	0.69	0.66	0.73	0.66	0.62	0.69	0.68	0.65	0.72
UTD	0.66	0.62	0.71	0.57	0.53	0.61	-	-	-	-	-	-	-	-	-
WuHanNLP	0.65	0.59	0.72	0.58	0.53	0.64	0.64	0.58	0.71	0.63	0.57	0.70	0.63	0.57	0.70
IIIT	0.62	**0.69**	0.56	0.51	0.57	0.47	0.61	0.67	0.55	0.58	0.64	0.52	0.59	0.66	0.54
XJNLP*	0.61	0.55	0.68	0.51	0.46	0.57	0.59	0.54	0.67	0.54	0.49	0.61	0.58	0.52	0.66
UNICA	0.50	0.39	0.71	0.43	0.34	0.61	0.49	0.38	0.70	0.47	0.37	0.66	0.47	0.37	0.67
Annotator agreement															
ann-ann	0.79	-	-	0.72	-	-	0.78	-	-	0.78	-	-	0.76	-	-
adj-ann	0.87	-	-	0.84	-	-	0.86	-	-	0.86	-	-	0.85	-	-

Table B.2: Reproduction of Table 3 from Bethard et al. (2017). Original caption: "System performance and annotator agreement on EVENT tasks: identifying the event expression's span (character offsets), contextual modality (ACTUAL, HYPOTHETICAL, HEDGED or GENERIC), degree (MOST, LITTLE or N/A), polarity (POS or NEG) and type (ASPECTUAL, EVIDENTIAL or N/A)".

	To document time			Narrative containers		
	F1	P	R	F1	P	R
Unsupervised domain adaptation						
LIMSI-COT	**0.51**	0.44	**0.60**	0.33	0.28	**0.40**
KULeuven-LIIR	0.49	**0.50**	0.48	0.32	0.33	0.30
GUIR	0.40	0.36	0.45	**0.34**	**0.52**	0.25
Hitachi	0.45	0.44	0.45	0.23	0.23	0.22
baseline	0.38	0.39	0.37	0.14	0.39	0.08
ULISBOA	0.41	0.37	0.45	-	-	-
WuHanNLP	0.41	0.39	0.43	-	-	-
Supervised domain adaptation						
LIMSI-COT	**0.59**	**0.53**	**0.66**	**0.32**	0.25	**0.43**
KULeuven-LIIR	0.56	0.52	0.61	0.28	0.23	0.35
GUIR	0.50	0.45	0.55	0.25	**0.59**	0.16
NTU-1	0.49	0.42	0.59	0.26	0.20	0.37
Hitachi	0.52	0.49	0.55	0.16	0.11	0.27
baseline	0.46	0.43	0.48	0.14	0.27	0.09
WuHanNLP	0.46	0.42	0.51	0.12	0.16	0.09
UTD	0.45	0.42	0.48	0.11	0.08	0.16
ULISBOA	0.44	0.39	0.51	-	-	-
IIIT	0.36	0.40	0.33	0.05	0.03	0.08
UNICA	0.20	0.15	0.28	-	-	-
Annotator agreement						
ann-ann	0.52	-	-	0.66	-	-
adj-ann	0.71	-	-	0.80	-	-

Table B.3: Reproduction of Table 4 from Bethard et al. (2017). Original caption: "System performance and annotator agreement on temporal relation tasks: identifying relations between events and the document creation time (DOCTIMEREL), and identifying narrative container relations (CONTAINS)".

Appendix C

i2b2 Task 1c Corpus: Conversion to Brat and CoNLL Formats

Column	Type	Description
1	Sentence ID	Sentence ID number in the current document
2	Token	The token itself
3	Begin Offset	Token begin character offset
4	End Offset	Token end character offset
5	i2b2 Offset	Token i2b2 offset (line:token)
6	Coreference	Coreference chain information encoded in a parenthesis structure

Table C.1: CoNLL file: column description.

Figure C.1: i2b2 task1c corpus: brat formatted sentence example.

```
#begin document (clinical-587);
...
14 One 520 523 14:0 -
14 month 524 529 14:1 -
14 prior 530 535 14:2 -
14 to 536 538 14:3 -
14 this 539 543 14:4 (18)
14 admission 544 553 14:5 -
14 ,554 555 14:6 -
14 the 556 559 14:7 (0
14 patient 560 567 14:8 0)
14 reports 568 575 14:9 -
14 not 576 579 14:10 -
14 feeling 580 587 14:11 -
14 well 588 592 14:12 -
14 with 593 597 14:13 -
14 worsening 598 607 14:14 (17
14 gastric 608 615 14:15 -
14 distress 616 624 14:16 17)
14 . 625 626 14:17 -
...
#end document
```

Figure C.2: i2b2 task1c corpus: CoNLL formatted sentence example.

Annexe D

Résumé Étendu

Des informations importantes sont contenues dans les dossiers patients électroniques. La majeure partie de ces informations est localisée dans des textes écrits en langue naturelle. Bien que le texte libre soit pratique pour exprimer des concepts médicaux complexes, il est très difficile de s'en servir dans le cadre d'applications telles que l'aide à la décision, la recherche clinique, l'analyse statistique ou le résumé automatique. Le besoin d'accéder à ces informations combiné à l'adoption massive du dossier patient électronique a favorisé le développement d'approches en Traitement Automatique des Langues (TAL) spécifiques au domaine clinique.

Les méthodes pour l'extraction d'information ont été appliquées avec succès à une variété de tâches durant les dix dernières années. Un exemple typique est l'assignation de codes de diagnostics tels que les codes ICD. Plusieurs jeux de données relatifs au sujet ont été distribués dans la communauté TAL afin de favoriser la recherche sur ce sujet. Par exemple, Pestian et al. (2007) proposent de travailler sur des rapports de radiologie tandis que Névéol et al. (2016) ont distribué un ensemble de certificats de décès annotés avec des codes ICD-10.

Un autre domaine de recherche actif concerne l'enrichissement sémantique des dossiers électroniques pour l'aide à la décision. Meystre et al. (2008) identifient plusieurs sujets : structuration automatique de documents qui consiste à segmenter le texte libre en sections selon un schéma défini préalablement, le résumé automatique qui permet d'avoir une vue concise des documents cliniques et la recherche de cas qui permet de chercher des patients répondant à des critères cliniques spécifiques. L'initiative i2b2 a mené plusieurs campagnes d'annotation portant sur le tabagisme (Uzuner et al. 2008), l'obésité et les comorbidités (Uzuner 2009), les médicaments (Uzuner et al. 2010) et les concepts, assertions et relations (Uzuner et al. 2011).

La surveillance est aussi un champ de recherche important. Un exemple d'application est la détection d'événements indésirables (Velupillai et al. 2015a). Ces événements peuvent être relatifs à des procédures médicales (Penz et al. 2007) ou à la prise de médicaments (Wang et al. 2009). Un autre cas d'utilisation concerne la surveillance des dossiers patients dans le

but de repérer des épidémies (Meystre et al. 2008) ou des infections nosocomiales (Velupillai et al. 2015a).

Tous ces projets de recherche n'auraient pas été possibles sans un accès aux données. Dans la mesure où les dossiers patients contiennent des informations personnelles confidentielles, la dé-identification des documents qu'ils contiennent est un prérequis à toute recherche. La dé-identification automatique des documents cliniques est un domaine de recherche actif qui fait l'objet de nombreux travaux de recherche (Grouin et Névéol 2014 ; Neamatullah et al. 2008 ; Stubbs et Uzuner 2015 ; Uzuner et al. 2007).

Établir une liste exhaustive des méthodes en TAL utilisées dans le domaine clinique est une tâche difficile. Nous renvoyons les lecteurs aux différentes revues de la littérature menées sur le sujet (Meystre et al. 2008 ; Velupillai et al. 2015a ; Wang et al. 2009).

Parmi toutes les informations médicales qui présentent un intérêt dans les dossiers patients, la chronologie médicale (Figure D.1) est une des plus importantes. Être en mesure d'extraire ces chronologies permettrait d'acquérir une meilleure compréhension de certains phénomènes cliniques tels que le déroulement des maladies ou l'effet longitudinal des médicaments (C. Lin et al. 2016a ; W. Sun et al. 2013c). De plus, cela permettrait d'améliorer les systèmes de question–réponse et de prédiction de résultats cliniques. Par ailleurs, accéder aux chronologies médicales est nécessaire pour évaluer la qualité du parcours de soins en le comparant aux recommandations officielles et pour mettre en lumière les étapes du parcours de soins auxquelles une attention particulière doit être fournie.

Dans notre thèse, nous nous concentrons sur la création de ces chronologies médicales en abordant deux questions connexes en TAL : l'extraction d'informations temporelles et la résolution de la coréférence dans des documents cliniques.

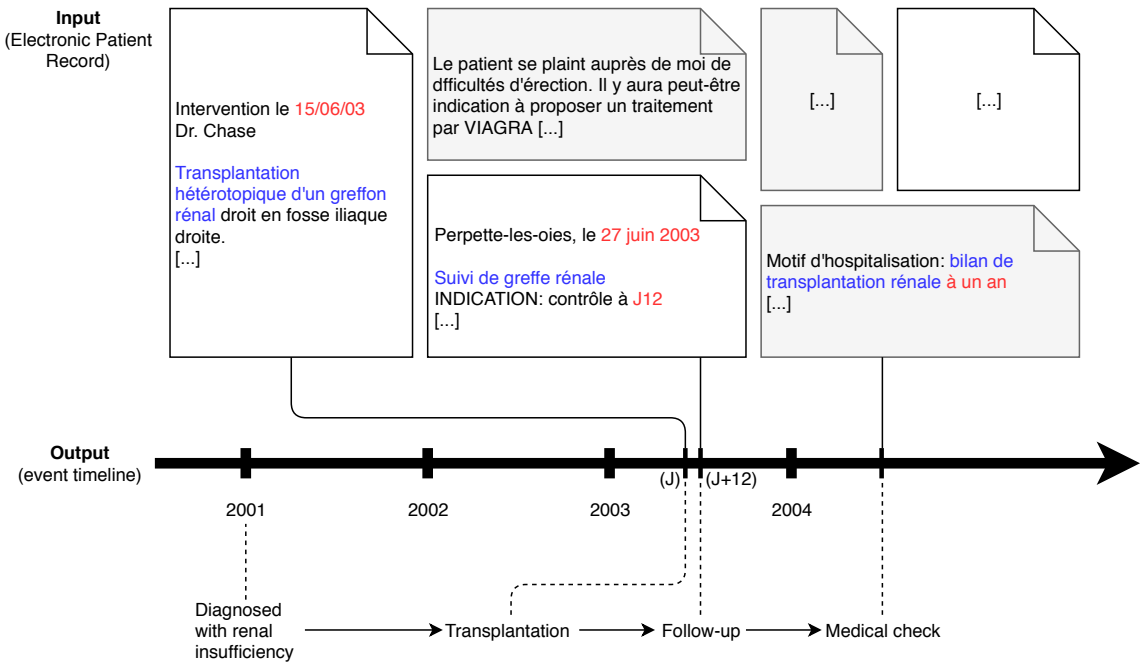

FIG. D.1 : Exemple de chronologie médicale.

D.1 Extraction d'informations temporelles

Intuitivement, la création de chronologies médicales requiert d'extraire les informations temporelles pertinentes dans les documents. L'extraction d'informations temporelles a déjà une longue histoire en TAL et s'est principalement développée dans le domaine journalistique. Les premiers travaux de recherche sur le sujet remontent aux conférences MUC (Sundheim 1993). À l'époque, la portée de l'effort se limitait à extraire les dates et les heures de reportages. Une étape importante a été franchie en 2005 avec la création de TimeML (Pustejovsky et al. 2005), le premier schéma complet pour l'annotation de la temporalité. Ce schéma permet de modéliser avec précision l'information temporelle dans le texte et a influencé la plupart des travaux de recherche suivants jusqu'à ce jour. Le domaine clinique a commencé à s'intéresser à l'extraction d'informations temporelles au début des années 2010 avec la publication du corpus i2b2 (W. Sun et al. 2013b). La grande majorité des corpus de données temporellement annotés dans les domaines général et clinique contiennent des documents rédigés en anglais. Cependant, plusieurs travaux se sont attelés à la création de ressources dans d'autres langues. Par exemple, Campillos et al. (2018) ont récemment annoté un ensemble de documents cliniques rédigés en français.

L'extraction d'informations temporelles dans le domaine clinique peut se décomposer en deux étapes principales. La première concerne l'extraction des mentions d'événements et des expressions temporelles. La seconde concerne l'extraction des relations temporelles. Dans l'Example 61, l'événement clinique (*neck pain*) et l'expression temporelle (*July*) sont liés par une relation temporelle (*begins-on*). Les définitions des événements, des expressions temporelles et des relations temporelles varient d'un corpus à l'autre, mais elles sont souvent dérivées du schéma TimeML. À la suite de ce processus en deux étapes, l'information temporelle acquise à partir des documents peut être agrégée pour former une chronologie.

(61) She has been experiencing [$_{\text{EVENT}}$ **neck pain**] since [$_{\text{DATE}}$ **July**].
→ neck pain BEGINS-ON July

Au cours des cinq dernières années, les approches basées sur des traits pour l'extraction automatique d'informations temporelles ont été progressivement remplacées par des approches neuronales. Cependant, les espoirs initiaux d'une amélioration majeure des performances se sont rapidement évanouis car les approches classiques basées sur les traits demeurent encore compétitives (Bethard et al. 2015 ; Bethard et al. 2016 ; Bethard et al. 2017). Un courant de recherche au sein de la communauté TAL essaye actuellement de combiner les deux types d'approches. Les performances ont été améliorées grâce à l'inclusion de traits catégoriels dans les approches neuronales. Par exemple, Dligach et al. (2017) utilisent les étiquettes morphosyntaxiques pour l'extraction des relations temporelles dans les documents cliniques.

D.2 Résolution de la coréférence

Un autre phénomène linguistique à prendre en considération lors de la construction de chronologies médicales est la coréférence. Les documents cliniques contiennent de multiples mentions des mêmes événements en raison du fait que le personnel médical doit faire un suivi de ces événements. Intuitivement, ce phénomène apporte du bruit lors de la création des chronologies médicales et doit être traité de façon appropriée.

Comme pour l'extraction d'informations temporelles, la résolution de la coréférence a été abordée principalement dans le domaine journalistique et les premières évaluations systématiques remontent aux mêmes conférences (Hirschman et Chinchor 1998 ; Sundheim 1995). Dans le domaine clinique, l'initiative i2b2 est à nouveau responsable de la diffusion du premier corpus de données cliniques annoté avec des chaînes de coréférence (Uzuner et al. 2012).

La résolution de la coréférence dans le domaine clinique peut être divisée en deux étapes principales. Premièrement, il faut extraire les portions de textes susceptibles d'êtres coréférentes. Habituellement, ces éléments entrent dans des catégories spécifiques telles que les médicaments ou les procédures médicales. Une fois que ces mentions ont été extraites, il faut regrouper celles qui font référence aux mêmes événements médicaux. Dans l'Exemple 62, toutes les mentions d'événements entre crochets font référence au même événement médical.

(62) The CXR revealed [**8 mm obstructing stone**] ... [**The renal stone**] was considered to be the cause of patient's symptoms ... We recommended surgical procedure to remove [**ureteropelvic stone**] ...

Les efforts de recherche portant sur la résolution de la coréférence dans le domaine clinique mettent en œuvre des approches fondées sur des traits (Uzuner et al. 2012). Cependant, les approches neuronales sont devenues de plus en plus populaires dans le domaine général (Lu et Ng 2018). De plus, nous notons que les approches hybrides incluant des traits catégoriels dans les réseaux neuronaux ont permis une amélioration significative des performances. Par exemple, K. Lee et al. (2017) ont encodé les informations du locuteur, du genre du texte et d'autres caractéristiques sous la forme de représentations denses dans une approche neuronale pour la résolution de la coréférence.

D.3 Interdépendance des domaines

Comme nous l'avons mentionné plus haut, l'extraction d'informations temporelles et la résolution de la coréférence sont deux sujets qui doivent être abordés lorsque l'on envisage de construire des chronologies médicales. Les mentions d'événements doivent être extraites et placées dans le temps. Simultanément, les mentions doivent être regroupées si elles font référence au même événement.

Au-delà de la nécessité évidente d'aborder les deux tâches simultanément, nous avançons que les deux sujets sont interdépendants. Intuitivement, deux mentions d'événements qui sont coréférentes partagent les mêmes informations temporelles. Cela a des implications lorsque l'on considère que les tâches sont effectuées dans un ordre spécifique (résolution de la coréférence *puis* extraction d'informations temporelles). Les informations apportées par la résolution de la coréférence pourraient être utilisées pour accroître la performance d'un système d'extraction d'informations temporelles.

Inversement, deux mentions d'événements qui partagent la même localisation temporelle et la même signification ont une forte probabilité d'être coréférentes. L'information temporelle supplémentaire qui serait disponible pour un système de résolution de la coréférence pourrait aider à améliorer la performance du système.

Sur la base de ces observations, il devient clair que l'extraction d'informations temporelles et la résolution de la coréférence doivent être traitées conjointement, non seulement parce qu'elles sont nécessaires pour la construction de chronologies médicales, mais aussi parce qu'il existe une certaine complémentarité entre les deux sujets.

D.4 Questions de recherche

L'extraction d'informations temporelles est un sujet complexe qui nécessite des corpus annotés soigneusement conçus. Ces ressources sont principalement en anglais et les approches développées dans la communauté sont biaisées en faveur de cette langue. Ceci motive notre première question de recherche : **est-il possible de concevoir une approche générique pour l'extraction d'informations temporelles qui pourrait être utilisée pour différentes langues ?**

Ces corpus contiennent souvent un riche ensemble de traits décrivant les entités annotées. De plus, il existe un grand nombre d'outils de prétraitement de documents cliniques qui permettent d'ajouter plus d'informations sur ces entités. Dans ce contexte, **comment cette diversité de traits catégoriels pourrait-elle être utilisée dans les approches neuronales ?** Cela soulève une question connexe : **comment l'utilisation de ces traits catégoriels influe-t-elle la performance de ces approches ?**

La résolution de la coréférence est un sujet actif dans le domaine clinique. Comme nous l'avons mentionné plus haut, l'extraction d'informations temporelles et la résolution de la coréférence sont liées. Cela nous amène à notre quatrième question de recherche : **quel type d'information temporelle pourrait être utile pour la résolution de la coréférence ?** Les approches développées pour la résolution de la coréférence sont complexes et il est très difficile de parvenir à améliorer les performances de modèles simples. Dans ce contexte, **comment cette information temporelle peut être intégrée dans une approche neuronale pour la résolution de la coréférence ?**

D.5 Contributions

Dans la première partie de cette thèse, nous abordons l'extraction d'informations temporelles dans les documents cliniques. Nous présentons quatre contributions sur ce sujet :

- **Une approche à base de traits pour l'extraction des conteneurs narratifs dans des documents cliniques**. Nous concevons une approche à base de traits pour l'extraction des relations temporelles. Nous testons notre système sur un corpus de documents écrits en anglais dans le contexte de l'édition 2016 de la campagne d'évaluation Clinical TempEval (Bethard et al. 2016).
- **Une abstraction de notre approche à base de traits**. Nous effectuons une évaluation empirique sur deux corpus de documents écrits en anglais et en français. Nous montrons qu'une approche similaire peut être utilisée pour les deux langues.
- **Une approche neuronale pour l'extraction d'informations temporelles qui intègre des traits catégoriels**. Nous abordons l'extraction d'événements, d'expressions temporelles et de relations temporelles dans les documents cliniques. Nous testons notre

approche sur un ensemble de documents écrits en anglais dans le contexte de l'édition 2017 de la campagne d'évaluation Clinical TempEval (Bethard et al. 2017).

- **Une étude empirique de l'effet des traits sur la performance de notre approche neuronale**. Nous utilisons les traits gold-standard disponibles dans le corpus, mais aussi des traits obtenus grâce à l'utilisation d'outils de prétraitement.

La deuxième partie de cette thèse est consacrée à la résolution de la coréférence dans les documents cliniques. Nous présentons deux contributions sur ce sujet :

- **Une approche neuronale pour la résolution de la coréférence dans le domaine clinique** inspirée des approches récentes dans le domaine général. Nous abordons la résolution de la coréférence à la fois sur les mentions gold-standard et les mentions prédites.
- **Une tentative d'élaboration d'un trait temporel dérivé de la relation temporelle qui existe entre les événements et les dates de création des documents**. Nous testons ce trait dans le contexte d'une étude empirique qui vise à mesurer comment les traits catégoriels et les modules neuronaux tels que les mécanismes d'attention et les représentations au niveau des caractères influent sur la performance.

En plus des contributions susmentionnées, nous concevons deux ressources. Tout d'abord, nous avons converti le corpus i2b2 task1c au format CoNLL. Cette transformation pourrait favoriser la recherche en TAL sur la résolution de la coréférence en permettant une meilleure reproduction des résultats et en facilitant le traitement des corpus. Deuxièmement, nous avons regroupé notre module neuronal d'étiquetage de séquences dans un outil open-source appelé YASET qui peut être utilisé pour toute tâche d'étiquetage de séquences en TAL.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Création de Chronologies d'Événements Médicaux: Extraction d'Informations Temporelles et Résolution de la Coréférence dans les Dossiers Patients Électroniques

Mots clés : traitement automatique des langues, extraction d'information temporelles, resolution de la coréference, chronologies médicales, dossiers patients électroniques

Résumé :
Les dossiers patients électroniques contiennent des informations importantes pour la santé publique. La majeure partie de ces informations est contenue dans des documents rédigés en langue naturelle. Bien que le texte texte soit pertinent pour décrire des concepts médicaux complexes, il est difficile d'utiliser cette source de données pour l'aide à la décision, la recherche clinique ou l'analyse statistique.
Parmi toutes les informations cliniques intéressantes présentes dans ces dossiers, la chronologie médicale du patient est l'une des plus importantes. Être capable d'extraire automatiquement cette chronologie permettrait d'acquérir une meilleure connaissance de certains phénomènes cliniques tels que la progression des maladies et les effets à long-terme des médicaments. De plus, cela permettrait d'améliorer la qualité des systèmes de question–réponse et de prédiction de résultats cliniques. Par ailleurs, accéder aux chronologies médicales est nécessaire pour évaluer la qualité du parcours de soins en le comparant aux recommandations officielles et pour mettre en lumière les étapes de ce parcours auxquelles une attention particulière doit être portée.
Dans notre thèse, nous nous concentrons sur la création de ces chronologies médicales en abordant deux questions connexes en traitement automatique des langues: l'extraction d'informations temporelles et la résolution de la coréférence dans des documents cliniques.
Concernant l'extraction d'informations temporelles, nous présentons une approche générique pour l'extraction de relations temporelles basée sur des traits catégoriels. Cette approche peut être appliquée sur des documents écrits en anglais ou en français. Puis, nous décrivons une approche neuronale pour l'extraction d'informations temporelles qui inclut des traits catégoriels.
La deuxième partie de notre thèse porte sur la résolution de la coréférence. Nous décrivons une approche neuronale pour la résolution de la coréférence dans les documents cliniques. Nous menons une étude empirique visant à mesurer l'effet de différents composants neuronaux, tels que les mécanismes d'attention ou les représentations au niveau des caractères, sur la performance de notre approche.

Title : Extracting Clinical Event Timelines: Temporal Information Extraction and Coreference Resolution in Electronic Health Records

Keywords : natural language processing, temporal information extraction, coreference resolution, clinical timelines, electronic health records

Abstract :
Important information for public health is contained within Electronic Health Records (EHRs). The vast majority of clinical data available in these records takes the form of narratives written in natural language. Although free text is convenient to describe complex medical concepts, it is difficult to use for medical decision support, clinical research or statistical analysis.
Among all the clinical aspects that are of interest in these records, the patient timeline is one of the most important. Being able to retrieve clinical timelines would allow for a better understanding of some clinical phenomena such as disease progression and longitudinal effects of medications. It would also allow to improve medical question answering and clinical outcome prediction systems. Accessing the clinical timeline is needed to evaluate the quality of the healthcare pathway by comparing it to clinical guidelines, and to highlight the steps of the pathway where specific care should be provided.
In this thesis, we focus on building such timelines by addressing two related natural language processing topics which are temporal information extraction and clinical event coreference resolution.
Our main contributions include a generic feature-based approach for temporal relation extraction that can be applied to documents written in English and in French. We devise a neural based approach for temporal information extraction which includes categorical features.
We present a neural entity-based approach for coreference resolution in clinical narratives. We perform an empirical study to evaluate how categorical features and neural network components such as attention mechanisms and token character-level representations influence the performance of our coreference resolution approach.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France