



HAL
open science

Leveraging Transformer-Based Language Models to Bridge the Gap Between Language and Specialized Domains

Abdine Hadi

► **To cite this version:**

Abdine Hadi. Leveraging Transformer-Based Language Models to Bridge the Gap Between Language and Specialized Domains. Artificial Intelligence [cs.AI]. Institut Polytechnique de Paris, 2024. English. NNT : 2024IPPAX020 . tel-04706229

HAL Id: tel-04706229

<https://theses.hal.science/tel-04706229v1>

Submitted on 23 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2024IPPAX020

Thèse de doctorat



Leveraging Transformer-Based Language Models to Bridge the Gap Between Language and Specialized Domains

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Polytechnique

École doctorale n°626 : l'École Doctorale de l'Institut Polytechnique de Paris
(ED IP Paris)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 02/04/2024, par **HADI ABDINE**

Composition du Jury :

Preslav Nakov Full Professor, Mohamed bin Zayed University of Artificial Intelligence	Président, Rapporteur
François Yvon Directeur de recherche, Sorbonne Université (ISIR)	Rapporteur
Oana Balalau Chargée de recherche, INRIA (Saclay)	Examinatrice
Éric Moulines Full Professor, École Polytechnique (CMAP)	Examineur
Christophe Cerisara Directeur de recherche, LORIA (Nancy)	Examineur
Jie Tang Full Professor, Tsinghua University	Examineur
Michalis Vazirgiannis Full Professor, École Polytechnique (LIX)	Directeur de thèse
Davide Buscaldi Maître de conférences, Université Sorbonne Paris Nord (LIPN)	Co-encadrant de thèse

ABSTRACT

The era of transformer-based language models has led the way in a new paradigm in Natural Language Processing (NLP), enabling remarkable performance across a wide range of tasks from both fields Natural Language Understanding (NLU) and Natural Language Generation (NLG). This dissertation delves into the transformative potential of transformer-based language models when applied to specialized domains and languages. It comprises four distinct research endeavors, each contributing to the overarching goal of enhancing language understanding and generation in specialized contexts.

To address the scarcity of non-English pretrained language models in both general and specialized domains, we explore the creation of two language models JuriBERT and GreekBART. JuriBERT is a set of French legal domain-specific BERT models tailored to French text, catering to the needs of legal professionals. JuriBERT is evaluated on two French legal tasks from the court of cassation in France. The findings underscore that certain specialized tasks can be better addressed with smaller domain-specific models compared to their larger generic counterparts. We equally introduce GreekBART, the first Greek Seq2Seq model. Being based on BART, these models are particularly well-suited for generative tasks. We evaluate GreekBART's performance against other models on various discriminative tasks and assess its capabilities in NLG using two Greek generative tasks from GreekSUM, a novel dataset introduced in this research. We show GreekBART to be very competitive with state-of-the-art BERT-based multi-lingual and mono-lingual language models such as GreekBERT and XLM-R.

We dive next into the domain of semantics by leveraging the transformer-based contextual embeddings to solve the challenging problem of Word Sense Induction (WSI). We propose a novel unsupervised method that utilizes invariant information clustering (IIC) and agglomerative clustering to enrich and cluster the target word representations. Extensive evaluation on two WSI tasks and multiple pretrained language models demonstrates the competitiveness of our approach compared to state-of-the-art baselines.

Finally, we introduce Prot2Text framework, a multi-modal approach for generating proteins' functions in free text by combining three modalities: protein structure, protein sequence and natural language. Prot2Text advances protein function prediction beyond traditional classifications. Integrating Graph Neural Networks (GNNs) and Large Language Models (LLMs) in an encoder-decoder framework. Empirical evaluation on a multi-modal protein dataset showcases the effectiveness of Prot2Text, offering powerful tools for function prediction in a wide range of proteins.

RÉSUMÉ

L'ère des modèles de langage basés sur des 'transformers' a ouvert la voie à un nouveau paradigme dans le traitement du langage naturel (NLP), permettant des performances remarquables dans un large éventail de tâches dans les domaines de la compréhension du langage naturel (NLU) et de la génération du langage naturel (NLG). Cette thèse se penche sur le potentiel de transformation des modèles de langage basés sur les 'transformers' lorsqu'ils sont appliqués à des domaines et des langues spécialisés. Elle comprend quatre projets de recherche, chacun contribuant à l'objectif global d'amélioration de la compréhension et de la génération du langage dans des contextes spécialisés.

Pour répondre à la rareté des modèles de langue non anglophones pré-entraînés dans les domaines généraux et spécialisés, nous explorons la création de deux modèles de langue : JuriBERT et GreekBART. JuriBERT est un ensemble de modèles BERT spécifiques au domaine juridique français, et qui répondent aux besoins des professionnels juridiques. JuriBERT est évalué sur deux tâches juridiques françaises provenant de la cour de cassation en France. Les résultats soulignent que certaines tâches spécialisées peuvent être mieux traitées avec de petits modèles spécifiques à un domaine qu'avec leurs homologues génériques de plus grande taille. Nous présentons également GreekBART, le premier modèle Seq2Seq grec. Basés sur BART, ces modèles sont particulièrement bien adaptés aux tâches génératives. Nous évaluons les performances de GreekBART par rapport à d'autres modèles sur diverses tâches discriminatives et évaluons ses capacités en NLG en utilisant deux tâches génératives grecques de GreekSUM, un nouvel ensemble de données introduit dans cette recherche. Nous montrons que GreekBART est très compétitif par rapport aux modèles linguistiques multilingues et monolingues basés sur BERT, tels que GreekBERT et XLM-R. À la fois JuriBERT et GreekBART sont les premiers modèles dans leurs domaines et langues respectifs. Avant JuriBERT, nous n'avions pas de modèle BERT spécialisé dans le domaine juridique français, tandis qu'avant - selon nos connaissances - GreekBART et GreekSUM, nous n'avions pas de modèle monolingue génératif pour la langue grecque ni de tâche générative pour le grec non plus. De plus, le seul modèle Seq2Seq multilingue (mBART) n'incluait pas la langue grecque dans son corpus de pré-entraînement. Les deux modèles sont open source et disponibles pour tous.

Nous examinons ensuite le domaine de la sémantique en tirant parti des représentations vectorielles contextuelles basées sur les 'transformers' pour résoudre le problème de l'induction du sens des mots (WSI). Nous proposons une nouvelle méthode non supervisée qui utilise le regroupement d'informations invariantes (IIC) et le regroupement agglomératif pour enrichir et regrouper les représentations des mots cibles. Une évaluation approfondie sur deux tâches WSI et de multiples modèles de langage pré-entraînés démontre la compétitivité de notre approche par rapport à l'état de l'art. De plus, nous avons appliqué une

nouvelle technique pour estimer un nombre dynamique de sens dans les mots cibles, basée sur la quantification de la polysémie des mots telle que présentée dans des recherches antérieures, et nous avons prouvé son utilité. Enfin, dans ce contexte, nous avons étudié l'impact de la profondeur de la couche de transformation sur les performances dans quatre modèles différents. Nos résultats fournissent des informations précieuses pour les chercheurs engagés dans des travaux futurs sur l'induction de sens des mots (WSI).

Enfin, nous présentons Prot2Text, une approche multimodale permettant de générer des fonctions de protéines en texte brut en combinant trois modalités pour la première fois : la structure des protéines, la séquence des protéines et le langage naturel. Prot2Text fait progresser la prédiction des fonctions des protéines au-delà des classifications traditionnelles. Prot2Text intègre des réseaux neuronaux graphiques (GNN) et des large modèles de langage (LLM) dans un cadre codeur-décodeur. Une évaluation empirique sur un ensemble de données protéiques multimodales montre l'efficacité de Prot2Text, qui offre des outils puissants pour la prédiction de la fonction d'une large gamme de protéines. Nous avons également publié un ensemble de données protéiques multimodales complet comprenant 256 690 structures protéiques, séquences et descriptions textuelles de fonctions extraites de SwissProt et AlphaFold. Les modèles Prot2Text, l'ensemble de données, les codes et une démo sont disponibles publiquement et en open source.

PUBLICATIONS

The following publications are included in parts or in an extended version in this thesis:

Iakovos Evdaimon, **Hadi Abdine**, Christos Xypolopoulos, Stamatis Outsios, Michalis Vazirgiannis, and Giorgos Stamou (2024). « GreekBART: The First Pretrained Greek Sequence-to-Sequence Model. » In: *Proceedings of the LREC-COLING 2024 conference*. Torino, Italy.

Stella Douka, **Hadi Abdine**, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles (Nov. 2021). « JuriBERT: A Masked-Language Model Adaptation for French Legal Text. » In: *Proceedings of the Natural Language Processing Workshop 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 95–101. DOI: [10.18653/v1/2021.nllp-1.9](https://doi.org/10.18653/v1/2021.nllp-1.9). URL: <https://aclanthology.org/2021.nllp-1.9>.

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis (2024). « Prot2Text: Multimodal Protein’s Function Generation with GNNs and Transformers. » In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence*.

Hadi Abdine, Moussa Kamal Eddine, Davide Buscaldi, and Michalis Vazirgiannis (2023). « Word sense induction with agglomerative clustering and mutual information maximization. » In: *AI Open* 4, pp. 193–201. ISSN: 2666-6510. DOI: <https://doi.org/10.1016/j.aiopen.2023.12.001>. URL: <https://www.sciencedirect.com/science/article/pii/S2666651023000232>.

Other contributions during the preparation of this dissertation:

Hadi Abdine, Yanzhu Guo, Virgile Rennard, and Michalis Vazirgiannis (June 2022a). « Political Communities on Twitter: Case Study of the 2022 French Presidential Election. » In: *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*. Marseille, France: European Language Resources Association, pp. 62–71. URL: <https://aclanthology.org/2022.politicalnlp-1.9>.

Hadi Abdine, Christos Xypolopoulos, Moussa Kamal Eddine, and Michalis Vazirgiannis (June 2022b). « Évaluation et Production de Plongements de Mots à Partir de Contenus Web Français à Grande Échelle. » In: *Conférence Nationale en Intelligence Artificielle 2022 (CNIA 2022)*. Actes CNIA 2022. Saint-Etienne, France. URL: <https://hal.science/hal-03866286>.

ACKNOWLEDGMENTS

First, I would like to thank my supervisor, Prof. **Michalis Vazirgiannis**, who opened the door for me to integrate the LIX laboratory and join his great team DaSciM. He chose me to conduct this Ph.D. and ensured the best environment and resources to work on this dissertation. I also would like to thank my thesis coadvisor, Prof. **Davie Buscaldi**.

Secondly, I would like to thank the two committee members, Prof. **François Yvon** and Prof. **Preslav Nakov** who kindly accepted to read my dissertation and to deliver a report evaluating my work.

Additionally, I would like to thank all the great jury members who kindly accepted to be in my Ph.D. defense committee. These members are: Prof. **Éric Moulines**, Prof. **Oana Balalau**, Prof. **Jie Tang**, and Prof. **Christophe Cerisara**. I would like to thank École Polytechnique and l'Agence Nationale de la Recherche (ANR-CHIA-0020-01) for funding my research.

Then I would like to thank all my colleagues at DaSciM, especially my collaborators, **Christos Xypolopoulos**, **Michail Chatzianastasis**, **Iakovos Evdaimon**, and **Stella Douka**.

Then, I want to thank **Moussa Kamal Eddine**, who was not only one of my collaborators who generously guided me at the beginning of my Ph.D. by delivering important tips, but also one of my best friends who was next to me during the whole journey of my Ph.D.

I am grateful to my friends in France for their constant daily support, particularly **Mohamad Al Assaad** and **Ahmad Chamma**. I am equally grateful to **Mahmoud Malass** the person who consistently provided me with encouragement all the way from Lebanon. With warm and sincere thanks to my dear friends **Sahar**, **Jana** and **Omar**

Finally, my greatest thanks go to my beloved family: my **father**, **mother**, **brothers**, **sisters**, **nieces** and **nephews**. They are the greatest blessing I have, always providing me with unwavering support and boundless love. Their enduring love remains the most treasured and profound gift, shaping the very essence of my academic and personal fulfillment.

Hadi ABDINE
Palaiseau, February 2024

CONTENTS

1	Introduction	1
1.1	Thesis Statement	6
1.2	Summary of Contributions	7
1.3	Software and Libraries	9
1.4	Outline of the Thesis	10
2	Preliminaries	11
2.1	Attention is All you Need	11
2.2	Pretrained Language Models	14
2.3	Protein Folding and Language Models	17
2.4	SemEval	20
2.5	Evaluation Measures	21
3	JuriBERT	25
3.1	Introduction	25
3.2	Related Work	26
3.3	Court of Cassation	27
3.4	JuriBERT	28
3.5	Downstream Evaluation Tasks	29
3.6	Results	32
3.7	Limitations	34
3.8	Conclusions and Future Work	34
4	GreekBART	37
4.1	Introduction and Related Work	37
4.2	GreekBART	40
4.2.1	Pre-training Corpus	40
4.2.2	Training details	41
4.3	GreekSUM	42
4.3.1	Motivation	42
4.3.2	Data collection	43
4.3.3	Post-processing	43
4.3.4	Analysis	43
4.4	Experiments	44
4.4.1	Discriminative tasks	44
4.4.2	Summarization	47
4.5	Conclusion	50
5	Word Sense Induction	51
5.1	Introduction	51
5.2	Related Work	53
5.3	Method	54
5.3.1	Dataset Setup	55
5.3.2	Vectors Extraction	55
5.3.3	Loss Function	55
5.3.4	Sense Embedding: Obtaining new word vectors	57

5.3.5	Clustering	57
5.4	Evaluation	58
5.4.1	SemEval-2010 task 14:	58
5.4.2	SemEval-2013 task 13:	59
5.4.3	Experiments	59
5.4.4	Results	60
5.5	Testing Various pre-trained Language Models	62
5.6	Ablation Study	63
5.7	Best LM Layer	63
5.8	Conclusion	65
5.9	Limitations	65
6	Prot2Text	67
6.1	Introduction	67
6.2	Related Work	68
6.3	Methodology	71
6.4	Experimental Results	74
6.5	Tokenization	75
6.6	Text Generation	76
6.7	Conclusion	80
6.8	Limitations and Future Work	81
7	Conclusion	83
7.1	Contributions and Limitations Discussion	83
7.2	Future Research Directions	85
7.3	Epilogue	85
	Bibliography	87
	Appendix	
A	Appendix : JuriBERT	107
B	Appendix : GreekBART Examples	111
B.1	GreekSUM Abstract	111
B.2	GreekSUM Title	116
C	Appendix : Various examples from the word sense induction datasets	121

LIST OF FIGURES

Figure 2.1	An example of a protein (Q6UFZ8) with its amino acid sequence and its 3D structure obtained from AlphaFold.	18
Figure 3.1	Ten recessive <i>matières</i> with the least number of examples in the test dataset.	30
Figure 3.2	Distribution of the 151 <i>matières</i> in the court of cassation data. The distribution reveals a significant imbalance of data among the classes.	31
Figure 5.1	The different sense-based clusters of the word bank with the most frequent words used in the corresponding contexts. We used PCA to project the cluster centroids into a 2D space. Each color corresponds to a cluster. The size of the points represents the frequency of the words in their corresponding cluster.	52
Figure 5.2	The pipeline of our method: For the word "live" chosen as the target, a list of sentences is provided. BART is used to generate the corresponding paraphrases. The hidden representation X_{live}^l of the target word is extracted from the layer l of a pre-trained language model. The dashed line denotes shared parameters.	54
Figure 5.3	The AVG scores of SemEval-2010 and SemEval-2013 WSI tasks using agglomerative clustering on all the layers of different pre-trained models.	64
Figure 6.1	Architecture of the proposed Prot2Text framework for predicting protein function descriptions in free text. . . .	71
Figure 6.2	Analyzing protein description lengths: Distribution of tokens per sample with threshold highlight at 256 tokens (in red).	75
Figure 6.3	Tracking Prot2Text _{BASE} BLEU score progression on validation set across training iterations. Higher is better. . . .	77
Figure 6.4	The test BLEU score for Prot2Text models as a function of the percentage identity using BLAST hit between the test and the train sets.	80
Figure 6.5	Ground-truth description vs text-free generated description: A textual comparison of the predefined description and generated text output for three different proteins from the test set.	81
Figure a.1	Accuracy, Precision, Recall and F1-Score of JuriBERT _{SMALL} on the chambers and sections classification task on the test dataset. The graph contains all eight classes.	107

Figure a.2	Confusion Matrix of JuriBERT _{SMALL} on the chambers and sections classification task on the test dataset. The graph includes accuracy and error rate for each class. . .	108
Figure a.3	Sample of Accuracy, Precision, Recall and F1-Score of JuriBERT _{SMALL} on the <i>matières</i> classification task on the test dataset. The graph contains 28 classes and the overall accuracy of all 148 classes.	109

LIST OF TABLES

Table 3.1	Architectural comparison of JuriBERT models. Where L is the number of transformer layers, H is the embedding dimension, and A_4 is the number of attention heads. . .	29
Table 3.2	Size of pre-training corpora used by different models. CamemBERT _{BASE} , JuriBERT models and JuriBERT-FP are encoder-only models, while BARThez is a pre-trained encoder-decoder model.	30
Table 3.3	Chambers and sections of the court of cassation, their data support and some of their subjects. The dataset does not differentiate between the different sections of the commercial chamber.	31
Table 3.4	Accuracy of models on the chambers and sections classification task. JuriBERT _{SMALL} has the highest accuracy despite not being the bigger model.	33
Table 3.5	Accuracy of models on the <i>matières</i> classification task. . .	34
Table 4.1	Datasets which consists of the GreekBART pre-training corpus (sizes in GB, before and after cleaning and deduplication).	41
Table 4.2	Sizes (column 2) are given in thousands of documents. Document and summary lengths are in words, while vocabulary sizes are in thousands of tokens.	42
Table 4.3	Degree of abstractiveness of GreekSUM compared with that of other datasets. It shows that GreekSUM follows XSum and OrangeSum, being more abstractive than traditional summarization datasets.	42
Table 4.4	Results on discriminative tasks. We present the mean accuracy, as well as the standard deviation.	46
Table 4.5	Results on GreekSUM. Except for ROUGE, we also provide the BertScore. The left-hand BERTScore has calculated using the M-BERT model (Devlin et al., 2019), while the right-hand one uses the Greek-BERT (Koutsikakis et al., 2020).	46

Table 4.6	Proportion of novel n-grams in the generated summaries. In addition, the length (number of words) of the generated summaries.	47
Table 4.7	The percentage of repeated words on the summaries.	48
Table 4.8	The results of the human evaluation study.	49
Table 5.1	The average perturbation percentage between the input text and the paraphrase. This percentage represents the proportion of changed unigrams.	59
Table 5.2	The average number of clusters obtained by using the polysemy scores on SemEval 2010 and SemEval 2013 test datasets.	60
Table 5.3	Evaluation of WSI models on SemEval 2010 task 14. The (+MIM) label indicates that mutual information maximization is applied to obtain the clustered vectors. Otherwise, the vectors from the pre-trained language models are directly used.	61
Table 5.4	Comparison of WSI-specific techniques on SemEval 2013 task 13. The (+MIM) label indicates that mutual information maximization is applied to obtain the clustered vectors. Otherwise, the vectors from the pre-trained language models are directly used.	61
Table 5.5	Results of different pre-trained language models on <i>SemEval-2010 Task 14</i> and <i>SemEval-2013 Task 13</i>	62
Table 5.6	Ablation study: Performance Comparison of Individual and Combined Loss Functions on SemEval-2010 task 14.	63
Table 5.7	The best layers of different pre-trained language models on the word sense induction tasks: <i>SemEval-2010 Task 14</i> and <i>SemEval-2013 Task 13</i>	65
Table 6.1	Test set results for different encoder models, including unimodal encoders such as vanilla-transformer, ESM2-35M, and RGCN, as well as multimodal encoders such as RGCN×vanilla-transformer and RGCN+ESM2-35M. All models share the same GPT-2 decoder.	78
Table 6.2	Test set results for different size variations of Prot2Text.	79
Table b.1	Example 1-GreekSUM Abstract	111
Table b.2	Example 2-GreekSUM Abstract	112
Table b.3	Example 3-GreekSUM Abstract	113
Table b.4	Example 4-GreekSUM Abstract	114
Table b.5	Example 5-GreekSUM Abstract	115
Table b.6	Example 1-GreekSUM Title	116
Table b.7	Example 2-GreekSUM Title	117
Table b.8	Example 3-GreekSUM Title	118
Table b.9	Example 4-GreekSUM Title	119
Table b.10	Example 5-GreekSUM Title	120
Table c.1	Random examples for the target word 'Access' from SemEval-2010 task 14 training set	121

Table c.2	Random examples for the target word 'Add' from SemEval-2013 task 13 training set	122
Table c.3	Random examples for the target word 'Access' from SemEval-2010 task 14 test set	122
Table c.4	Random examples for the target word 'Add' from SemEval-2013 task 13 test set	122

INTRODUCTION

The introduction of transformer-based language models represents a pivotal moment in the realm of Natural Language Processing (NLP), ushering in a paradigm shift that has profoundly influenced the landscape of language modeling and understanding. The revolutionary Transformer architecture, introduced by Vaswani et al. (2017), marked a departure from conventional models by harnessing self-attention mechanisms. This innovation enables the simultaneous processing of entire word sequences, breaking away from the sequential constraints of earlier models. The transformative impact of transformers extends across both Natural Language Understanding (NLU) and Natural Language Generation (NLG), fostering unprecedented advancements in linguistic capabilities. The rapid development of transformer-based language models has revolutionized the field of natural language processing (Yang et al., 2023), leading to a state of the art in many NLP tasks and positively affected other domains such as bioinformatics (Lin et al., 2023; Rives et al., 2021). Leveraging pre-trained models or large language models (LLMs) today is the common approach in nearly all natural language processing tasks. The act of publicly sharing the pre-trained models has facilitated the research community in their efforts to advance the field. This is evident in the development of new models that are either partially or completely initialized with the parameters from the released checkpoints (Luo et al., 2023). Transformer models also extend their footprint into diverse fields such as bioinformatics (Brandes et al., 2022), computer vision (Dosovitskiy et al., 2021), and neuroscience (Whittington, Warren, and Behrens, 2022). Nevertheless, despite the significant advancements in the field, there are still numerous challenges that need to be addressed and numerous applications that can benefit from the power of pre-trained language models. In the upcoming section, we will delve into an exploration of challenges encountered in the utilization of pretrained models, while also discussing potential applications that illuminate the versatility and implications of these advanced language models.

INEQUITY IN LANGUAGES AND SPECIFIC-DOMAIN MODELING

The recent effort to scale up large language models (Touvron et al., 2023a,b) has allowed them to learn a wide range of natural language tasks using few-shot in-context learning. This approach involves showing the model a few input-output examples as context before the test input, allowing the model to predict the target answer without any gradient update. While most large language models were pre-trained on multilingual data in addition to a massive English corpus, they have shown impressive abilities in languages other than English, but they perform best in languages with abundant resources, such as

French. Additionally, in some cases, these models may still require translating the inputs into English and then translating the responses back into the native language (Huang et al., 2023), disregarding in such a way some important information specific to certain languages which leads to underperformance in some specialized tasks. Nguyen et al. (2023) additionally discovered that the models might mistakenly produce incorrect language and face difficulties in handling low-resource non-Latin scripts because of the fragmented tokenization process, which involves breaking short texts into excessively long byte-level tokens.

Furthermore, the issue of language inequity is only partially resolved by multilingual models. One limitation is that these models typically support a limited number of languages, usually around 100, whereas there are approximately 7000 languages worldwide. For instance, the BART model was first released for the English language (Lewis et al., 2020), then, larger multilingual versions: mBART-25 and mBART-50 (Lewis, 2022), pre-trained on 25 and 50 languages respectively without the Greek language in their pre-training corpus, were released. Later, in 2021, Eddine, Tixier, and Vazirgiannis (2021) showed that mBART-25 would underperform compared to its monolingual French counterpart BART_{fr} in specific tasks that BART was originally created for. Thus, the importance of monolingual pre-trained language models.

In addition to language inequity in the field of pre-trained language models, we mention also language inequity in labeled datasets. For example, until the pre-training of BART_{fr} (Eddine, Tixier, and Vazirgiannis, 2021), no French summarization dataset existed, and the authors had to collect a novel French abstractive summarization dataset.

Addressing this problem, with the lack of generative models for the Greek language and the corresponding evaluation dataset, we introduce later GreekBART and GreekSUM (Iakovos Evdaimon et al., 2024) the first seq2seq Greek model based on BART along with a new Greek abstractive summarization dataset.

On the other hand, similarly to the comparison between the performance of monolingual and multilingual models, different work showed the superiority of domain-specific pre-trained models on certain domain-specific datasets compared to general-domain language models. Multiple compelling reasons drive the specialization of pre-trained language models in specific domains. There are notable variations in the manner of communication and language usage across various domains, positions, and activities. These can range from medical prescriptions and legal statements to online conversations. Gaining expertise and proficiency in these specific styles often necessitates extensive training that may span several years, with a substantial focus on practical and specialized knowledge. Furthermore, various organizations have their unique strategy that determines the optimal response to maximize their utility function for specific tasks. These models cannot be replaced by a single general-purpose pre-trained language models without customization (Ling et al., 2023). Additionally, domain-specific tasks require accurate domain knowledge, which cannot be easily attained through general pre-trained language models. Additionally, it was also shown that these specific language models outperform the general

ones on the specific domain tasks, despite being smaller.

For example, BloombergGPT (Wu et al., 2023), a 50B pre-trained language model specifically tailored to financial data, has demonstrated remarkable performance in financial tasks (in English), surpassing larger general domain language models such as OPT-66B (Zhang et al., 2022a) and BLOOM-176B (Workshop et al., 2023) by a significant margin. Despite its smaller size, BloombergGPT’s specialization in financial contexts enables it to capture and understand the unique intricacies of the financial domain. This focused training equips the model with a greater ability to process and generate contextually relevant information for finance-related tasks. The superior performance of BloombergGPT underscores the efficacy of domain-specific pre-training, showcasing how tailored models can yield superior results in specialized applications, even when competing against larger, more generalized counterparts.

However, most of the existing specialized language models are targeted towards the English language, presenting even a bigger language inequity problem than the one in general-domain language models. For instance, until the pre-training of our legal domain JuriBERT models (Douka et al., 2021), to our knowledge, no domain-specific model existed for the French language. In this context, we retain the following key points.

- Many languages are under-represented in the realm of pre-trained language models, especially in the monolingual setting.
- Multilingual models do not achieve optimal performance on specific downstream tasks for the languages they cover compared to monolingual models.
- Proper evaluation is hindered by the absence of datasets tailored to specific tasks in many languages.
- Domain-specific language models proved their importance and superiority in their fields while being smaller.
- The lack of pre-trained models is even greater in the area of domain-specific pre-trained models in most languages other than English.

WORD MEANING REPRESENTATION

Neural language models have the capability to produce word vector representations that contain extensive information about the language. These representations are learned by the models through exposure to vast quantities of unannotated text or what is known as self-supervised learning¹. Neural language models rely on the word distributional hypothesis (Harris, 1954), which asserts that words with similar meanings are found in comparable contexts. As a result, the word embeddings generated capture the distributional similarity, meaning that words that appear in similar contexts are represented by vectors that are closely positioned in the vector space. One of the first language models to produce such word embeddings is Word2Vec (Mikolov et al., 2013). Word2Vec is a technique that learns distributed representations of words

1. A term coined by Yann LeCun.

from large amounts of text data. It is a simple multi-layer perceptron layer trained based on predicting masked tokens using their surrounding context. Different extensions and variants followed Word2Vec to improve word representations such as GloVe (Pennington, Socher, and Manning, 2014) and FastText (Bojanowski et al., 2016). However, these word embeddings were limited by being **static** vectors which results in their inability to catch different senses of ambiguous or polysemous words since each word is represented by a single vector no matter the context it occurs in. To overcome this issue and create what is called 'sense embedding', different approaches rely on the combination of different static word embeddings for the words that occur in one expression. Nevertheless, their ability to capture subtle differences in meaning that result from changes in context is still restricted.

Following Word2Vec, relying on deep recurrent neural networks (Peters et al., 2018) and the attention mechanism (Vaswani et al., 2017), new, relatively large, language models pre-trained on a huge amount of textual data have emerged (Devlin et al., 2019; He et al., 2021; Liu et al., 2019). They now constitute the predominant paradigm in computational linguistics and other fields. These large models can produce what are called **contextualized** word embeddings. The potential of contextualized embeddings, which depict word instances or tokens, is extensive since each token/word can possess multiple word embeddings depending on the context in which it is found. These transformer-based token level contextualised representations present an intriguing chance to explore the semantic domain composed of word occurrences.

Additionally, due to various non-linguistic factors, the lists of senses for words with multiple meanings are highly subjective (Kilgarriff, 1997). The establishment of boundaries between word senses is not a standardized process, and the granularity of senses proposed can differ across different resources. Furthermore, due to the swift advancement of social media platforms, the meanings and usage of words are constantly evolving. They are acquiring new connotations and losing certain meanings based on emerging events and contexts which further increases the variability of the current fixed sense inventories (Hovy et al., 2006; McCrae et al., 2021).

As a result, we study in this dissertation how well transformer-based pre-trained language models represent the sense embedding of polysemous words, and if we can, without relying on any existing sense inventory, detect the different possible senses of polysemous words using only the contextualized word embeddings. In addition, to improve the quality of semantic representations, we examine the unsupervised enrichment of these contextual word embeddings by maximizing mutual information between two synthetic contexts that represent the same meaning for a target word. Studying these contextual embeddings can also make these language models better by enhancing their understanding of word semantics. The experiments mainly relied on word sense induction (WSI), an unsupervised task that aims to predict the possible senses of a polysemous word without relying on any inventory, to accomplish this objective, since it represents the unsupervised detection of word meanings. The experiments are carried out mainly in the English language, since the main WSI tasks, Se-

mEval 2010 task 14 (Manandhar and Klapaftis, 2009) and SemEval 2013 task 13 (Jurgens and Klapaftis, 2013) are also designed for the English language. However, following the same challenge of language inequity described earlier, all our approaches are designed with multilinguality in mind, where they can be applied in all languages simply by changing the language model used to extract the word representations. For example, CamemBERT (Martin et al., 2020) and BARThez (Eddine, Tixier, and Vazirgiannis, 2021) can be used in place of BERT (Devlin et al., 2019) and (Lewis et al., 2020) to apply the same method to the French language. However, there is still a lack of evaluation datasets for languages other than English, as there is no WSI dataset for French, for instance.

A FUTURE OF CROSS-DOMAINS APPLICATIONS

Multimodal learning has been an important research area in recent years (Baltrusaitis, Ahuja, and Morency, 2019). The Internet and intelligent devices have evolved significantly in recent years, leading to a surge in the transmission of multimodal data. As a result, there is a growing number of emerging application scenarios that involve multiple modes of communication (Xu, Zhu, and Clifton, 2023). Furthermore, with the rise of deep learning, deep neural networks have significantly advanced the progress of multimodal learning. In particular, the transformer architecture has introduced new possibilities and challenges to multimodal learning. Recent achievements of large language models and their multimodal variations highlight the potential of transformers to build foundational models for multimodal learning (Alayrac et al., 2022; Yi-Lin Sung, 2022).

Furthermore, massive multimodal datasets have been suggested due to the rapid advancement of Internet applications such as social networks and online retail, for example, COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), etc. In addition to the published datasets, others could also be constructed from huge existing resources. For example, a multimodal dataset could be extracted from UniprotKB, the Universal Protein Knowledgebase, which is a comprehensive resource in the field of bioinformatics. It is a central hub that provides information on protein sequences and their functional annotations. UniProtKB is a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). The database contains a vast collection of protein sequences, with each entry featuring detailed information such as the protein's function, structure, subcellular location, and related literature references. Such datasets and multimodal learning could open the door to evolution in the bioinformatics domain.

Inspired by the success of the transformer-based pre-trained language models, other modalities make use of the transformer architecture, which accelerates the way toward transformer-based pre-trained multimodal models. For instance, vision transformer (ViT) (Dosovitskiy et al., 2021) represents a departure from traditional convolutional neural networks (CNNs), which have been the

standard architecture for image recognition tasks. Vision transformers were introduced to address limitations in the scalability of CNNs and to enable more efficient processing of visual information. ViTs extend this concept to images by dividing an input image into fixed-size patches, linearly embedding these patches, and then processing them with transformer blocks. This allows the model to capture both local and global contextual information in the image. MolBERT (Fabian et al., 2020) is another important example in the domain of science, it is a bidirectional language model that uses the BERT architecture and pre-trained on the SMILES string of 1.6M molecules. MolBERT introduced a significant positive impact on molecules-related tasks and achieved a new state of the art.

With the success of transformer-based pre-trained models in different modalities, researchers started to co-train different modalities using transformers. An example is contrastive language-image pre-training (CLIP) (Radford et al., 2021), which facilitates cross-modal understanding between images and text. It combines a ViT vision encoder with a transformer-based language encoder to learn joint representations of images and their associated textual descriptions. CLIP led the way later to a revolution in image generation when its pre-trained image and text encoders were used with a diffusion process to create models such as DALL-E 2 (Ramesh et al., 2022). Other applications also benefitted from the multimodal models such as molecule-language translation with MolT5 model (Edwards et al., 2022).

Finally, with the rise of protein language models (Brandes et al., 2022; Lin et al., 2023; Rives et al., 2021) and protein folding models (Senior et al., 2020), a new important multimodal application that combines protein data and textual data arises. The problem of comprehending the function of proteins is a crucial issue in the field of biological sciences. Proteins serve as fundamental components for almost all biological functions. Precise prediction of the function of proteins is vital for understanding biological systems and has various applications, including drug discovery. This enables researchers to identify and focus on specific proteins that play a significant role in disease pathways (Ha et al., 2021). To test one of the potentially important applications that combine these modalities, this dissertation introduces Prot2Text (Hadi Abdine et al., 2024), which consists of the co-training of protein structure, protein amino acid sequence, and natural language textual data.

1.1 THESIS STATEMENT

In the realm of artificial intelligence, the development and deployment of language models have significantly advanced natural language processing capabilities. This dissertation contributes pipelines, models, and datasets with a combined goal of expanding linguistic representation and addressing specific challenges in various specialized domains. In particular, we contribute the following.

- The first seq2seq model for the Greek language (GreekBART), the first language model specialized in the French legal domain (JuriBERT), and

the first abstractive summarization dataset for the Greek Language (GreekSUM).

- A new fully unsupervised framework based on pre-trained language models to solve the word sense induction task (WSI). We prove the effectiveness of our approach by testing it on two different WSI datasets. It achieves a new state-of-the-art performance in one task while being competitive in the second.
- Prot2Text, the first multimodal framework that combines both graph neural networks (GNN) and pre-trained language model to generate protein functions in free text style using both protein structure and amino acid sequence. We also introduce a comprehensive multimodal protein dataset containing the protein details along with their function description.

We elaborate on these three contributions below.

1.2 SUMMARY OF CONTRIBUTIONS

Pretrained Language Models

Language inequity manifests prominently in the realm of language models, where certain languages benefit from dedicated models in both generalized and specific domains, while others remain underserved. Dominant languages often enjoy well-tailored models, resulting in superior performance and usability, while less represented languages are left with limited or generic models, perpetuating a linguistic imbalance. To address this, we contribute the following².

- We proposed the first Greek seq2seq model, GreekBART, and the first French language model specialized in the French legal domain, JuriBERT. GreekBART is based on the $BART_{BASE}$ architecture (Lewis et al., 2020), while JuriBERT is a set of four model ranges from a tiny model of 6M parameters to a base model of 110M parameters based on the BERT pre-training (Devlin et al., 2019).
- We automatically evaluated our proposed models against different baselines, including models with higher capacities, and showed that, in different configurations, they outperform them or have competitive results.
- We proposed GreekSUM, a Greek abstractive summarization set. At the time of this work, there was no abstract summary dataset for Greek. GreekSUM is a Greek equivalent to XSUM (Narayan, Cohen, and Lapata, 2018) to evaluate our GreekBART model, it contains two tasks, GreekSUM Title which consists of pairs of news article and the corresponding title,

2. In this work, I was mainly responsible of supervising (interns) and designing the projects (the pipelines of JuriBERT and GreekBART) including the choice of the models and the pre-training corpus while the interns (with a high knowledgeable skills and effort) wrote and ran the pre-training code and the experiments. With the help of Christos Xypolopoulos, I collected the GreekSUM datasets. For JuriBERT, I constructed the pre-training corpus along with downstream datasets with the corresponding finetuning code. In addition, I was mainly involved in the analysis of the results.

and GreekSUM Abstract that consists of pairs of news article and the corresponding abstract.

- We publicly release our JuriBERT models, GreekBART model, and GreekSUM dataset so that the NLP community can use them in future research.

Word Sense Induction with Mutual Information Maximization

With the advancements in the capabilities of pre-trained language models, engaging in the word sense induction task presents an exciting opportunity to unravel the nuances of language, as it involves automatically discerning and categorizing the various meanings a word can assume in different contexts. Transformer-based pre-trained language models can significantly enhance word sense induction systems by leveraging their contextual understanding, enabling more accurate sense disambiguation; conversely, integrating word sense induction into these models can introduce crucial enhancements by fostering a deeper semantic grasp and improving the contextual appropriateness of the text. Our contribution³ to this area is listed below.

- Introducing an innovative unsupervised method that utilizes pre-trained language models, hierarchical clustering, and mutual information maximization. Our approach overcomes certain limitations observed in previous efforts while demonstrating competitive performance.
- Applying a technique to estimate a dynamic number of senses in target words, based on the quantification of the word polysemy as presented in previous research (Xypolopoulos, Tixier, and Vazirgiannis, 2021).
- Investigating the impact of the depth of the transformer layer on performance in four different models, detailed in Section 5.7. Our findings provide valuable information for researchers engaged in future work on word sense induction (WSI).

Prot2Text: Multimodal Protein Function Generator

Motivated by the imperative need to accurately predict protein functions for a comprehensive understanding of biological systems and millions of unknown proteins, in addition to various possible applications. Given the flexibility and advancement of the transformer architecture, we propose a novel approach to overcome the limitations of traditional protein function prediction methods. Our contributions⁴ cover:

- The introduction of the **Prot2Text** framework, a multimodal approach that integrates Graph Neural Networks (GNNs) and Pretrained Language Models to generate detailed protein function descriptions in free text.

3. In this work, I designed, developed and ran all models and experiments.

4. In this work, I was responsible for designing the transformer-related parts (protein sequence encoding, the fusion module, and text generation) in addition to combining different model parts and running all experiments, while another Ph.D. student was responsible for graph protein construction and the graph encoding part. In addition, I was the main contributor for the multimodal dataset construction and the development of the demo.

- The release of a comprehensive multimodal protein dataset comprising 256,690 protein structures, sequences, and textual function descriptions extracted from SwissProt (Bairoch and Apweiler, 1996) and AlphaFold (Senior et al., 2020). This dataset is publicly available, facilitating benchmarking and fostering advancements in protein function prediction methods. The CD-HIT clustering algorithm (Li and Godzik, 2006) ensures that the proteins in the test set have a maximum sequence alignment of 40% to those in the training set.
- Proposition of various baselines for protein text generation and demonstration that the integration of graph-protein and sequence-protein information leads to better generation capabilities.
- Public release five pre-trained multimodal models of different sizes: *Prot2Text_{SMALL}*, *Prot2Text_{BASE}*, *Prot2Text_{MEDIUM}*, and *Prot2Text_{LARGE}* that use both the structural information and the sequence of the protein. In addition to *ESM2Text_{BASE}* that uses only the protein sequence in case of unavailability of the predicted folded protein.
- A Web app containing a demonstration of protein description generation using *Prot2Text_{BASE}* and *ESM2Text_{BASE}* is available, together with access to all models and the created dataset at nlp.polytechnique.fr/prot2text.

1.3 SOFTWARE AND LIBRARIES

The following are the main libraries utilized in the context of this thesis:

- Pytorch (Paszke et al., 2019). A Python library that enables instant execution of dynamic tensor computations, incorporating automatic differentiation and GPU acceleration. It is utilized for tasks such as computer vision and natural language processing, initially created by Meta AI and currently under the Linux Foundation umbrella. It is software that is both free and open-source.
- Transformers (Wolf et al., 2020). Hugging Face and the community maintains a library for PyTorch, TensorFlow, and JAX that offers cutting-edge machine learning capabilities. This library includes numerous pre-trained models that can be used for various tasks involving text, vision, and audio.
- Datasets (Lhoest et al., 2021). A library maintained by Hugging Face for easily accessing and sharing datasets for Audio, Computer Vision, and Natural Language Processing (NLP) tasks.
- Fairseq (Ott et al., 2019). An open-source sequence modeling toolkit developed by MetaAI that allows researchers and developers to train custom models for translation, summarization, language modeling, and other text generation tasks.
- Scikit-Learn (Pedregosa et al., 2011). A machine learning library for the Python programming language including different algorithms for clustering, classification and regression.

- Pytorch-Geometric (Fey and Lenssen, 2019). A library built upon PyTorch to easily write and train Graph Neural Networks (GNNs) for a wide range of applications related to structured data.
- Pandas (team, 2020). A foundational Python library for data analysis and statistics.
- Numpy (Harris et al., 2020). A fundamental package for scientific computing in Python.
- Matplotlib (Hunter, 2007). A library for creating static, animated, and interactive visualizations in Python.

Note that we made all source code and preprocessed data (Except for court of cassation private data in Chapter 3) publicly available for reproducibility and for fostering research on the topics covered by this thesis.

1.4 OUTLINE OF THE THESIS

The upcoming chapters of this dissertation are structured as follows. In Chapter 2, we provide some introductory information and fundamental knowledge that will be helpful for understanding the remaining content. Chapters 3, 4, 5, and 6 are dedicated to presenting our contributions in the field of NLP, addressing various challenges and applications. Specifically, in Chapters 3 and 4, we introduce our two pre-trained models, JuriBERT and GreekBART. Chapter c focuses on our solution for word sense induction using transformer-based pre-trained models. Lastly, in chapter 6, we present Prot2Text, a multimodal model proposed by us to generate the functions of proteins in free text. Chapter 7 concludes the dissertation and provides some limitations and suggestions for potential future research topics.

This chapter presents an explanation of key concepts and essential background information necessary to understand the remainder of the thesis. Initially, we provide an introduction to the transformer architecture, which is predominantly utilized in this dissertation. Then, we present a brief history of pre-trained language models. Next, we present the use of the transformer architecture for proteins. Then, we briefly present SemEval, more specifically, the word sense induction tasks within SemEval. Finally, we present the evaluation metrics used in this dissertation.

2.1 ATTENTION IS ALL YOU NEED

The Transformer Architecture

The Transformer architecture is a neural network architecture introduced in the paper "Attention is All You Need" by Vaswani et al. (2017). It has become a fundamental model for various natural language processing (NLP) tasks and other sequence-to-sequence tasks. The Transformer architecture relies on attention mechanisms to capture relationships between different elements in a sequence, allowing it to process input data in parallel and handle long-range dependencies more effectively than traditional sequential models such as recurrent neural networks (RNN).

The transformer architecture consists of two primary elements: the encoder and the decoder. Both elements have essential functions in processing input data and producing significant output.

The transformer encoder consists of the following key components:

- **Input Embeddings:** The input sequence, which could be a sentence or any ordered set of data, is first transformed into embeddings. Each element in the sequence (e.g., word or token) is represented as a vector in a high-dimensional space of dimension d using the embedding matrix $W_{embedding} \in \mathbb{R}^{k \times d}$ where k is the number of vocabulary words.
- **Positional Encoding:** provides the model with information on the position of each element in the sequence.
- **Multi-Head Self-Attention:** The encoder consists of multiple layers, and each layer contains a multi-head self-attention mechanism. This allows the model to weigh different parts of the input sequence differently, capturing dependencies and relationships between elements in the same sequence. Note that all the inputs here to compute the attention weights (*query*, *key*, and *value*) come from the same sequence. This is why the attention mechanism applied here is called *self-attention*.

- **Feedforward Neural Network:** After attention mechanisms, each sub-layer in the encoder includes a feedforward neural network. This network processes the information obtained from the attention layer and learns complex representations.
- **Residual Connections and Layer Normalization:** Residual connections are employed around each sub-layer (attention and feedforward) to facilitate the flow of information. Layer normalization is applied to stabilize the training process.

On the other hand, the transformer decoder consists of the following key components:

- **Output Embeddings:** Similar to the encoder, the output sequence is first transformed into embeddings. These embeddings serve as the initial representation of the decoder.
- **Positional Encoding:** Just like in the encoder, positional encoding is added to the output embeddings to convey information about the position of each element.
- **Masked Multi-Head Self-Attention:** The decoder also contains a multi-head self-attention mechanism. However, a masking strategy is applied during training to ensure that each position can only attend to previous positions, preventing information leakage from future elements.
- **Encoder-Decoder Cross-Attention:** Another crucial component of the decoder is the encoder-decoder attention layer. It allows the decoder to attend to different parts of the input sequence (from the encoder), allowing the model to effectively align source and target information. The *query* to compute the attention weights in this layer comes from the self-attention output of the decoder representing the target sequence. While the *key* and the *value* represent the input sequence from the encoder.
- **Feedforward Neural Network:** Similar to the encoder, the decoder includes a feedforward neural network that processes the information obtained from attention layers.
- **Residual Connections and Layer Normalization:** Residual connections and layer normalization are applied around each sub-layer in the decoder, similar to the encoder.

Positional Embedding

Since transformers do not inherently understand the order of elements in a sequence, as we can see in the scaled dot-product attention equation, positional encoding must be added to word embeddings to provide the model with information about the position of each element in the sequence and avoid getting the same word representations if the order of words is shuffled. In the positional encoding layer, for each position p in the input sequence, a unique

positional encoding vector PE is created and added to the embedding of the word at position p where:

$$PE_i(p) = \begin{cases} \sin(p/10000^{2i/d}) & \text{if } i \text{ is even} \\ \cos(p/10000^{2i/d}) & \text{if } i \text{ is odd} \end{cases}$$

where i represents the dimension index in the embedding space.

Attention Mechanism

The fundamental concept of an attention mechanism is similar to a spotlight that can be utilized by a neural network to concentrate on various sections of the input data while making predictions. It enables the model to dynamically assign different weights to different elements within the input sequences, which proves to be advantageous in situations where certain parts of the input have a greater impact on the output compared to others. These weights are calculated dynamically during model training and inference. The model learns to determine the importance of each element in the context of the specific task it is performing.

One key advantage of the transformer is its adaptability to varying input lengths. Unlike fixed-size neural networks that require a predefined input size, transformer can handle sequences of different lengths by dynamically adjusting the weights. Through the attention mechanism, transformer is particularly effective in capturing long-range dependencies in sequences. Traditional neural networks, especially those that rely on fixed-size windows, may struggle to capture relationships between elements that are far apart in the sequence. Attention mechanisms, through their dynamic weighting, can give importance to elements regardless of their position in the sequence.

Scaled Dot-Product Attention

Scaled dot-product attention is the function used within the attention mechanism of the transformer architecture. It is a method used to compute attention scores between different elements of a sequence (self-attention) or between two different sequences (cross-attention), allowing the model to focus on the relevant parts of the input sequence. The "scaled" part comes from a scaling factor applied to the dot product of the query and key vectors, which helps to control the magnitude of the resulting attention scores.

For given sequences S_1 and S_2 (matrices in $\mathbb{R}^{n_1 \times d}$ and $\mathbb{R}^{n_2 \times d}$ where n_1 and n_2 are the length of each sequence respectively), each token is associated with three vectors: Query (Q), key (K), and value (V). These vectors are learned during the training process. Where $Q = S_1 \cdot W_Q$, $K = S_2 \cdot W_K$, and $V = S_2 \cdot W_V$. W_Q , W_K , and W_V are trainable matrices of dimension $d \times d$.

The scaled attention score between two tokens in S_1 and S_2 , say token S_1^i and token S_2^j , is calculated by taking the dot product of Q_i with K_j and then divided by the square root of the embedding dimension d . Finally, the scaled dot

products are passed through a softmax activation function to obtain attention weights. This process is repeated for all pairs of elements in the query. The attention weights are then used to compute a weighted sum of the corresponding value (V) vectors. This weighted sum represents the contribution of different elements of S_2 to the output corresponding to the token S_1^i :

$$Attention_Output(Q_i, K, V) = \sum_j Softmax\left(\frac{Q_i \cdot K_j}{\sqrt{d}}\right) \cdot V_j$$

Multi-Head Attention

Multi-head attention is a version of the scaled dot-product attention used in the transformer architecture. It allows the model to jointly attend to different parts of the input sequence with multiple sets of attention weights, enabling it to capture diverse patterns and relationships. The idea behind multi-head attention is to have the model learn different attention patterns or "heads" in parallel:

$$H_i = Attention_Output(S_1 \cdot W_Q^i, S_2 \cdot W_K^i, S_2 \cdot W_V^i)$$

where i is the index of the attention head and W_Q^i , W_K^i , and W_V^i are trainable matrices of dimension $d \times d/h$. Then concatenate or linearly combine their outputs:

$$Multihead_Attention_Output(Q, K, V) = Concat(H_1, H_2, \dots, H_h) \cdot W_O$$

where h is the number of attention heads and W_O is a trainable matrix of dimension $d \times d$.

2.2 PRETRAINED LANGUAGE MODELS

Pre-trained language models have emerged as a transformative technology in the field of natural language processing (NLP), significantly impacting the way we approach language understanding and generation tasks. These models are trained on massive datasets containing various linguistic patterns, allowing them to learn rich representations of language. The importance of pre-trained language models lies in their ability to generalize well across a wide range of NLP tasks, alleviating the need for task-specific feature engineering and enabling breakthroughs in various applications.

Pre-trained language models follow a transfer learning paradigm, where a model is initially trained on a large corpus for a language modeling task using some self-supervised techniques (i.e. using part of unannotated input as a label) and then fine-tuned on specific downstream tasks. This approach has proven highly effective in leveraging the knowledge gained from general linguistic patterns to improve performance on specific tasks.

In addition, pre-trained language models exhibit versatility across a spectrum of NLP tasks, including sentiment analysis, named entity recognition, machine translation, question-answering, and more. This versatility is attributed to their

capacity to capture diverse linguistic patterns during pre-training. In the following subsections, we detail some of the most successful pre-trained models that are mainly used in this dissertation.

GPT

Generative pre-trained transformer (GPT) is a series of powerful and widely used language models developed by OpenAI. GPT models represent a class of transformer-based models that pre-train only a left-to-right decoder as a general language model. GPT models are pre-trained using the causal language modeling (CLM) objective. CLM is a type of language modeling in which the model is trained to predict autoregressively the next token in sequence based only on the past tokens. The training objective typically involves maximizing the likelihood of the next token given the current context:

$$Objective_{CLM} = \operatorname{argmax}_{\Theta} \sum_{t=2}^T \log P(y_t | x_1, x_2, \dots, x_{t-1}; \Theta)$$

where T is the length of the sequence, $X = (x_1, x_2, \dots, x_T)$ is the input sequence, $Y = (y_1, y_2, \dots, y_T)$ is the target sequence. For GPT models, Y is the same sequence as X , and Θ are the learnable weights of the model.

The original GPT model (Radford et al., 2018) was introduced in 2018. It is one of the earliest transformer-based language models. It demonstrated the effectiveness of pre-training large-scale transformer models for various NLP tasks by fine-tuning the model for 12 different language understanding tasks. Later, GPT-2 (Radford et al., 2019) was introduced, a more advanced version of GPT that has more trainable parameters. The authors showed that as long as general language models have very high capacities, they can reach reasonable performance on many specific natural language processing tasks. Due to its impressive language generation capabilities, it was initially considered too risky to be released at full capacity due to concerns about potential misuse. Afterward, four variants of GPT-2 were released.

Finally, GPT-3 was unveiled in 2020, the largest and most advanced model in the GPT series. With a staggering 175 billion parameters. GPT-3 (Brown et al., 2020) demonstrated remarkable language understanding and generation capabilities. It can perform a wide range of natural language tasks, including text completion, translation, question answering, and even creative writing. The weights of the model were not made public; then, in late 2022, GPT-3 was the backbone of ChatGPT, which revolutionized multiple research areas, not only NLP, and started a new research era based on large language models (LLMs). It is worth mentioning that different open source pre-trained left-to-right transformer decoders were released after GPT, such as LLaMA (Touvron et al., 2023a,b) and Mistral (Jiang et al., 2023).

BERT

BERT (Devlin et al., 2019), which stands for "Bidirectional Encoder Representations of Transformers", is a transformer-based language model introduced by Google in a 2018 research paper. BERT is designed to capture bidirectional context in a given text, allowing it to understand the meaning of words in the context of both their preceding and following words. This is in contrast to previous models that were unidirectional and considered only the context on one side of a word.

The BERT training uses two objectives. The masked language modeling (MLM) objective. It involves masking or hiding certain words in the input sequence and training the model to predict these masked words based on the context provided by the surrounding words. And the next sentence prediction (NSP) objective. The specific steps of the objective function are as follows:

1. Randomly mask some percentage of the words in the input sequence. These masked words are then replaced with a special token, such as [MASK].
2. Predict the original identity of the masked tokens using the bidirectional context provided by the other words in the sequence.
3. Sample pairs of sentences during training and concatenate them with a special separator token [SEP].
4. Predict whether the second sentence of the pair follows the first sentence (a binary classification task).

Liu et al. (2019) showed later that the NSP task does not improve performance in the downstream tasks, and thus only pre-trained RoBERTa on the MLM task:

$$Objective_{MLM} = \operatorname{argmax}_{\Theta} \sum_{i \in \mathcal{M}} \log P(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_T; \Theta)$$

where T is the length of the sequence, $X = (x_1, x_2, \dots, x_T)$ is the input sequence, \mathcal{M} is a set of randomly chosen token to be masked, and Θ are the learnable weights of the model.

While GPT is only a pre-trained transformer decoder, BERT is an encoder-only pre-trained language model.

BART

Bidirectional Auto-Regressive Transformer (BART) (Lewis et al., 2020), is a sequence-to-sequence model introduced by Facebook AI research (FAIR). BART is designed to handle various natural language processing tasks, including text generation, translation, and comprehension. It utilizes a denoising auto-encoder architecture, both encoder and decoder of the transformer architecture are used. The model is trained to reconstruct a corrupted or noisy version of the input sequence. This training approach encourages the model to capture meaningful representations and relationships within the data.

BART combines bidirectional encoder representations with an auto-regressive

decoder. The bidirectional encoder helps capture contextual information, while the auto-regressive decoder generates the output sequence step by step.

Similarly to BERT, with one of the noise functions, BART uses a masked language model objective during pre-training. It involves randomly masking or deleting some tokens in the input sequence and training the model to predict these tokens based on the context provided by the other tokens.

In addition, BART is also trained on reconstructing correct documents from noisy documents in which token shuffling (the order of some tokens in the sequence is randomly permuted) or sentence permutation (randomly permutes the order of entire sentences within the) was applied.

2.3 PROTEIN FOLDING AND LANGUAGE MODELS

Protein language models (PLMs) and protein folding are two closely related areas in bioinformatics and computational biology, each addressing different aspects of understanding and predicting protein behavior.

The primary structure of a protein is a sequence of amino acids and its function and behavior are intricately linked to this sequence. Protein language models aim to capture the language-like patterns within these amino acid sequences, while protein folding uses them to assume their functional three-dimensional shape.

Analogy between Proteins and Natural Language

Inspired by natural language processing, protein language models use techniques such as recurrent neural networks (RNNs), transformers, or other deep learning architectures to learn representations of amino acid sequences. They also follow similar pre-training objectives and transfer learning techniques to perform protein-specific downstream tasks.

Similarly to natural language, proteins can be represented as a sequence of amino acids. Amino acid sequences are fundamental components of proteins, and they play a critical role in the structure and function of these biological molecules. Understanding amino acid sequences involves recognizing the linear order of amino acids in a protein, where each amino acid is represented by a specific letter or abbreviation. Amino acids are often compared to the alphabets of a language. In the genetic code, each amino acid is represented by a specific three-letter or one-letter code (e.g., alanine is represented as Ala or A). The sequence of these codes in a protein constitutes the "words" or "tokens" that make up the language of proteins.

Finally, as in natural language, the order of these amino acids is also important. The arrangement of these sequences imparts specific functions to the protein. Although there are more than 500 amino acids in nature, by far the most important are the 22 α -amino acids incorporated into proteins. Only these 22 appear in the genetic code of life. These 22 amino acids serve as the building blocks of proteins (or, in other words, the *vocabulary* of proteins). Each amino acid has a specific chemical structure and unique properties and is represented by a letter

of the alphabet. An example of a protein with its amino acid sequence can be seen in Figure 2.1

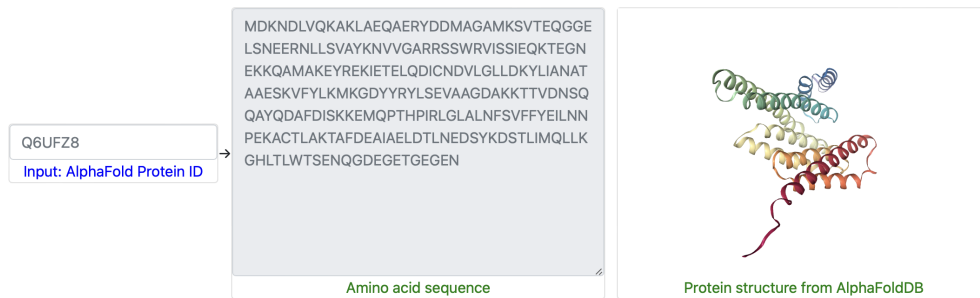


Figure 2.1 – An example of a protein (Q6UFZ8) with its amino acid sequence and its 3D structure obtained from AlphaFold. Each amino acid is represented by one letter.

PLM

Following the revolution of the transformer architecture and its ability to learn rich representations, in addition to the availability of millions of unannotated proteins sequences, different transformer-based protein language models were introduced along with remarkable performance in tasks such as predicting protein-protein interactions, identifying potential drug targets, understanding protein evolution, and annotating functional sites in proteins.

ProteinBERT (Brandes et al., 2022) and ESM (Lin et al., 2023; Rives et al., 2021) are some examples of a protein language model based on the BERT architecture. They have been pre-trained on a large corpus of protein sequences and can be fine-tuned for specific downstream tasks. On the other hand, ProtGPT2 (Ferruz, Schmidt, and Höcker, 2022) is a PLM based on the GPT architecture that can specifically help in the protein design task.

AlphaFold

Understanding protein folding is crucial to deciphering protein function and designing drugs targeting specific proteins. Predicting protein folding is a challenging problem due to the vast conformational space that proteins can explore. The relationship between an amino acid sequence of a protein and its folded structure is complex and is not fully understood. Computational methods, such as molecular dynamics simulations and machine learning techniques, are used to predict protein folding. Machine learning models, including deep learning approaches, use known protein structures to predict the folding of new sequences.

AlphaFold (Senior et al., 2020), developed by DeepMind, is a notable example of a deep learning model that has shown remarkable success in predicting protein structures. It achieved outstanding performance in the Critical Assessment of Structure Prediction (CASP) competition. AlphaFold is built on a deep neural

network that uses convolutional neural networks (CNNs)/attention mechanisms inspired by transformer architectures. During training, AlphaFold is exposed to a dataset of known protein structures, learning to correlate amino acid sequences with corresponding three-dimensional structures. The model learns to capture complex relationships between amino acids and their spatial arrangements. One of the key innovations of AlphaFold is its ability to predict interresidue distances in a protein. These distance predictions serve as crucial input for the subsequent generation of 3D models. AlphaFold predicts the 3D coordinates of the atoms of a protein using distance information and incorporating it into a refinement process. The model iteratively refines its predictions to achieve accurate and biologically significant protein structures.

Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as a powerful framework for modeling and analyzing graph-structured data, with their effectiveness demonstrated in various domains including social networks, recommendation systems, and, notably, in the field of computational biology (Kipf and Welling, 2017; Scarselli et al., 2009). GNNs excel in propagating and refining features across graph structures through iterative information exchange among nodes, leading to a nuanced encoding of the graph's structure and semantics.

Specifically, for proteins, GNNs are revolutionizing the field of protein encoding by leveraging their ability to capture complex relationships in data. In protein encoding, GNNs represent amino acids and their interactions as nodes and edges in a graph structure, enabling the network to learn intricate patterns and dependencies crucial for understanding protein function and structure. Through layers of neural computations, GNNs iteratively refine their understanding of protein sequences, effectively encoding them into high-dimensional representations. This approach not only enhances the accuracy of protein function prediction but also facilitates drug discovery and personalized medicine by elucidating the molecular mechanisms underlying diseases. By harnessing the power of GNNs, researchers are unlocking new frontiers in protein biology, offering insights that were previously inaccessible with traditional computational methods (Chen et al., 2023; Wang et al., 2022; Zhang et al., 2022b).

When delving into the realm of Graph Neural Networks (GNNs), it's essential to explore the diverse types and architectures that have emerged to tackle various challenges in graph-based learning. In the following, we present some types of GNNs.

The Graph Convolutional Network (GCN) adapts traditional convolutional principles for graph data, employing a technique where each node updates its features by aggregating features from its neighbors. This method facilitates efficient semi-supervised learning on graphs, making GCNs particularly effective for tasks like node classification (Kipf and Welling, 2017).

Expanding on GCNs, Relational Graph Convolutional Networks (RGCNs) handle graphs with multiple types of relationships. RGCNs use distinct convolutional filters for different relationship types, enhancing their utility in

complex, heterogeneous networks (Schlichtkrull et al., 2018).

Graph Attention Networks (GATs) integrate attention mechanisms that dynamically prioritize information from different neighboring nodes, adapting to nodes' varying relevance, which improves performance in diverse applications (Veličković et al., 2018).

Graph Isomorphism Networks (GINs) address the graph isomorphism problem by learning unique embeddings for different graph structures, thus distinguishing between non-isomorphic graphs. This capability makes GINs highly effective for graph classification tasks (Xu et al., 2019).

Together, these GNN architectures illustrate the flexibility and comprehensive capabilities of GNNs, highlighting their potential to transform a wide array of graph-based data analysis challenges. Each model leverages core neural network techniques to specialize in extracting insights from complex datasets, reflecting the vast applications of GNNs.

2.4 SEMEVAL

Semantic evaluation (SemEval) is an ongoing series of workshops and evaluation campaigns in the field of natural language processing (NLP) and computational linguistics. SemEval provides a platform for researchers to develop and evaluate systems for various NLP tasks. It typically involves organizing shared tasks in which participants in the research community submit their solutions to specific challenges.

The tasks covered in SemEval workshops are diverse and can include sentiment analysis, named entity recognition, semantic role labeling, question answering, and various other aspects of natural language understanding. Each task is carefully defined and participants are provided with training and testing datasets to evaluate the performance of their systems.

One main task in SemEval is word sense induction (WSI), which focuses on the automatic grouping or clustering of word instances that share a common sense or meaning, without relying on predefined sense inventories. The goal is to discover sense distinctions or groupings directly from the data. In the context of SemEval, a typical WSI task involves providing participants with a dataset containing occurrences of a target word in context. Participants are then tasked with clustering these instances according to their underlying senses or meanings. Evaluation metrics assess the quality of the induced senses, considering factors such as coherence within clusters and separation between different senses.

SemEval provides a standardized platform for researchers to develop and evaluate their WSI systems, fostering the comparison of different approaches and the advancement of techniques in the field. The tasks and datasets used in SemEval WSI challenges vary between editions, ensuring that a diverse range of words and contexts is considered. We mainly mention SemEval 2010 task 14 (Manandhar and Klapaftis, 2009) and SemEval 2013 task 13 (Jurgens and Klapaftis, 2013) for WSI.

Overall, the WSI in SemEval contributes to the broader goal of improving our

understanding of word meanings and sense distinctions in natural language, which is crucial for various NLP applications. More details about these two tasks of word sense induction will be discussed in Chapter 5.

2.5 EVALUATION MEASURES

In the field of machine learning, automatic evaluation metrics play a crucial role in advancing the field. A commonly used approach to assess the performance of a model is to partition the dataset into three subsets: training, validation, and testing. The training set is used to update the model learnable parameters. The validation set is used to select the best checkpoint of the model according to a certain metric. And the testing set is used to report the performance of the trained model using a significant evaluation metric. In this section, we present the evaluation metrics used to measure the performance of the different approaches developed in this dissertation.

Accuracy, Precision, Recall and F1-Score

Given a multiclass classification task with examples and their ground truth labels, there are four types of possible predictions: Accuracy, precision, recall, and F1 score defined as following:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN}; F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP is the number of true positive predictions, TN is the number of true negative predictions, FP is the number of false positive predictions, and FN is the number of false negative predictions.

BLEU Score

The BLEU (Bilingual Evaluation Understudy) score (Papineni et al., 2002) is a metric used to evaluate the quality of machine-generated text, especially in the context of machine translation. BLEU is widely used in natural language processing and machine translation research.

The BLEU score measures the similarity between a machine-generated text and one or more human references. It operates by comparing n-grams between the generated text and the reference texts. The primary components of the BLEU score include the precision for different sizes of n-grams and the brevity penalty, where:

$$Precision_n = \frac{\text{number of matching n-grams in the generated and reference text}}{\text{number of n-grams in the generated text}}$$

$$Brevity_Penalty = \begin{cases} 1 & \text{if generated length} \geq \text{reference length} \\ \exp\left(1 - \frac{\text{reference length}}{\text{generated length}}\right) & \text{if generated length} < \text{reference length} \end{cases}$$

$$BLEU = Brevity_Penalty \times \left(\prod_n^N Precision_n \right)^{1/N}$$

The BLEU score is typically reported as a percentage, and higher BLEU scores indicate better agreement between the machine-generated text and the reference texts. The choice of N (the maximum size of n -grams considered) can vary, and common choices include BLEU-1, BLEU-2, BLEU-3, and BLEU-4. The brevity penalty addresses the issue of shorter translations that receive higher precision scores and adjusts the final BLEU score accordingly.

ROUGE Score

ROUGE (Lin, 2004) is a modified recall measure based on the n -grams overlap between the generated sequence and one or more reference sequences.

$$ROUGE_n = \frac{\text{number of matching } n\text{-grams in the generated and reference text}}{\text{number of } n\text{-grams in the reference text}}$$

ROUGE-L is an alternative version of ROUGE that takes into account the Longest Common Subsequence (LCS) between the reference and the generated sentence instead of the overlapping n -grams.

BERT Score

BERTScore (Zhang* et al., 2020) is a recently introduced natural language generation metric that takes into account the semantics of words rather than the exact matching to compute the score between the generated sequences and the reference sequences. BERTScore initially calculates the cosine similarity between the token representations in both sequences: Reference $X = x_1, x_2, \dots, x_{n_1}$ of length n_1 and generated $Y = y_1, y_2, \dots, y_{n_2}$ of length n_2 . It then employs a greedy matching approach to pair each token with its most similar counterpart in the other sequence. The token representations are obtained using pre-trained language models such as BERT.

BERTScore computes a precision, recall and F_1 -score as follows:

$$P_{BERT} = \frac{1}{n_1} \sum_{x_i \in X} \max_{y_j \in Y} x_i^T y_j; R_{BERT} = \frac{1}{n_2} \sum_{y_i \in Y} \max_{x_j \in X} y_i^T x_j$$

$$F1_{BERT} = 2 \times \frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}}$$

V-measure

V-measure (Rosenberg and Hirschberg, 2007) assesses the quality of a clustering solution by explicitly measuring its homogeneity and its completeness. Homogeneity h refers to the degree to which each cluster consists of data points that belong primarily to a single gold standard (GS) class, while completeness c refers to the degree to which each GS class consists of data points assigned primarily to a single cluster.

Paired F-Score

This metric assesses the quality of a clustering solution by explicitly measuring precision and recall. Precision can be defined as the number of common instance pairs (pairs are formed between instances from the same cluster in the clustering solution and the ones of the same class in the GS) between the two sets (clustering solution and GS) to the total number of pairs in the clustering solution, while recall can be defined as the number of common instance pairs between the two sets to the total number of pairs in the gold standard. Finally, precision and recall are combined to produce the harmonic mean

Fuzzy B-Cubed

B-Cubed (Bagga and Baldwin, 1998) based on precision and recall, which estimate the fit between two clustering solutions, X and Y at the item level. For an item i , precision reflects how many items that share a cluster with i in X appear in its cluster in Y ; conversely, recall measures how many items sharing a cluster in Y with i also appear in its cluster in X . The final B-Cubed value is the harmonic mean of the two scores. To generalize B-Cubed to fuzzy covers, a method inspired by Amigó et al. (2009) is used, it introduces a correctness term C such that:

$$C(i, j, X) = \sum_{k \in I_X(i) \cup I_X(j)} 1 - |w_k(i) - w_k(j)|$$

Where X is the clustering solution, i and j are the different words and w is the weight of the instance to belong to a cluster. Finally, the fuzzy b-cubed will be the F-score where:

$$P = \frac{\text{Min}(C(i, j, X), C(i, j, Y))}{C(i, j, X)}$$

and,

$$R = \frac{\text{Min}(C(i, j, X), C(i, j, Y))}{C(i, j, Y)}$$

Fuzzy Normalized Mutual Information

Mutual information measures the dependence between two random variables. In the context of the evaluation of cluster solutions, mutual information treats sense labels as random variables and measures the level of agreement in which instances are labeled with the same senses (Danon et al., 2005). Formally, mutual information is defined as:

$$I(X : Y) = H(X) - H(X|Y)$$

Where $H(X)$ denotes the entropy of a variable X that represents a partition, that is, the sets of instances assigned to each sense. The mutual information must be normalized to compare between systems with ease, the normalized factor used is $\text{Max}(H(X), H(Y))$ (Vinh, Epps, and Bailey, 2009). To extend the definition of mutual information score to fuzzy cover, a new definition

of mutual information is proposed: represent each cluster X_i as a continuous random variable, with the entire fuzzy cover denoted as the variable $X_{1\dots k}$ and then find:

$$H(X_i) = \sum_{i=1}^n p(w_i) \log_2 p(w_i)$$

and,

$$H(X_k, Y_l) = \sum_{i=1}^n \sum_{j=1}^m p(w_i, w_j) \log_2 p(w_i, w_j)$$

Finally,

$$H(X_k|Y_l) = H(X_k, Y_l) - H(Y_l)$$

where $p(w_i)$ is the probability that an instance is labeled with the rating w_i . Thus, find $I(X; Y)$ in the fuzzy setting.

Language models have proven to be very useful when adapted to specific domains. Nevertheless, little research has been done on the adaptation of domain-specific BERT models in the French language. In this chapter, we focus on creating a language model adapted to the French legal text with the goal of helping law professionals. We conclude that specific tasks derive enhanced benefits from domain-specific language models pre-trained on tailored datasets rather than relying on generic counterparts trained on large amounts of data in terms of both computational power and performance. We explore the use of smaller architectures in domain-specific sublanguages and their benefits for French legal text. We prove that domain-specific pre-trained models are competitive with their equivalent generalized ones in the legal domain in terms of performance while having fewer parameters. Finally, we release JuriBERT (Douka et al., 2021), a new set of BERT models adapted to the French legal domain.

3.1 INTRODUCTION

Domain-specific language models have evolved the way we learn and use text representations in natural language processing. Instead of using general-purpose pre-trained models that are highly skewed towards generic language, we can now pre-train models that better meet our needs and are highly adapted to specific domains, like medicine and law. In order to achieve that, models are trained on large-scale raw text data, which is a computationally expensive step, and then are used in many downstream evaluation tasks, achieving state-of-the-art results in multiple explored domains.

The majority of domain-specific language models and trained word embeddings so far have been applied to the English language. For languages such as French, most existing models are trained on generic datasets. For instance, Abdine et al. (2021) published French word vectors from large-scale generic web content that surpassed previous pre-trained static word embeddings such as the French FastText embeddings (Grave et al., 2018) trained on generic data as well. Furthermore, Martin et al. (2020) introduced CamemBERT, a French monolingual language model that is used for generic everyday text and proved its superiority compared to other multilingual models. Meanwhile, domain-specific language models for French are in the wane. There is an even greater shortage when it comes to the legal field. Sulea et al. (2017) mentioned the importance of using state-of-the-art technologies to support law professionals and provide them with guidance and direction. Given this need, we introduce JuriBERT, a new set of BERT models pre-trained on French legal text. We explore the use of smaller models architecturally when we are dealing with

specific sub-languages, like French legal text. Therefore, we publicly release JuriBERT¹ in four different sizes online.

3.2 RELATED WORK

Previous work on domain-specific text data has indicated the importance of creating domain-specific language models. These models are either adaptations of existing generalized models, for example, BERT-Base by Devlin et al. (2019) trained on general purpose English corpora, or pre-trained from scratch on new data. In both cases, domain-specific text corpora are used to adjust the model to the peculiarities of each domain.

A remarkable example of adapting language models is the research done by Lee et al. (2019) who introduced BioBERT, a domain-specific language representation model pre-trained on large-scale biomedical corpora. BioBERT outperformed BERT and other previous models on many biomedical text mining tasks and showed that pre-training on specific biomedical corpora improves performance in the field. Similar results were presented by Beltagy, Lo, and Cohan (2019) that introduced SciBERT and showed that pre-training on scientific-related corpus improves performance in multiple domains, and by Yang, Uy, and Huang (2020) who showed that FinBERT, pre-trained on financial communication corpora, can outperform BERT on three financial sentiment classification tasks.

Moving on to the legal domain, Bambroo and Awasthi (2021) worked on LegalDB, a DistilBERT model (Sanh et al., 2019) pre-trained on English legal-domain specific corpora. LegalDB outperformed BERT in legal document classification. Elwany, Moore, and Oberoi (2019) also proved that pre-training BERT can improve classification tasks in the legal domain and showed that acquiring large-scale English legal corpora can provide a major advantage in legal-related tasks such as contract classification. Furthermore, Chalkidis et al. (2020) introduced LegalBERT, a family of English BERT models that outperformed BERT in a variety of datasets in text classification and sequence tagging. Their work also showed that an architecturally large model may not be necessary when dealing with domain-specific sublanguages. A representative example is Legal-BERT-Small, which is highly competitive with larger versions of LegalBert. We intend to further explore this theory with even smaller models.

Despite the increasing use of domain-specific models, we have been mainly limited to the English language. In contrast, in the French language, little work has been done on the application of text classification methods to support law professionals, with the exception of Sulea et al. (2017) that managed to achieve state-of-the-art results in three legal domain classification tasks. It is also worth mentioning Garneau et al. (2021) who introduced CriminelBART, a fine-tuned version of BART_{thez} (Eddine, Tixier, and Vazirgiannis, 2020). CriminelBART is specialized in criminal law by using French Canadian legal judgments. Overall, no previous work has adapted a BERT model in the legal domain using French legal text.

1. You can find the models in <http://nlp.polytechnique.fr/resources#juribert>

3.3 COURT OF CASSATION

The Court of Cassation is the highest court in the French judiciary. Sitting in the historic law courts of Paris on the *Île de la Cité* (center of Paris), this institution has a key role: unifying and monitoring the interpretation of law. The Court thus ensures that everyone has equal treatment before the judge. As its decisions establish the main legal principles that structure our society and concern many aspects of our daily lives, the *Cour de cassation* plays an essential role in the functioning of democracy.

The court is made up of six chambers. The six chambers of the court divide the appeals among themselves according to the legal nature of the disputes to be decided. Each chamber is subdivided into sections, which are, in turn, specialized, and into which its various responsibilities are divided. The judges who sit in the chambers are called councilors. The First President assigns the councilors to the chambers, taking into account the volume of litigation to be handled, but also directing the best specialists to one or another of the chambers. Each chamber is headed by a president. The Public Prosecutor's Office assigns a senior public prosecutor to each chamber. Clerks are assigned to each chamber. The Bureau of the court of cassation, made up of the First President, the Chamber Presidents, the Public Prosecutor, and three First Advocates General, may decide to extend the jurisdiction of the civil chambers. Chamber hearings are open to the public. The chambers are the following:

1. The first civil chamber: rules in particular on disputes relating to personal and family law, consumer protection, associations, movable property, intellectual property, private, international law, etc.
2. The second civil chamber: settles disputes concerning civil procedure, social security, over-indebtedness, lawyers' fees, elections, etc.
3. The third civil chamber: settles disputes concerning real estate, construction, co-ownership, residential leases, environment, and pollution...
4. The Commercial, Financial, and Economic Chamber: settles disputes in the areas of banking, stock exchange, credit insurance, competition, goodwill, transport of goods, collective proceedings, industrial property (patents, trademarks)...
5. The social chamber: rules on disputes concerning labor law, employment and training, collective labor relations, staff representation, termination of employment, etc.
6. The criminal chamber: decides disputes relating to: Crimes, misdemeanors, contraventions, criminal procedure, execution of sentences...

More details about these chambers could be found on the court of cassation website².

Multiple types of textual data from the court of cassation in order to train JuriBERT as detailed in the following section. In addition, the pleadings

2. <https://www.courdecassation.fr/la-cour/lorganisation-de-la-cour-de-cassation/les-six-chambres-de-la-cour-de-cassation>

documents with their assignment to four chambers of the court of cassation: The first civil chamber, the second civil chamber, the third civil chamber and the commercial chamber are used as a downstream task as described in section 3.5

3.4 JURIBERT

We introduce a new set of BERT models pre-trained from scratch in legal-domain specific corpora. We train our models on the Masked Language Modeling (MLM) task. This means that given an input text sequence, we mask tokens with probability 15% and the model is then trained to predict these masked tokens. We follow the example of Chalkidis et al. (2020) and choose to train significantly even smaller models, including Bert-Tiny and Bert-Mini. The architectural details of the models we pre-trained are presented in Table 3.1. We also choose to further pre-train *CamemBERT_{BASE}* on French legal text in order to better explore the impact of using domain-specific corpora in pre-training.

Training Data

For the pre-training we used two different datasets of French legal text. The first dataset contains data crawled³ from the *Légifrance*⁴ website and consists of a raw French legal text. The *Légifrance* text is then cleaned from non-French characters. We also use 253,194 court decisions from different courts, such as the court of cassation, the courts of appeal, and *the counsel of Prud'hommes*. In addition, we include the claimant pleadings of the court of cassation that consist of 123,361 long documents from different court cases. All personal and private information, including names and organizations, has been removed from the documents for the privacy of stakeholders. The combined datasets provide us with a collection of raw French legal text of size 6.3 GB that we will use to pre-train our models.

Legal Tokenizer

In order to pre-train a new BERT model from scratch, we need a new Tokenizer. We trained a ByteLevelBPE tokenizer with newly created vocabulary from the training corpus. The vocabulary is restricted to 32,000 tokens in order to be comparable to the CamemBERT model from Martin et al. (2020) and minimum token frequency of 2. We used a RobertaTokenizer as a template to include all the necessary special tokens for a Masked Language Model. Our new Legal Tokenizer encodes the data using 512-sized embeddings.

3. We used Heritrix, a crawler that respects the robots.txt exclusion directives and META nofollow tags. See <https://github.com/internetarchive/heritrix3>

4. <https://www.legifrance.gouv.fr/>

Model	Architecture	Number of Parameters
JuriBERT-Tiny	$L=2, H=128, A=2$	6M
JuriBERT-Mini	$L=4, H=256, A=4$	15M
JuriBERT-Small	$L=6, H=512, A=8$	42M
JuriBERT-Base	$L=12, H=768, A=12$	110M
JuriBERT-FP	$L=12, H=768, A=12$	110M

Table 3.1 – Architectural comparison of JuriBERT models. Where L is the number of transformer layers, H is the embedding dimension, and A_4 is the number of attention heads.

Pretraining Details

For the pretraining of the JuriBERT model, we used both the crawled *Légifrance* data and the Pleadings dataset, thus creating a 6.3GB collection of legal texts. The encoded corpus was then used to pre-train a BERT model from scratch. Our model was pre-trained in four different architectures. As a result, we have JuriBERT_{TINY} with two layers, 128 hidden units and two attention heads (6M parameters), JuriBERT_{MINI} with four layers, 256 hidden units and four attention heads (15M parameters), JuriBERT_{SMALL} with six layers, 512 hidden units and eight attention heads (42M parameters) and JuriBERT_{BASE} with 12 layers, 768 hidden units and 12 attention heads (110M parameters). JuriBERT_{BASE} uses the exact same architecture as CamemBERT_{BASE}.

JURIBERT-FP Apart from pre-training from scratch, we also decided to further pre-train CamemBERT_{BASE} on the training data. Our goal is to compare its performance with the *from scratch* JuriBERT models to further explore the impact of using specific-domain corpora during pre-training. JuriBERT-FP uses the same architecture and number of parameters as CamemBERT_{BASE} and JuriBERT_{BASE}.

TECHNICAL DETAILS All the models were pre-trained for 1M steps. A learning rate of $1e-4$ was used along with an Adam optimizer ($\beta_1=0.9, \beta_2=0.999$) with weight decay of 0.1 and a linear scheduler with 10,000 warm-up steps. All models were pre-trained with batch size of eight for JuriBERT_{BASE} and JuriBERT-FP that used batches of size of four. For the pre-training, we used an Nvidia GTX 1080Ti GPU with 11GB of memory.

3.5 DOWNSTREAM EVALUATION TASKS

In order to evaluate our models we will be using two legal text classification tasks provided by the court of cassation (in French: *cour de cassation*), the highest court of the French judicial order.

The subject of the first task is assigning the court claimant pleadings (in French *Mémoires ampliatifs*), to a chamber and a section of the court. This leads

Model	Pre-training Corpora	Number of Parameters
CamemBERT _{BASE}	138GB	110M
BARThez	66GB	165M
JuriBERT	6.3GB	6M, 15M, 42M, 110M
JuriBERT-FP	(+)6.3GB	110M

Table 3.2 – Size of pre-training corpora used by different models. CamemBERT_{BASE}, JuriBERT models and JuriBERT-FP are encoder-only models, while BARThez is a pre-trained encoder-decoder model.

to a multiclass classification task with eight different imbalanced classes. In Table 3.3 we can see the eight classes that correspond to the different chambers and sections of the Court, as well as their support in the data. The classes represent four chambers: the first civil chamber (C1) that deals with topics like Civil Contract Law and Consumer Law, the second civil chamber (C2) with topics like Insurance Law and Traffic accidents, the third civil chamber (C3) dealing with Real property and Construction Law among other topics and the Commercial, Economic and Financial Chamber (CO) for Commercial Law, Banking and Credit Law and others. Each chamber has two or more sections dealing with different topics.

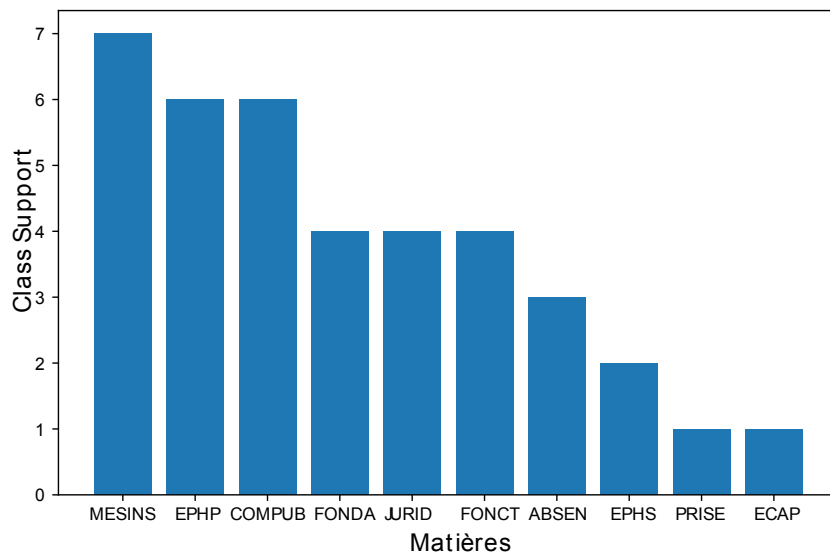


Figure 3.1 – Ten recessive *matières* with the least number of examples in the test dataset. The *matières* are represented by their ID. Where MESINS=Mesures d’instruction, EPHP=Elections aux conseils de prud’homme, COMPUB=Commandes publiques, FONDA=Fondation, JURID=Juridictions, FONCT=Fonctionnaires et agents publics, ABSEN=Absence, EPHS=Elections aux conseils de prud’homme, PRISE=Prise à partie, and ECAP=Elections aux chambres d’agriculture.

The second task is to classify the claimant’s pleadings into a set of 151 subjects (*matières* in French). Figure 3.2 shows the support of the *matières* in the data. As we can see in Figure 3.1, the ten recessive *matières* have between seven and one examples in our data set. We decided to remove the last three *matières* as they have less than three examples and therefore it is not possible to split them in train, test, and development sets.

Class	Support
Commercial Chamber (CO)	28 198
First Civil Chamber, Section 1 (C1_S1)	14 650
First Civil Chamber, Section 2 (C1_S1)	16 730
Second Civil Chamber, Section 1 (C2_S1)	11 525
Second Civil Chamber, Section 2 (C2_S2)	9 975
Second Civil Chamber, Section 3 (C2_S3)	13 736
Third Civil Chamber, Section 1 (C3_S1)	16 176
Third Civil Chamber, Section 2 (C3_S2)	12 282

Table 3.3 – Chambers and sections of the court of cassation, their data support and some of their subjects. The dataset does not differentiate between the different sections of the commercial chamber.

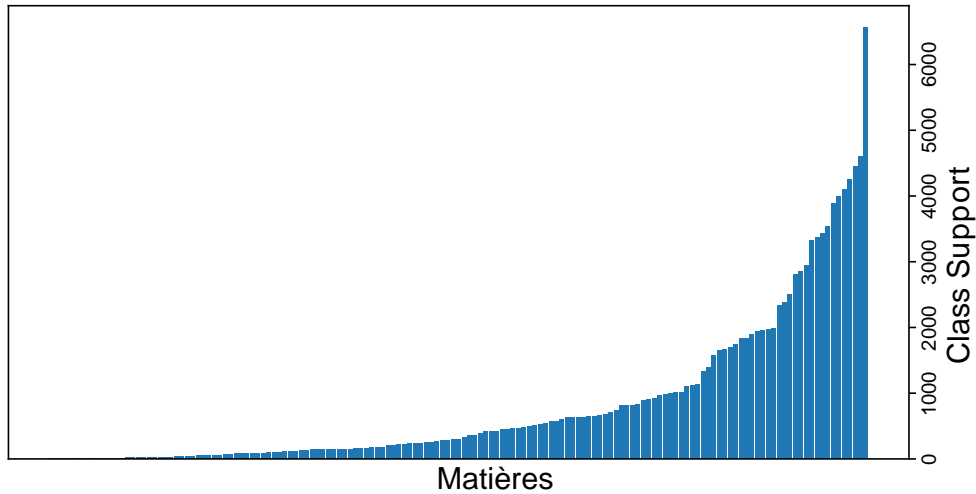


Figure 3.2 – Distribution of the 151 *matières* in the court of cassation data. The distribution reveals a significant imbalance of data among the classes.

FINE-TUNING DETAILS Our models were fine-tuned for the downstream evaluation task using the same classification head as Devlin et al. (2019), which consists of a dense layer with function *tanh* followed by a dense layer with softmax activation function and dropout layers with fixed dropout rate of 0.1. We

applied grid search to the learning rate in a range of $\{2e-5, 3e-5, 4e-5, 5e-5\}$. We used an Adam optimizer along with a linear scheduler that provided the training with 100k warm-up steps. We train for a maximum of 30 epochs with patience of two epochs on the early stopping callback and checkpoints for the best model. For the classification, we use only the paragraphs starting with 'ALORS QUE' from the pleadings dataset, as they include all the important information for the correct chamber and section. This was suggested by a legal expert from the court of cassation, as the average size of a *mémoire ampliatif* is extremely large, from 10 to 30 pages long. By using the 'ALORS QUE' paragraphs, we have text sequences with an average size of 800 tokens. For the chambers and sections classification task, we split the data into 14% development and 16% test data. For the *matières* classification, we split the data into 17% development and 14% test data and stratify in order to have all classes represented in each subset. Both tasks use a fixed batch size of 4. For the fine-tuning, we used an Nvidia GTX 1080Ti GPU.

3.6 RESULTS

The results of the downstream evaluation tasks are presented in Tables 3.4 and 3.5. We compare our models with CamemBERT_{BASE}, and with BART_{hez}, a sequence-to-sequence model dedicated to the French language. CamemBERT_{BASE} has been pre-trained on 138GB of French raw text from the OSCAR corpus. Despite the difference in the size of the pretraining corpora, with our model using only 6.3GB of legal text, JuriBERT_{SMALL} managed to outperform the bigger CamemBERT_{BASE} model. This further shows the importance of domain-specific language models in natural language processing and transfer learning. Despite our expectations, the performance of JuriBERT_{BASE} does not exceed the performance of its smaller equivalent models. We attribute this peculiarity to the fact that larger models usually need more computational resources and more time and data to converge. A recent study (Hoffmann et al., 2022) published approximately one year after the pre-training of JuriBERT models, proved that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size, the number of training tokens should also be doubled by training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens. They also proved that an **oversized** model performs worse than smaller models that are trained on a suitable number of tokens. They additionally found that for the English language, the optimal number of tokens to train one parameter is around 20 trainable tokens. After examination of our pre-training corpora, it contains 1.08 billion tokens. On the other hand, depending on the study of Hoffmann et al. (2022), JuriBERT_{BASE} needs approximately 2.2B tokens in the training set to be optimally pre-trained. This explains why JuriBERT_{SMALL} outperforms JuriBERT_{BASE} since according to the same formula, JuriBERT_{SMALL} requires around 0.84B tokens to be optimally pre-trained. We note that there is no equivalent study for the French language.

Model	Lrate	Dev Accuracy	Test Accuracy
CamemBERT _{BASE}	$2e - 5$	82.75	83.22
BARThez	$3e - 5$	83.70	83.49
JuriBERT _{TINY}	$3e - 5$	82.00	81.58
JuriBERT _{MINI}	$3e - 5$	83.08	82.62
JuriBERT _{SMALL}	$3e - 5$	83.86	83.95
JuriBERT _{BASE}	$3e - 5$	82.26	82.51
JuriBERT-FP	$2e - 5$	83.07	83.28

Table 3.4 – Accuracy of models on the chambers and sections classification task. JuriBERT_{SMALL} has the highest accuracy despite not being the bigger model.

JuriBERT_{SMALL} also outperforms BARThez in the chambers and sections evaluation task, which is pre-trained on 66GB of French raw text and usually used for generative tasks. On the *matières* classification task, BARThez is the dominant model, with JuriBERT_{SMALL} being second.

JuriBERT-FP outperforms JuriBERT_{BASE} and achieves similar results to the base version of CamemBERT on the chambers and sections classification task. This shows that further pre-training a general purpose language model can have better results than training from scratch. However, it did not outperform JuriBERT_{SMALL} in both tasks, which can be attributed to the same small number of training tokens. Unfortunately, there are no smaller versions of CamemBERT available to further test this theory. On the *matières* classification task, JuriBERT-FP still outperforms JuriBERT_{BASE}. On the contrary, it performs worse than CamemBERT_{BASE}. Along with the state-of-the-art results of BARThez, which align with the results of JuriBERT_{SMALL}. This leads us to believe that in order to achieve better results in more complex tasks, JuriBERT models require a bigger pre-training corpus. Overall, JuriBERT_{SMALL} achieves equivalent results with previous larger generic language models with an accuracy of 83.95% for the first task and 71.80% for the second task for the test data. JuriBERT_{SMALL}, JuriBERT_{MINI} and even JuriBERT_{TINY} all outperform JuriBERT_{BASE}, proving that smaller models architecturally can achieve comparable, if not better, results when we are training on a small domain-specific dataset. A larger model not only requires more resources to be trained, but is also not as efficient as its smaller equivalents when the number of training tokens is small. This is of great importance for researchers with limited resources available. Furthermore, JuriBERT-FP achieves better results than JuriBERT_{BASE} in both tasks. This leads us to infer that pre-training from an existing language model can be a major advantage, as opposed to randomly initializing the model’s weights. More detailed results are presented in the Appendix a.

Model	Lrate	Dev Accuracy	Test Accuracy
CamemBERT _{BASE}	$3e - 5$	71.64	71.66
BARThez	$2e - 5$	72.17	72.09
JuriBERT _{TINY}	$2e - 5$	61.36	61.48
JuriBERT _{MINI}	$2e - 5$	70.01	70.41
JuriBERT _{SMALL}	$2e - 5$	71.67	71.80
JuriBERT _{BASE}	$3e - 5$	70.28	70.38
JuriBERT-FP	$2e - 5$	70.99	71.21

Table 3.5 – Accuracy of models on the *matières* classification task.

3.7 LIMITATIONS

JuriBERT_{BASE} and JuriBERT-FP have been pre-trained using smaller batch sizes than the other models due to limited resources. We acknowledge that this may have affected their performance compared to the other models. However, we believe that their lower performance can also be attributed to their size, as larger models are computationally heavier and thus require more resources to converge.

Acquiring large-scale legal corpora, especially for a language other than English, has proven to be challenging due to their confidential nature. For this reason, JuriBERT models were fine-tuned on two downstream evaluation tasks that contain data from the pre-training dataset collection. Further testing shall be required to validate the performance of our models on different tasks.

The differences in performance between the generic language models and the newly created JuriBERT models are very small. More specifically, only JuriBERT_{SMALL} manages to outperform CamemBERT_{BASE} and Barthez with a difference in accuracy of 0.73%. We attribute this limitation to the use of much fewer pre-training data. However, we emphasize that JuriBERT manages to achieve similar results despite the difference in pre-training corpora size. Thus, we expect JuriBERT to achieve better results in the future, provided that we further pre-train with more data.

3.8 CONCLUSIONS AND FUTURE WORK

We introduce a new set of domain-specific BERT models pre-trained from scratch on French legal text. We conclude that specific tasks derive enhanced benefits from domain-specific language models pre-trained on tailored datasets rather than relying on generic counterparts trained on large amounts of data in terms of both computational power and performance. We also show the superiority of much smaller models when training on small specific sub-language corpus like the French legal text that contains only one billion tokens. It becomes apparent that large architectures may, in fact, not be necessary when such domains are targeted. This is important for researchers with lower resources

available, as smaller models are fine-tuned much more quickly on downstream tasks. In future work, we plan to further explore the potential of JuriBERT in other tasks and, as a result, prove its superiority over the task-specific one. In addition, since we have our data as documents and not as sentences, it would be beneficial to train a BART model for the French legal domain. The French legal BART model can mainly perform generative tasks such as summarizing pleading and text cases that can be too long or even extract useful information from big legal text.

The era of transfer learning has revolutionized the fields of computer vision and natural language processing, bringing powerful pre-trained models with exceptional performance in a variety of tasks. Specifically, Natural Language Processing tasks have been dominated by transformer-based language models. In the Natural Language Inference and Natural Language Generation tasks, the BERT model and its variants, as well as the GPT model and its successors, demonstrated exemplary performance. However, the majority of these models are pre-trained and evaluated primarily for the English language or on a multilingual corpus. In this chapter, we introduce GreekBART, the first Seq2Seq model based on the BART base architecture and pre-trained on a large-scale Greek corpus. We evaluated and compared GreekBART with BART random, Greek-BERT, and XLM-R on a variety of discriminative tasks. In addition, we examine its performance on two NLG tasks from GreekSUM, a newly introduced summarization dataset for the Greek language. The model, the code, and the new summarization dataset are publicly available.

4.1 INTRODUCTION AND RELATED WORK

The field of machine learning has entered a new era with the establishment of transfer learning, providing new possibilities, especially in the areas of Computer Vision (Krizhevsky, Sutskever, and Hinton, 2017) and Natural Language Processing. Transfer learning has become a new trend that is so rare to train a model for computer vision or natural language processing tasks from scratch, dealing with the issue of insufficient training data for real-world machine learning applications. Tasks are solved by reusing pre-trained models which are trained on enormous amounts of data, and the resulting models have reached state-of-the-art performance. Transformer (Vaswani et al., 2017) based pre-trained models, as BERT (Devlin et al., 2019) and its variants, are broadly used in Natural Language Processing, as have been shown to be effective in many tasks.

BART (Lewis et al., 2020) is a denoising auto-encoder for pre-training sequence-to-sequence models. It is trained by corrupting the text with an arbitrary noising function and learning a model to reconstruct the original text. It uses a standard transformer-based neural machine translation architecture and a standard seq2seq architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT (Radford et al., 2018)). This means that the encoder’s attention mask is fully visible, like BERT, and the decoder’s attention mask is causal, like GPT-2 (Radford et al., 2019). The unsupervised pre-trained BART learns a language model, giving us the possibility to adapt it to a particular NLP task. Therefore, large-scale labeled data sets are not

required for fine-tuning. This type of model is suitable for machine translation, question-answering, and especially, text summarization tasks, but that does not mean that BART is insufficient in sequence classification tasks; on the contrary, it is also quite effective in that type of tasks.

In the last few years, a lot of research has been conducted on other languages, except for the English language. For example, CamemBERT (Martin et al., 2020) and BARThez (Kamal Eddine, Tixier, and Vazirgiannis, 2021) for the French language, CAMELBERT (Inoue et al., 2021) and AraBART (Eddine et al., 2022) for Arabic language, BART for Japanese language (Kim and Komachi, 2021), BETO (Cañete et al., 2020) and NASes (Ahuir et al., 2021) for Spanish and Catalan languages, and BARTpho (Tran, Le, and Nguyen, 2021) for Vietnamese language. Recently, a variety of multilingual language models have been presented, covering multiple languages by being pre-trained on a large-scale corpus of different languages, trying to learn the language model of multiple languages at once. In particular, M-BERT (Devlin et al., 2019) is a case of a multilingual pre-trained language model, which consists of the multilingual version of BERT, pre-trained in the top 100 languages with the largest Wikipedias. Another case of a popular multilingual model is XLM (Conneau and Lample, 2019), which is a transformer-based multilingual language model pre-trained on Wikipedias of 15 languages. This model was trained in two auxiliary tasks, Masked Language Modeling and the Translation Language Modeling task. Training a cross-lingual language model can be very beneficial for low-resource languages, as all languages are processed with the same shared vocabulary. Conneau et al. 2020 introduced XLM-R, an improved version of XLM based on the RoBERTa model. The model was trained with a cross-lingual masked language modeling objective on 2.5TB data in 100 languages from Common Crawl (Conneau et al., 2020; Wenzek et al., 2020), increasing the amount of training available data for low-resource languages by two orders of magnitude on average. Finally, mBART (Liu et al., 2020) is the multilingual version of BART and is pre-trained on a subset of 25 languages from the same data set as XLM-R. In mBART, we use its 250K sentencepiece (Kudo and Richardson, 2018) model, which was trained using monolingual data for 100 languages from XLM-R, supporting languages beyond the original 25 mBART was trained on. The parameters of mBART25 are roughly 610M. Later, an extension of mBART was proposed in additional 25 languages (*e.g.* total 50 languages) mBART50 (Tang et al., 2020), increasing the number of parameters to approximately 680M. Except for mBART and mBART50, all other aforementioned multilingual models support the Greek language. mBART25 and mBART50 are not pre-trained on modern Greek, but it is included in their vocabulary. Nevertheless, multilingual models cannot compete with the performance of monolingual models in most NLP tasks. In recent months, another BART related model that is in the spotlight in the NLP research area is ChatGPT¹. ChatGPT is built on top of GPT-3 architecture (Brown et al., 2020), so it is a transformer-based language model that has been pre-trained on massive amounts of text data and fine-tuned for conversational AI applications. Like BART, ChatGPT is capable of generating high-quality sequences of text,

1. <https://openai.com/blog/chatgpt>

making it suitable for tasks such as text summarization and question answering. However, unlike BART, ChatGPT is specifically designed for conversational applications, making it well-suited for chatbots and other dialogue systems. In addition, ChatGPT’s architecture is unidirectional, which means that it can generate text in a left-to-right sequence, making it more suitable for tasks such as language generation and dialogue.

Compared to languages that are widely spoken, Greek has fewer linguistic resources available. Especially, the available research on deep learning models for Greek is still very undeveloped. However, there have been some efforts to develop datasets, models, knowledge bases, and frameworks for Greek NLP. Outsios et al. 2018 presented the production of Greek word embeddings, where a large corpus of about 50GB (contains 120 million sentences), crawled from about 20 million URLs, was used for their work. Later, Lioudakis, Outsios, and Vazirgiannis 2020 presented an ensemble method, continuous bag-of-skip grams, to extract word representations for Greek. Recently, Koutsikakis et al. 2020 used Greek-BERT, the first transformer-based language model, based on BERT, for the Greek language. The model was pre-trained on a 29GB dataset, achieving state-of-the-art performance in several NLP tasks in Greek. It is worth noting that Papantoniou and Tzitzikas 2020 have provided a complete survey of the work that has been performed in NLP for the Greek language.

In this contribution, we try to handle the issue that the multilingual models are not sufficient to compete with the monolingual ones and the limited available deep learning models for the Greek language. Thus, we propose the first pre-trained Seq2Seq monolingual model for the Greek language. The model is called GreekBART, as we pre-trained the BART-base architecture on a large monolingual Greek corpus. Despite the existence of the Greek-BERT (Koutsikakis et al., 2020), our model exceeds the possibilities of the Greek-BERT, focusing on generative tasks. GreekBART is evaluated on two different generative tasks and on four discriminative tasks. Our main contributions are as follows.

- We introduce the pre-trained Seq2Seq model for the Greek language, based on the BART-base architecture (Lewis et al., 2020), and pre-trained on a large corpus of 87.6 GB. We examine the performance of our model in four discriminative tasks (*i.e.* two classification tasks, one sentiment analysis task and one natural language inference task) and in two generative tasks.
- We present the first summarization dataset in Greek, GreekSUM, introducing two generative tasks and a classification task by processing this dataset.
- We compare GreekBART against popular language models, already pre-trained or not on Greek. In the case of the discriminative tasks, we collate our model, a BART-random model, Greek-BERT (Koutsikakis et al., 2020) and XLM-R (Conneau et al., 2020). We also inspect the differences, in terms of performance, between the GreekBART (*i.e.* our model), BART-random model, mBART25 (Liu et al., 2020) and mBART50 (Tang et al., 2020) on two novel generative tasks.

- We publish our code and models², providing access to everyone who wants to further extend the applications of our work or take advantage of our contributions in favor of his/her work.

4.2 GREEKBART

Our proposed model is based on BART (Lewis et al., 2020) a denoising auto-encoder. We use the *BASE* architecture, with six encoder and six decoder layers. In addition, 768 hidden dimensions are used, twelve attention heads in both the encoder and the decoder, and a normalization layer is added on top of both the encoder and the decoder (Liu et al., 2020). The purpose of these additional layers is to stabilize the training when FP16 precision (Micikevicius et al., 2017) is applied. The use of FP16 precision speeds up the pre-training of the model. In total, our model has roughly 181M parameters. Generally, we follow a similar methodology as Kamal Eddine, Tixier, and Vazirgiannis 2021, in which a monolingual model in a language different from English is pretrained, following the methodologies of BART (Lewis et al., 2020) and mBART (Liu et al., 2020).

4.2.1 Pre-training Corpus

The pre-trained corpus is produced by the following corpora: (a) the Greek part of Wikipedia³; (b) the Greek part of the European Parliament Proceedings Parallel Corpus (EuroParl)⁴ (Koehn, 2005); (c) the Greek part of OSCAR⁵ (Abadji et al., 2022), a clean version of CommonCrawl⁶; (d) the Greek Web Corpus, crawled from about 20 million Greek-language URLs⁷ (Ousios et al., 2018). In particular, we use the same datasets as the Greek-BERT (Koutsikakis et al., 2020) model, including also the dataset of Ousios et al. 2018 in order to have a larger corpus that will be well suited for the pre-training of BART model. Moreover, by choosing these datasets we cover a wide variety of Greek language areas, including formal and informal text, news articles, encyclopedic information, and political conversations. This diverse range of text types helps to ensure that the pre-training of the BART model is robust and able to handle different styles and registers of Greek language use. Overall, the choice of datasets helps to ensure that the Greek BART model is well-equipped to handle a wide range of natural language processing tasks in the Greek language.

We preprocessed each of the aforementioned corpora by removing URLs, emojis, tags, and hashtags. Also, we erase comments and some observed noisy sentences which do not provide any additional contextual meaning. The noisy sentences differ from dataset to dataset, so we had to detect them "manually". Furthermore, for all corpora except Wikipedia's dataset, we got rid

2. <https://github.com/iakovosevdaimon/GreekBART>

3. <https://dumps.wikimedia.org/elwiki/>

4. <https://www.statmt.org/europarl/>

5. <https://oscar-corpus.com/>

6. <https://commoncrawl.org/>

7. <http://nlp.polytechnique.fr/resources-greek>

Corpus	Size before deduplication	Size after deduplication
OSCAR	51.7	44.6
Greek Web Corpus	38.4	30.9
Wikipedia	0.9	0.9
EuroParl	0.5	0.5
Total	91.5	76.9

Table 4.1 – Datasets which consists of the GreekBART pre-training corpus (sizes in GB, before and after cleaning and deduplication).

of documents that contained less than one thousand characters. In the case of Wikipedia, we removed documents with fewer than 30 characters. Generally, we did not remove non-Greek characters, because we supposed that it will not prevent the GreekBART from understanding the language model, as their amount is insignificant. We deduplicated each corpora and then concatenated all of them in one corpus. Again, we duplicated the merged dataset for the final time. The deduplication process was done using the Runiq package⁸. To generate our vocabulary, we used SentencePiece⁹ (Kudo and Richardson, 2018) which implements byte pair encoding (BPE) (Sennrich, Haddow, and Birch, 2016). So, any type of pre-tokenization was not necessary. We fixed the size of the vocabulary to 50K subwords, and the SentencePiece model was trained on a 20GB random sample of the pre-training corpus. We set the character coverage at 99.95%. The total corpus size was 76.9/87.6GB before/after SentencePiece tokenization.

4.2.2 Training details

We adhere to the same pre-training process as BART. Thus, GreekBART tries to reconstruct the corrupted input by minimizing the loss of cross-entropy between the decoder output and the original input. Two types of noise are applied in the input text. First, we employ the text fill technique, where a number of text spans are replaced by a special token, called [MASK], masking 30% of text. A Poisson distribution with ($\lambda = 3.5$) is used to determine the length of the spans. Sentence permutation is the second perturbation method, where the sentences of the input document are shuffled randomly. We pre-trained GreekBART on Jean Zay, using a batch size equal to 768,000 tokens per GPU, as we set the update frequency to 128. We used Adam Optimizer (Kingma and Ba, 2015) with $\epsilon = 10^{-6}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, with a learning rate starting from 6.10^{-4} and decreasing linearly as a function of the training step. We use a warm-up of 6% of the total number of training steps. In the first 12 epochs, we fixed the dropout to 0.1, for epochs 12 to 16 we decreased it to

8. <https://github.com/whitfin/runiq>

9. <https://github.com/google/sentencepiece>

Dataset	train/val/test	avg. doc length		avg. summary length		vocabulary size	
		words	sentences	words	sentences	docs	summaries
CNN	90.3/1.22/1.09	760.50	33.98	45.70	3.58	34	89
DailyMail	197/12.15/10.40	653.33	29.33	54.65	3.86	564	180
NY Times	590/32.73/32.73	800.04	35.55	45.54	2.44	1233	293
XSum	204/11.33/11.33	431.07	19.77	23.26	1.00	399	81
OrangeSum Title	30.6/1.5/1.5	315.31	10.87	11.42	1.00	483	43
OrangeSum Abstract	21.4/1.5/1.5	350	12.06	32.12	1.43	420	71
GreekSUM Title	146.046/10/10	355.49	14.26	9.95	1.05	663	91
GreekSUM Abstract	129.159/10/10	368.97	14.76	24.55	1.46	629	127

Table 4.2 – Sizes (column 2) are given in thousands of documents. Document and summary lengths are in words, while vocabulary sizes are in thousands of tokens.

Dataset	% of novel n-grams in gold summary				LEAD			EXT-ORACLE		
	unigrams	bigrams	trigrams	4-grams	R-1	R-2	R-L	R-1	R-2	R-L
CNN	16.75	54.33	72.42	80.37	29.15	11.13	25.95	50.38	28.55	46.58
DailyMail	17.03	53.78	72.14	80.28	40.68	18.36	37.25	55.12	30.55	51.24
NY Times	22.64	55.59	71.93	80.16	31.85	15.86	23.75	52.08	31.59	46.72
XSum	35.76	83.45	95.50	98.49	16.30	1.61	11.95	29.79	8.81	22.65
OrangeSum Title	26.54	66.70	84.18	91.12	19.84	08.11	16.13	31.62	17.06	28.26
OrangeSum Abstract	30.03	67.15	81.94	88.3	22.21	07.00	15.48	38.36	20.87	31.08
GreekSUM Title	26.7	67.9	84.5	91.4	14.68	04.46	14.37	23.36	07.39	23.12
GreekSUM Abstract	20.6	50.8	65.3	73.0	17.11	06.17	16.69	34.18	14.17	33.93

Table 4.3 – Degree of abstractiveness of GreekSUM compared with that of other datasets. It shows that GreekSUM follows XSum and OrangeSum, being more abstractive than traditional summarization datasets.

0.05, and finally we set it to zero for epochs 16 to 20. We did all the experiments using the Fairseq library¹⁰ (Ott et al., 2019).

4.3 GREEKSUM

Transformer-based Seq2Seq models, including BART, can also perform not only extractive but abstractive summarization. This type of summarization is one of the most central and challenging evaluation tasks in NLP. However, there is no available summarization dataset for the Greek language. Therefore, we created the first dataset in the Greek language, well-suited to the abstractive summarization task.

4.3.1 Motivation

Our main goal was to create a Greek version equivalent to the OrangeSum dataset¹¹ (Kamal Eddine, Tixier, and Vazirgiannis, 2021) and the XSum dataset (Narayan, Cohen, and Lapata, 2018). OrangeSum was produced by scraping articles, their single-sentence title, and their brief abstract from the "Orange

10. <https://github.com/facebookresearch/fairseq>

11. <https://github.com/Tixierae/OrangeSum>

Actu" website¹². The title and abstract of each article are written by the author of the article. Well-performed models on OrangeSum, as well as XSum, require a high degree of abstractiveness.

4.3.2 Data collection

We followed a similar approach, scraping the "News24/7" website¹³. News24/7 is one of the leading news websites in Greece, part of the 24 MEDIA digital publishing group¹⁴. We collected data from web pages that span from October 2007 to June 2022, covering five major categories: politics, society, economy, culture, and the world. Each article had a one-sentence title and a succinct abstract, features which were extracted, yielding two summarization tasks. GreekSUM Title and GreekSUM Abstract. The average length of these two novel tasks' gold summaries is 9.95 and 24.55 words, respectively (see Table 4.2).

4.3.3 Post-processing

Initially, we filtered the scrapped pages, removing all empty articles and articles whose titles were shorter than 2 words or whose abstracts were less than 5 words. Secondly, we filtered the duplicated articles (*i.e.* articles with the same body, or with the same title, or with the same abstract), since an article can belong to more than one category, and thus can be crawled multiple times. Finally, we noticed that several abstracts looked more like introductions than actual summaries of the article. Therefore, we eliminated 10% of the articles with the highest proportion of novel unigrams in the abstracts. This corresponded to a threshold of 46.7% novel unigrams. For both proposed summarization tasks, we reserved 10k pairs for testing, 10k for validation, and all the remaining pairs for training. The GreekSUM dataset released can be reproduced by using our code¹⁵.

4.3.4 Analysis

In Table 4.2 we compare GreekSUM with OrangeSum, XSum, and the well-known CNN, DailyMail, and NY Times datasets (Hermann et al., 2015). We can observe that GreekSUM and OrangeSum datasets are very equivalent in terms of average documents and summaries length. In addition, GreekSUM has a scale similar to XSum. Inspecting Table 4.3, it is noticeable that extractive methods (*i.e.* LEAD and EXT-ORACLE) do not perform so well in GreekSUM; thus our dataset is less biased towards extractive models. Due to the poor performance of the two extraction methods, it seems that GreekSUM is more abstractive than the traditional summarization datasets (*i.e.* CNN, DailyMail, NY Times).

12. <https://actu.orange.fr/>

13. <https://www.news247.gr/>

14. <https://www.24media.gr/>

15. <https://github.com/iakovosevdaimon/GreekSUM>

However, the summaries and titles of GreekSUM do not display a degree of novelty as high as those of OrangeSum and XSum. In the GreekSUM dataset, there are 20.6% novel unigrams in the abstracts and 26.7% novel unigrams in the titles compared to 30% in the OrangeSum Abstract, 26.5% in the OrangeSum Title, and 35.7% in XSum. Therefore, we can conclude that the summaries of GreekSUM are not as abstract as we would like them to be.

4.4 EXPERIMENTS

In this section, we present the results of all experiments. Basically, we have two types of downstream tasks, discriminative tasks and summarization tasks. In the case of discriminative tasks, we compare GreekBART to BART-random, Greek-BERT (Koutsikakis et al., 2020), and the XLM-R model (Conneau et al., 2020). Except for BART-random, the other models are already pre-trained on the Greek language. So, we evaluate the performance of our model against the current state-of-the-art monolingual model pre-trained only on the Greek language as well as against a widely used multilingual model. We fine-tuned all the above-mentioned models on the downstream tasks.

For the summarization task, we set side by side the GreekBART, the BART-random and the two versions of mBART (Liu et al., 2020; Tang et al., 2020). mBART₂₅ and mBART₅₀ are built upon the *LARGE* architecture of BART, and they are pre-trained on 25 and 50 languages, respectively, excluding the Greek language. Therefore, we performed zero-shot learning for the summarization task. On the other hand, the BART-random model uses the same architecture and vocabulary as GreekBART; however, it is trained from scratch on the downstream tasks.

4.4.1 Discriminative tasks

Except for generative tasks, the BART model also achieves remarkable results in discriminative tasks (Lewis et al., 2020). In the case of sequence classification, a classification head is added on top of the model and the input is fed into both the encoder and the decoder. The representation of the final decoder token is used by the newly introduced multiclass linear classifier. We examined the performance of the models (*i.e.* Greek-BERT, XLM-R, BART-random, GreekBART) on four discriminative tasks. More precisely, we evaluated our model on two classification tasks, one task of sentimental analysis, and a Natural Language Inference (NLI) task.

4.4.1.1 Training details

In all experiments, we fine-tuned the models with a learning rate chosen from $\{10^{-4}, 5 \cdot 10^{-5}, 10^{-5}\}$, based on the best validation score. We repeat each experiment three times with different seeds and record the mean and standard deviation of their accuracy on the test set for each aforementioned task.

4.4.1.2 NCC task (*News Category Classification task*)

For the first classification task, we used the novel summarization dataset (GreekSum, see section 4.3) which we scraped from the news website News24/7¹⁶. We considered the five distinct subjects that an article may fall into politics, society, economy, culture, and world. These categories serve as labels for the classification task that our model is trained to perform. Essentially, the model is fed the content of an article and learns to predict which category it belongs to (*i.e.* subject). We fine-tuned all examined models for 5 epochs, using a batch size equal to 32. For the XLM-R model, we set the learning rate to 5.10^{-5} while for the rest of the models, the learning rate is equal to 10^{-4} . The training set consists of 146,046 samples, whereas both the validation and the test set have 10,000 instances exactly like the two summarization datasets (*i.e.* GreekSUM Abstract and GreekSUM Title).

In the second classification task, we used the proposed Greek classification dataset of Lioudakis, Outsios, and Vazirgiannis 2020, which was created from articles from Makedonia newspaper. The dataset contains 8,005 articles from 18 different categories: Sports, Reportage, Economy, Politics, International, Television, Arts-Culture, Letters, Opinions, Interviews, Weather, Society, Advertisements, Biographies, Others, Articles, Police and Zodiacs. We reserved 70% of the dataset for train and the remaining 30% for both validation and test. So, the train set consists of 5,610 samples, whereas the test set and the validation set consist of 1,191 and 1,204 instances, respectively. All models are fine-tuned for 20 epochs, with a batch size of 16 and a learning rate equal to 5.10^{-5} . Due to the small size of the dataset, we trained the models for more epochs and smaller batch sizes.

4.4.1.3 Natural Language Inference

Cross-lingual Natural Language Inference Corpus (XNLI) (Conneau et al., 2018) contains pairs of sentences. The objective of this task is to determine whether the first sentence, also known as the premise, entails, contradicts, or is neutral in relation to the second sentence, referred to as the hypothesis. The XNLI corpus contains 5,000 test and 2,500 validation pairs, and 340k training pairs from the MultiNLI corpus (Williams, Nangia, and Bowman, 2018). The dataset has been translated from English to 14 languages, including Greek. Unfortunately, a large number of the training pairs are of extremely poor quality, as they are produced by machine translation. This condition can affect the performance of models. We fine-tuned for 5 epochs, using 32 batches, and a learning rate equal to 5.10^{-5} .

4.4.1.4 Sentimental Analysis task

We used a publicly available sentimental analysis dataset¹⁷ about Greek movie reviews. We preprocessed the dataset by mainly removing emojis and

16. <https://www.news247.gr/>

17. <https://www.kaggle.com/datasets/nikosfragkis/greek-movies-dataset>

Model	NCC		Sentiment Analysis	XNLI
	News24/7 (ours)	Makedonia (Lioudakis, Outsios, and Vazirgiannis, 2020)		
Greek-BERT	92.61 \pm 0.19	89.45 \pm 0.84	86.39 \pm 0.06	78.6 \pm 0.62
XLM-R	93.1 \pm 0.51	89.6 \pm 0.29	85.43 \pm 0.05	78.2 \pm 0.59
BART-random	91.33 \pm 0.17	80.17 \pm 0.09	80.87 \pm 0.12	60.1 \pm 0.43
GreekBART (ours)	93.2 \pm 0.29	91.1 \pm 0.43	85.43 \pm 0.19	78.67 \pm 0.25

Table 4.4 – Results on discriminative tasks. We present the mean accuracy, as well as the standard deviation.

	GreekSUM Abstract				GreekSUM Title				
	R-1	R-2	R-L	BertScore	R-1	R-2	R-L	BertScore	
LEAD	17.11	06.17	16.69	72.61/63.56	14.68	04.46	14.37	70/57.13	
EXT-ORACLE	34.18	14.17	33.93	73.89/65.43	23.36	07.39	23.12	70.02/57.33	
BASE	BART-random	13.85	04.47	13.65	72.44/63.27	11.55	03.27	11.42	74.47/62.22
	GreekBART (ours)	16.5	06.13	16.21	73.03/ 64.46	15.35	05.02	15.18	75.78/63.98
LARGE	mBART25	15.07	05.8	14.82	72.75/64.08	16.09	05.58	15.93	76.81/65.38
	mBART50	15.53	06.	15.31	73.07/64.43	16.1	05.59	15.96	76.81/65.38

Table 4.5 – Results on GreekSUM. Except for ROUGE, we also provide the BertScore. The left-hand BERTScore has calculated using the M-BERT model (Devlin et al., 2019), while the right-hand one uses the Greek-BERT (Koutsikakis et al., 2020).

hashtags. Each instance consists of a review and a rating. To distinguish between positive and negative reviews, we established a threshold of 3 out of 5. Ratings above this threshold were classified as positive reviews, while those at or below 3 out of 5 were classified as negative reviews. In an effort to create a balanced dataset, we aimed to include a similar number of positive and negative reviews. For the purpose of our task, we only retained the reviews and ratings, excluding any additional information. We divide the data set into train, validation, and test set. The train set consists of 104,157 samples, while the validation and test contain 22,320 and 22,318 instances, respectively. We set the learning rate and the batch size equal to $5 \cdot 10^{-5}$ and 16 respectively. We fine-tuned the models for five epochs.

4.4.1.5 Results

Table 4.4 reports the accuracy on the test set for the four different tasks. We compare our model with Greek-BERT (Koutsikakis et al., 2020), XLM-R (Conneau et al., 2020), and BART-random. For all models, their corresponding *BASE* architecture is used. Among the models, we observe that GreekBART is the best in almost all discriminative tasks, except for the sentiment analysis task, where Greek-BERT achieved the best performance. Generally, it is common for BERT models to perform better than BART models in this kind of task. The performance of our model (*i.e.* GreekBART) verifies the results of the BART paper (Lewis et al., 2020) that models based on that architecture perform well on both generative and discriminative tasks.

	GreekSUM Abstract					GreekSUM Title					
	unigrams	bigrams	trigrams	4-grams	length	unigrams	bigrams	trigrams	4-grams	length	
Gold	20.6	50.8	65.3	73.0	24.55	26.7	67.9	84.5	91.4	9.95	
BASE	BART-random	9.6	43.0	64.5	76.8	20.27	21.6	69.4	89.1	95.8	9.37
	GreekBART (ours)	7.4	23.5	34.5	42.2	23.63	14.9	50.1	69.3	79.9	9.78
LARGE	mBART ₂₅	6.2	20.0	29.4	36.0	26.22	12.8	46.6	65.6	76.2	10.67
	mBART ₅₀	6.5	21.8	32.3	39.7	23.95	12.8	46.6	65.6	76.2	10.67

Table 4.6 – Proportion of novel n-grams in the generated summaries. In addition, the length (number of words) of the generated summaries.

4.4.2 Summarization

We evaluate our model in two distinct summarization tasks, in which the model learns to predict the title and the abstract of an article based on its corresponding content. In both generative tasks, we fine-tuned GreekBART for 30 epochs with a learning rate equal to 5.10^{-5} , which was heated for 6% of the training steps and then linearly decreased to 0. We used the same set of hyperparameters as those of GreekBART to train mBART₂₅ and mBART₅₀. For BART-random, we trained the model for 60 epochs. To produce the summaries of the test set, we used ROUGE-L (Lin, 2004) to select the checkpoint that was associated with the best validation score. In addition, we incorporated two extractive techniques as baselines: EXT-ORACLE and LEAD (Narayan, Cohen, and Lapata, 2018). The LEAD technique generates a summary by extracting the first N sentences from the document, with N set to 1 in our case. On the other hand, EXT-ORACLE selects the set of sentences from the document that maximizes a specific score, with ROUGE-L being the score used in our implementation. In particular, we extracted the one sentence from the document with the highest ROUGE-L score. In Table 4.5, we report the ROUGE-1, ROUGE-2, ROUGE-L scores (Lin, 2004) and two different BERTScores (Zhang et al., 2019b), using the M-BERT model (Devlin et al., 2019) and the Greek-BERT model to calculate contextual embeddings. BERTScore is a recently proposed metric that makes use of the contextual representations of the predicted and gold sentences. BERTScore focuses on semantic similarity between tokens of reference and hypothesis, trying to understand the meaning of what you have generated and what was supposed to be generated. We report BERTScore because ROUGE can mainly capture n-gram overlap, which is inadequate for the abstractive summarization setting. Some examples of the generated summarizations are available in the Appendix b.

4.4.2.1 Quantitative results

In Table 4.5 we compare the performance of our models fine-tuned on the summarization task. Although GreekBART is a BART-BASE model and it is compared to BART-LARGE models, it is capable of better performance than all other models in the task of GreekSUM abstract. Only mBART₅₀ achieves a slightly higher BERTScore than GreekBART when evaluated using the M-BERT model. On the other hand, both mBART models surpass our model in the GreekSUM title task. However, even in that task the performance of GreekBART

		Repetitions (%)
Abstract	Gold	7.77
	BART-random	28.12
	GreekBART (ours)	12.19
	mBART ₂₅	12.7
	mBART ₅₀	10.03
		Repetitions (%)
Title	Gold	0.91
	BART-random	8.76
	GreekBART (ours)	3.62
	mBART ₂₅	2.52
	mBART ₅₀	2.52

Table 4.7 – The percentage of repeated words on the summaries.

is comparable to one of the two mBART models, both in terms of ROUGE and BERTScore. Our evaluation indicates that mBART₅₀ and GreekBART are the most promising models for the two summarization tasks. Specifically, mBART₅₀ performs better overall in both generative tasks, being the top-performing model in the GreekSUM title task and second-best in the GreekSUM Abstract task, according to its ROUGE and BERTScores. On the other hand, GreekBART excels in the GreekSUM abstract task but ranks third-best in the GreekSUM title task. Generally, it is remarkable that both mBART models, which are not pre-trained on the Greek language, are capable to achieve a good performance due to the size of GreekSUM dataset, which contains more than 100k training samples. It is clear that BART-random has the poorest performance by a significant margin. Finally, it is interesting that mBART₅₀ has a better performance than mBART₂₅ in terms of both ROUGE and BERTScore, while their only difference is the number of languages on which they are pre-trained. This situation warrants further investigation, as it is possible that some of the additional 25 languages supported by mBART₅₀ have roots in the Greek language, potentially contributing to a better understanding of the language model.

4.4.2.2 Qualitative results

As shown in Table 4.6, GreekBART is more abstractive than the two mBART models, as its generated summaries display a higher degree of novel n-grams. In general, none of the models surpasses the LEAD method in terms of ROUGE scores. Furthermore, the ROUGE scores of the models suggest that the machine-generated summaries tend to be extractive, since the gold summaries are also predominantly extractive in nature. This situation is confirmed by the proportion of novel n-grams that are introduced (Table 4.6), where few new words are introduced in the GreekSUM gold summaries, influencing, therefore, the training of the models examined, forcing them to generate more extractive summaries. Moreover, Table 4.6 shows that the length of all generated summaries is

	System	Score
	Gold	45.24
BASE	BART-random	-72.62
	GreekBART (ours)	10.71
LARGE	mBART ₂₅	-03.57
	mBART ₅₀	20.24

Table 4.8 – The results of the human evaluation study.

pretty close to the length of ground truth summaries. According to Table 4.7 the generated summaries of mBART₅₀ contain the smallest percentage of repetitions, with GreekBART following. The rate of repeated words on mBART₅₀ summaries is close to the one of ground-truth summaries. Finally, we notice that BART-random introduces many new words; however, they are irrelevant.

4.4.2.3 Human Evaluation

In order to further understand and validate quantitative results, we conducted a human evaluation study, using the best-worst scaling (Louviere, Flynn, and Marley, 2015). We chose 11 native Greek speakers from various age groups, ranging from 18 to 60 years old, with varying educational backgrounds and levels. Following the Narayan, Cohen, and Lapata 2018 method, we randomly selected 14 documents from the GreekSUM abstract test set and for each document we generated all possible pairs of human-authored summaries (Gold), GreekBART, BART random, mBART₂₅, and mBART₅₀, resulting in a total of 140 pairs for all documents. Thus, each pair of summaries consists of two summaries generated by two different models. Volunteers were presented with a document and a pair of summaries, and they should decide which one is the best summary and which was the worst, based on the accuracy (does the summary contain accurate facts?), the informativeness (is important information captured?) and the fluency (is the summary written in well-formed Greek?). Each pair of summary was randomly assigned to three participants and the score of the system was determined by calculating the percentage of times it was selected as the *best* summary, minus the percentage of times it was selected as the *worst* summary. Thus, the maximum score that a model can achieve is 100, while the minimum score can be -100. The results of the human evaluation study are presented in Table 4.8. Gold reaches first place, followed by mBART₅₀ and GreekBART. According to the evaluators, Gold is by far the most preferred summary, while the mBART₅₀ score is remarkably higher than that of GreekBART, verifying our assumptions based on the quantitative results. Finally, the high negative score of BART-random indicates that its summaries were considered worse in the majority of cases.

4.5 CONCLUSION

We introduced GreekBART, the first pre-trained Seq2Seq model specifically for the Greek language. Also, we created the first summarization dataset for the Greek language. Our model was shown to outperform previous state-of-the-art models in 3 out of 4 discriminative tasks and to be on par with BART-LARGE models on summarization tasks. Moreover, we presented the capabilities of zero-shot learning, training from scratch a multilingual BART model on summarization tasks, even though it was not pre-trained on the Greek language. As a future work, we can consider the creation of a more abstractive summarization dataset, and the investigation of any correlation between the Greek language and one or more of the 25 extra languages of mBART50. Finally, it would be interesting to try to boost the performance of mBART50 on summarization tasks by applying an affordable language-adaptive phase in order to further pre-train it on the Greek language for a logical number of epochs.

ETHICS STATEMENT

The collection of the GreekSUM dataset was performed using a Python crawler that respected the *robots.txt* of <http://www.news247.gr>. As the dataset is used only for evaluation purposes, the content follows the legal instructions listed on the webpage.

For the GreekBART training, we used a cluster of GPUs consisting of 2 NVIDIA V100 GPUs for 20 days. As the majority of language models based on the BART architecture, the energy resources required for pre-training models are currently very high and need to be tackled soon (Strubell, Ganesh, and McCallum, 2019).

LIMITATIONS

The proposed GreekSUM dataset that we used for the evaluation of our model is limited to news articles from one webpage only. Thus, the ability to abstract summarize GreekBART is assessed on one single domain only. This is due to the fact that there is a lack of non-English benchmarks and tasks. This is also applicable in discriminative tasks, where the only available ones for Greek are either sentence classification or natural language inference. Although other evaluation datasets are not existing for the Greek language (i.e. Word Sense Disambiguation) or are not available to the public (i.e., Named Entity Recognition dataset).

On the other hand, GreekBART is only compared with extractive summarization methods or with large multilingual language models for the summarization task. Since it is the first base model for this language and since the base mBART model does not exist publicly, a fair in-depth comparison of GreekBART with other summarization systems could not be conducted.

WORD SENSE INDUCTION

Word sense induction (WSI) is a challenging problem in natural language processing that involves the unsupervised automatic detection of a word's senses (i.e. meanings). Recent work achieved significant results on the WSI task by pre-training a language model that can exclusively disambiguate word senses. In contrast, others employ off-the-shelf pre-trained language models with additional strategies to induce senses. In this chapter, we propose a novel unsupervised method based on hierarchical clustering and invariant information clustering (IIC). The IIC loss is used to train a small model to optimize the mutual information between two vector representations of a target word occurring in a pair of synthetic paraphrases. This model is later used in the inference mode to extract a higher-quality vector representation to be used in hierarchical clustering. We evaluated our method on two WSI tasks and in two distinct clustering configurations (fixed and dynamic number of clusters). We empirically show that our approach is at least on par with the state-of-the-art baselines, outperforming them in several configurations. The code and the data to reproduce this work are available to the public¹.

5.1 INTRODUCTION

The automatic identification of a word's senses is an open problem in natural language processing, known as "word sense induction" (WSI). WSI is closely related to the word sense disambiguation task (WSD). While the latter relies on a predefined sense inventory (i.e. WordNet (Feinerer and Hornik, 2020; Fellbaum, 1998; Wallace, 2007)) and aims to classify the word's sense in context, the former focuses on clustering a collection of sentences according to the target word senses. For example, Figure 5.1 shows the different clusters obtained using our approach² on 3000 sentences that contain the word *bank* collected from Wikipedia. Note that in this case, the senses and their number are not predefined, which highlights the difference between WSI and WSD.

Word senses are more beneficial than simple word forms for various tasks, including Information Retrieval (Pantel and Lin, 2002). Word senses are typically represented as a fixed list of definitions from a manually constructed lexical database. However, lexical databases lack important domain-specific senses. For example, these databases often lack explicit semantic or contextual links between concepts and definitions (Agirre et al., 2009). Hand-crafted lexical databases also frequently fail to convey the precise meaning of a target word in a specific context (Véronis, 2004). To address these issues, the WSI intends to learn in an unsupervised manner the various meanings of a given

1. <https://github.com/hadi-abdine/wsi-mim>

2. with RoBERTa_{LARGE} (Liu et al., 2019) as underlying model

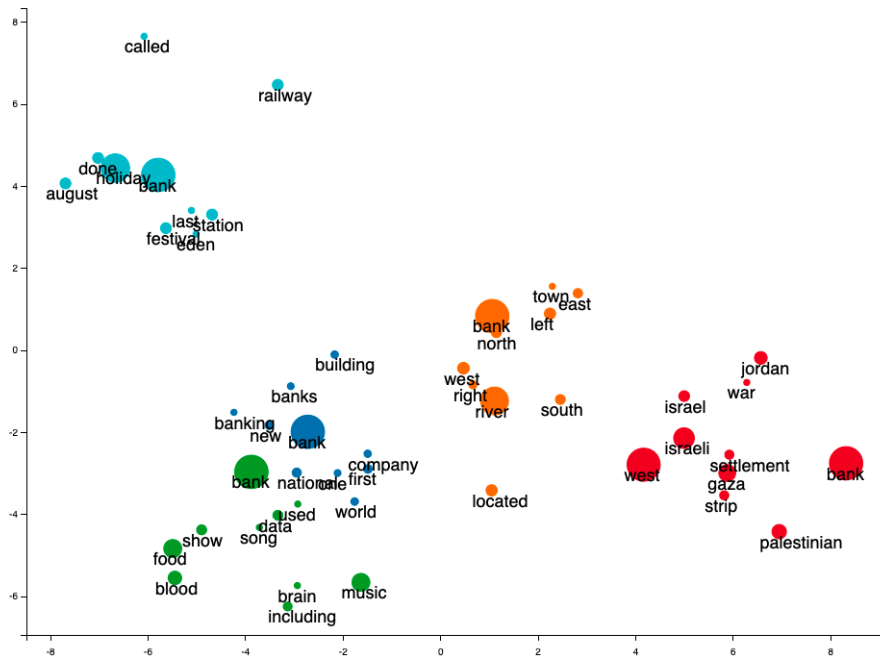


Figure 5.1 – The different sense-based clusters of the word **bank** with the most frequent words used in the corresponding contexts. We used PCA to project the cluster centroids into a 2D space. Each color corresponds to a cluster. The size of the points represents the frequency of the words in their corresponding cluster.

word. Although current state-of-the-art methods reasonably tackle this problem, they have significant limitations that should be addressed. For example, in their approaches, Ansell, Bravo-Marquez, and Pfahringer (2021) and Amrami and Goldberg (2019) choose a fixed number of senses regardless of the target word without an explicit justification for their choices. On the other hand, the approach of Ansell, Bravo-Marquez, and Pfahringer (2021) requires the pre-training of a new language model with a fixed vocabulary specific to the task. Applying their approach to a new vocabulary or a new language will be computationally expensive, which can impede the process of experimentation. This chapter includes the following contributions.

- 1) We propose a new unsupervised method leveraging pre-trained language models, hierarchical clustering, and mutual information maximization. Our approach addresses some limitations of the previous efforts while providing competitive performance.
- 2) We apply a new method to estimate the dynamic number of senses for target words. This method relies on word polysemy quantification (Xypolopoulos, Tixier, and Vazirgiannis, 2021).
- 3) We study the variation of performance w.r.t. the depth of the selected layer. Our findings in Section 5.7, which cover four different models, are valuable for researchers conducting future work on WSI.

5.2 RELATED WORK

Previous work on WSI use generative statistical models to solve this task. In general, they approach this task as a topic modeling problem using Latent Dirichlet Allocation (LDA) (Chang, Pei, and Chen, 2014; Goyal and Hovy, 2014; Komninos and Manandhar, 2016; Lau et al., 2012; Wang et al., 2015). AutoSense Amplayo, Hwang, and Song, 2019, one of the most recent best-performing LDA methods, is based on two principles. First, the senses are represented as a distribution over topics. Second, the model generates a pair composed of the target word and its neighboring word, thus separate the topic distributions into fine-grained senses based on lexical semantics. AutoSense throws away the garbage senses by removing topic distributions that do not belong to any instance. Furthermore, it adds new ones according to the generated (target, neighbor) pairs, which means that fixing the number of senses by the model is not required. While most of the WSI methods fix the number of clusters for all the words, in our work we explore two setups for the number of clusters, fixed and dynamic. Other work (Corrêa and Amancio, 2018; Song et al., 2016) use the Word2Vec static word embedding (Mikolov et al., 2013) to get the representations of polysemous words before applying the clustering method. After the emergence of contextual word embeddings, pre-trained language models such as ELMo (Peters et al., 2018) (based on BiLSTM) and BERT (Devlin et al., 2019) (based on the transformers) (Vaswani et al., 2017) are used with additional techniques to induce senses of a target word. (Amrami and Goldberg, 2018) and (Amrami and Goldberg, 2019) use consecutively ELMo and BERT_{LARGE} to predict probable substitutes for target words. Next, it gives k representatives for each instance where each contains multiple possible substitutes drawn randomly from the word distribution predicted by the language model. Each representative is a vector obtained from TF-IDF. The representatives are then clustered using agglomerative clustering, where the number of clusters is fixed to 7. Finally, each instance will be assigned to one or more clusters according to the corresponding cluster of each of its representatives. Instead of using the word substitutes approach, our work uses the contextual word embedding extracted from pre-trained language models.

PolyLM (Ansell, Bravo-Marquez, and Pfahringer, 2021) is one of the most recent techniques for induction of words sense that uses an MLM (Masked Language Model) to induce senses. PolyLM took a novel approach to the problem of learning word senses. It uses the transformer architecture to predict eight probabilities for each word, where each probability represents the probability that a word will be assigned to one of eight different senses. It is built on two assumptions: the chance of a word being predicted in a masked place is proportional to the total of its distinct senses, and for a particular context, one of the word's senses is more likely to be used. The model has the drawback of assuming the same fixed number of senses for all words.

5.3 METHOD

Our method (Hadi Abdine et al., 2023) consists of four main steps: First, we construct a synthetic dataset of pairs, each consisting of a sentence paired with a randomly perturbed version as explained in Section 5.3.1. Second, we extract the pair of hidden state representations of the target word using a pre-trained language model (e.g., RoBERTa). In our experiments, we use RoBERTa_{LARGE} primarily. Furthermore, we consider two widely adopted language models, BERT_{LARGE} and DeBERTa_{XLARGE}^{mli} in the ablation study. Third, we train an MIM (Mutual Information Maximization) model where: (1) Considering an instance of the hidden state representations pairs, we trained the network using two objectives: maximizing the mutual information and minimizing the match loss between the output of the two vectors. (2) The best instance of the model is chosen according to the smaller loss on the pre-defined test set. (3) We consider the output of the first layer as the new vector representation for the target word. Fourth, for each target word in the evaluation datasets, we apply the agglomerative clustering method on the new vector representations to obtain our clustering solution. To choose the predefined number of clusters, we follow two approaches: (i) Fix the number of senses (clusters) to 7 as in Amrami and Goldberg (2018, 2019) and (ii) Use a dynamic number of clusters based on the polysemy score (Xypolopoulos, Tixier, and Vazirgiannis, 2021) of each target word.

The main steps are detailed in the following subsections.

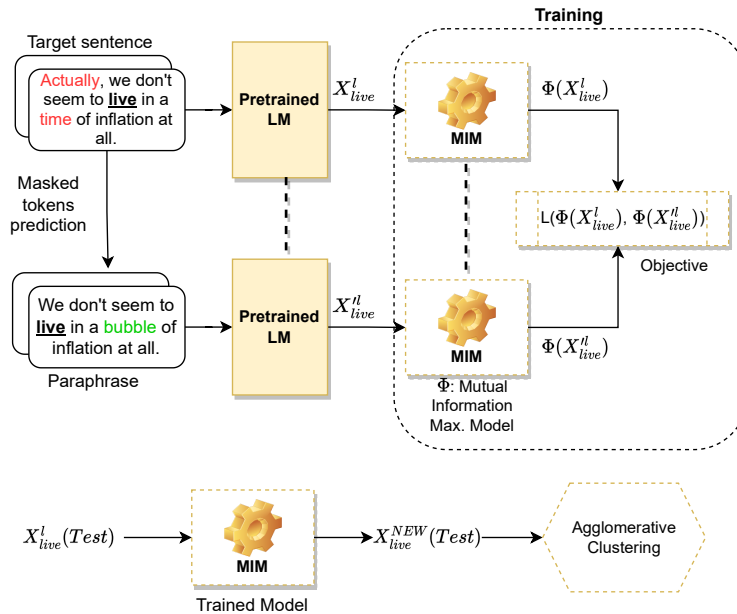


Figure 5.2 – The pipeline of our method: For the word "live" chosen as the target, a list of sentences is provided. BART is used to generate the corresponding paraphrases. The hidden representation X_{live}^l of the target word is extracted from the layer 1 of a pre-trained language model. The dashed line denotes shared parameters.

5.3.1 Dataset Setup

BART (Lewis et al., 2020) is a denoising autoencoder for pre-training sequence-to-sequence models. The training process involves using a model to reconstruct a modified version of the original sentences using a random noising function.. It is based on a standard Transformer-based neural machine translation architecture which can be seen as a generalization of BERT (due to the bidirectional encoder), GPT (Radford and Narasimhan, 2018) (with the left-to-right decoder), and other recent pre-training schemes. BART can be used also as a generative model given an input, i.e. sentence completion, translation, summarization, etc.

GENERATING RANDOMLY PERTURBED REPLICATES In order to apply our method to text input, we need to create a pair of sentences where the target word has the same meaning. To fulfill this, a function is needed to introduce random perturbations to the input sentence while preserving the meaning. The sentence and its perturbed version are keeping the same sense of the target lemma. Thus, we can generate a pair of sentences that belong to the same cluster. First, we masked 40% of the original sentence while preventing -in most cases- masking the target word. Second, we predicted the masked tokens using $BART_{BASE}$ with a beam size of one.

5.3.2 Vectors Extraction

The training set is used to train the parameters of a small network while the test set is used to perform the induction of the senses. Using the best layer of each pre-trained language model, we extracted representations of the target word from the different training and test instances. The best layer for each pre-trained language model is chosen based on the best performance on BERTScore (Zhang* et al., 2020) with WMT16 To-English Pearson³. At this stage, if the target word is broken down into multiple tokens, we computed the average vector of the corresponding word pieces. Note that while generating the perturbation on the input text using $BART_{BASE}$, there is a small probability that the paraphrase might not contain the target word. Thus, all the sentences in the training set with their corresponding paraphrases deprived of the target word are removed.

5.3.3 Loss Function

We seek to minimize a loss function L with two components, each of which is explained in the following.

$$L = L_{IIC} + L_M \quad (5.1)$$

3. https://docs.google.com/spreadsheets/d/1RK0Vpse1B98Nnh_E0C4A2BYn8_201tmPODpNWu4w7xI/edit?usp=sharing

5.3.3.1 Invariant Information Clustering Loss

Invariant information clustering IIC (Ji, Henriques, and Vedaldi, 2019) is a clustering objective that learns a neural network from scratch to perform unsupervised image classification and segmentation. The model learns to cluster unlabeled data based on maximizing the mutual information score between the unlabeled sample and a transformation of the input. Therefore, both the input and its corresponding transformation surely contain the same information and belong to the same class/cluster. Maximizing the mutual information is robust to clustering degeneracy, where a single cluster tends to dominate the predictions or some clusters tend to disappear as in k-means. Also, it helps to avoid noisy data from affecting the predictions by overclustering. In addition, maximizing the mutual information between two samples of the same cluster increases the model ability to capture dependencies between variables. Thus, leading to better latent space word representation that separates better word senses. The objective function is as follows:

$$\max_{\Phi} I(\Phi(x), \Phi(x')) \quad (5.2)$$

Where Φ is the classification neural network, x is the input, and $x' = g(x)$ is the transformation (random perturbation of the input) of x (i.e. rotation, maximizing, minimizing, etc.). This is equivalent to maximizing the predictability of $\Phi(x)$ from $\Phi(x')$ and vice versa. The mutual information function is defined by:

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} P_{X,Y}(x, y) \log \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)} \quad (5.3)$$

The loss of invariant information clustering is therefore defined by:

$$L_{IIC} = -I(\Phi(x), \Phi(x')) \quad (5.4)$$

We adopt the IIC loss to the NLP domain by changing the nature of the random perturbation introduced to the input.

5.3.3.2 Match Loss

In our approach, we use a synthetic dataset generated by a language model (BART) as explained in Section 5.3.1. The idea of using AI-generated data to train other deep learning may lead to model collapse (Kothapalli, Tirer, and Bruna, 2023). Where the output of the model's last layer might represent the same embeddings for all the different train sentences in some cases. To address this issue, we encourage similarity between the output of the last layer $\Phi(x)$ and $\Phi(x')$ by adding a match loss. This loss is proportional to the cosine similarity between the two outputs, and is inspired by (Ansell, Bravo-Marquez, and Pfahringer, 2021), which the authors have proven to be effective with the following:

$$L_M = -0.1 \sum \frac{\Phi(x) \cdot \Phi(x')}{\|\Phi(x)\| \|\Phi(x')\|} \quad (5.5)$$

The match loss function guides the neural network to learn embeddings such that similar inputs (x and x') are mapped closer together in the embedding space. By minimizing the negative cosine similarity between these embeddings while considering their magnitudes, the network learns to generate embeddings that are more similar for pairs of inputs that are deemed similar based on the task objective.

This type of loss function is often used in metric learning scenarios, where learning similarity or distance between data points is essential, to create embeddings that capture similarity relationships effectively, which is crucial in our case.

5.3.4 *Sense Embedding: Obtaining new word vectors*

The architecture of our MIM model is very simple. It is made up of three projection layers with the ReLU activation function. The final layer is equipped with the softmax function to obtain a vector of probability distribution that can be used as input to our loss. The hidden size of one linear layer is set to the double of RoBERTa_{LARGE} hidden state dimension which is 1024.

For each target word, we train a model while providing the pairs of extracted representations belonging to the same cluster. In other terms, the target word's representations from the original sentence and the sentence with lexical perturbation, respectively.

The training consists of 8 runs over 5 epochs with a batch size of 32 using Adam optimizer (Kingma and Ba, 2015). The learning rate starts with $2e-5$ and then is linearly reduced to zero during the remaining training time. The best model results from the epoch, minimizing the validation loss. The validation set represents 10% pairs of sentences drawn randomly from the train dataset. Once training is complete, the hidden state representation of the first layer is extracted for each test word vector of the original sentence. Thus, the target word has a new projected representation.

5.3.5 *Clustering*

To cluster the instances into senses, we used the agglomerative clustering method. Agglomerative clustering is a hierarchical clustering technique used in unsupervised machine learning to group similar data points into clusters. It is a bottom-up approach where each data point starts as its own cluster and then iteratively merges with the closest clusters based on a defined distance metric and a linkage function. The linkage function takes the distance information and groups pairs of objects into clusters based on their similarity. Next, these newly formed clusters are linked to each other to create larger clusters. This process is iterated until all objects in the original data set are linked in a hierarchical tree. The advantage of agglomerative clustering lies in its ability to produce a hierarchical structure of clusters, allowing flexibility in selecting the number of clusters based on the problem at hand. Agglomerative clustering is widely used in various fields, including biology, image segmentation, social network analysis,

and market segmentation, among others, where understanding hierarchical relationships among data points is beneficial.

For a fair comparison, the same setup as in (Amrami and Goldberg, 2018, 2019) is used along with the cosine distance as the distance metric and the average linkage. To choose the number of clusters (senses) of each target word, we follow two approaches: (i) Fix the number of senses as in (Amrami and Goldberg, 2018, 2019; Ansell, Bravo-Marquez, and Pfahringer, 2021). (ii) Use a dynamic number of clusters between four and eight based on its polysemy score obtained using the unsupervised word polysemy quantification function (Xypolopoulos, Tixier, and Vazirgiannis, 2021). For dynamic clustering, we use the best configuration in the paper with dimensionality D equal to 3 and level L equal to 8.

5.4 EVALUATION

Several competitions were organized to systematically evaluate various methods applied for WSI, including *SemEval-2007 task 02* (Agirre and Soroa, 2007), *SemEval-2010 task 14* (Manandhar and Klapaftis, 2009) and *SemEval-2013 task 13* (Jurgens and Klapaftis, 2013). The two tasks of *SemEval-2010* and *SemEval-2013* are considered the benchmark for WSI. In this section, we publish and analyze the mean and the standard deviation over eight runs of the previously described model on the two mentioned tasks: *SemEval-2010 task 14* and *SemEval-2013 task 13*.

5.4.1 *SemEval-2010 task 14*:

On one hand, the main objective of the *SemEval-2010* WSI challenge is to compare unsupervised word-sense induction systems. It provides a mapping mechanism for evaluating WSI systems using the WSD dataset. The target word dataset consists of 100 tagged words, 50 nouns, and 50 verbs extracted from OntoNet (Hovy et al., 2006). In the test set, each target word has around one hundred instances to be clustered. To learn its senses, a training set containing approximately 10,000 instances is provided for each target word. The training set is created using a semi-automatic web-based method. For each sense of the target word in WordNet (Fellbaum, 1998), the query grabs all the sentences containing its corresponding stems and lemmas using Yahoo! search API. Each instance in the test dataset in this task is labeled with one sense only.

The performance in this task is measured with V-Measure (Rosenberg and Hirschberg, 2007) (biased toward a high number of clusters) and F-Score (biased toward low number of clusters). We report the overall performance (AVG) defined as the geometric mean of these two metrics. Where:

- the V-measure assesses the quality of a clustering solution by explicitly measuring its homogeneity and completeness. Homogeneity refers to the degree to which each cluster consists of data points primarily belonging to a single golden standard class, while completeness refers to the degree that each golden standard class consists of data points primarily assigned to a single cluster.

In F-Score, precision can be defined as the number of common instance pairs

(pairs are formed between instances from the same cluster) between the two sets (clustering solution and gold standard) to the total number of pairs in the clustering solution, while recall can be defined as the number of common instance pairs between the two sets to the total number of pairs in the gold standard. Finally, precision and recall are combined to produce the harmonic mean called the F-score.

5.4.2 *SemEval-2013 task 13*:

On the other hand, *SemEval-2013 task 13* is a task to evaluate Word Sense Induction and Disambiguation systems in a context where instances are tagged with many senses whose applicability is weighted accordingly (Fuzzy Setting). The task focuses on disambiguating the senses for 50 target lemmas: 20 nouns, 20 verbs, and 10 adjectives. The ukWac corpus (Baroni et al., 2009) is provided as a training corpus. It contains a large number of instances crawled from the Web and can be filtered by the lemma, the POS tag, and many more filters⁴. Test data are drawn from the Open American National Corpus (Ide and Suderman, 2004) in a variety of genres and from the spoken and written portions of the corpus.

The performance of this task is measured with Fuzzy B-Cubed (F-BC)(Bagga and Baldwin, 1998). It is a generalized version of B-Cubed that deals with the fuzzy setting and Fuzzy Normalized Mutual Information (F-NMI). The former estimates the fit between two clustering solutions based on how many items share a cluster in both solutions. The latter is a generalized version of mutual information that deals with multi-sense annotation. We also report on the overall performance (AVG).

5.4.3 *Experiments*

To prepare the training set for *SemEval 2010 task 14*, we randomly chose 3,500 sentences from the training dataset provided for this task for each target word. For *SemEval-2013 task 13*, we extracted for each target word tagged up to 3500 random sentences from ukWac. Note that if some of the target words in *SemEval-2013 task 13* do not have 3,500 sentences in ukWac, we extracted all possible sentences.

Dataset	Train	Test
SemEval-2010 Task 14	3.02%	13.5%
SemEval-2013 Task 13	16.05%	9.95%

Table 5.1 – The average perturbation percentage between the input text and the paraphrase. This percentage represents the proportion of changed unigrams.

Following, we generate the paraphrases for both datasets by integrating the random perturbation described in section 5.3.1. The average percentage of

4. https://corpora.dipintra.it/public/run.cgi/first_form

perturbation for each dataset is presented in Table 5.1.

The instances in the datasets of *SemEval-2010 task 14* and *SemEval-2013 task 13* contain some of the target words with morphological variability. Therefore, lemmatization is required to identify the target lemma during the vector extraction phase. Given this word and its POS tag, we use the WordNetLemmatizer from the *NLTK* library to find its position inside both the sentence and its paraphrase followed by extracting the corresponding RoBERTa, BERT, and DeBERTa vectors. These vectors are used to train the model as described earlier.

To infer the meaning of an instance in *SemEval-2010*, we first apply the agglomerative clustering method on the extracted RoBERTa_{LARGE} vectors of the target word in the SemEval instances (Section 5.3.2). The aforementioned step studies the effect of our word vectors enriching method. Second, for the model to be tested, we forward the test word vectors to the trained model and extract the corresponding hidden state of the first layer. This state is considered as the new word representation (sense embedding) of dimensionality 2,048.

Finally, we applied agglomerative clustering on the new word representations implementing our clustering solution. We assign each instance to a single cluster.

The results of the evaluation on both *SemEval-2010* and *SemEval-2013* tasks are presented in tables 5.3 and 5.4, respectively, providing the comparison to other WSI systems.

In the *SemEval 2013* task, there is the possibility that a word has multiple senses with a corresponding degree of applicability. Thus, once agglomerative clustering was applied, we convert the cosine similarity distances between each target word’s representation and the centroids of the different clusters to a vector of probabilities using the Softmax function. These probabilities are considered to be degrees of applicability of the senses. The average number of clusters for each dataset in the dynamic setting is presented in table 5.2.

Dataset	Average # of clusters
SemEval-2010 Task 14	6.73
SemEval-2013 Task 13	5.36

Table 5.2 – The average number of clusters obtained by using the polysemy scores on SemEval 2010 and SemEval 2013 test datasets.

5.4.4 Results

Table 5.3 shows the performance of our approach compared to other baselines on *SemEval 2010 task 14*. The best performing system, among the baselines, is Amrami and Goldberg (2019) providing the highest F score of 71.3%. With our method, RoBERTa_{LARGE} outperforms all baselines in both settings: Fixed and dynamic number of clusters. This finding highlights the importance of our MIM approach, which allows for an improvement of 2.5 absolute points

over the previous state-of-the-art in terms of average score. In addition, we observe that the only model that uses dynamic clustering among the baselines (AutoSense) is largely outperformed by the other methods using a fixed number of clusters. However, given that WSI is an unsupervised task, the fixed number of clusters is supposed to be arbitrary, and there is no guarantee that using the same number of clusters on other datasets would be optimal. Our proposed dynamic approach to choose the number of clusters leads to better performance alongside RoBERTa_{LARGE}.

Model	# Clusters	V-Measure	F-score	AVG
RoBERTa _{LARGE} ¹⁷	7	39.8	67.18	51.71
RoBERTa _{LARGE} ¹⁷ (+MIM)	7	46.26±0.51	68.18±0.4	56.16±0.42
RoBERTa _{LARGE} ¹⁷	Dynamic	37	67.42	49.94
RoBERTa _{LARGE} ¹⁷ (+MIM)	Dynamic	45.06±0.92	68.79±0.33	55.67±0.54
PolyLM _{BASE} (Ansell, Bravo-Marquez, and Pfahringer, 2021)	8	40.5	65.8	51.6
PolyLM _{SMALL} (Ansell, Bravo-Marquez, and Pfahringer, 2021)	8	35.7	65.6	48.4
Amrami and Goldberg (2019)	7	40.4	71.3	53.6
AutoSense (Amplayo, Hwang, and Song, 2019)	Dynamic	9.8	61.7	24.59

Table 5.3 – Evaluation of WSI models on SemEval 2010 task 14. The (+MIM) label indicates that mutual information maximization is applied to obtain the clustered vectors. Otherwise, the vectors from the pre-trained language models are directly used.

Model	# Clusters	F-BC	F-NMI	AVG
RoBERTa _{LARGE} ¹⁷	7	64.1	19.28	35.16
RoBERTa _{LARGE} ¹⁷ (+MIM)	7	62.49±0.48	21.5±0.62	36.67±0.64
RoBERTa _{LARGE} ¹⁷	Dynamic	64.2	16.11	32.16
RoBERTa _{LARGE} ¹⁷ (+MIM)	Dynamic	64.8±0.29	19.95±0.63	35.95±0.56
PolyLM _{BASE} (Ansell, Bravo-Marquez, and Pfahringer, 2021)	8	64.8	23	38.3
PolyLM _{SMALL} (Ansell, Bravo-Marquez, and Pfahringer, 2021)	8	64.5	18.5	34.5
Amrami and Goldberg (2019)	7	64	21.4	37
Amrami and Goldberg (2018)	7	57.5	11.3	25.4
AutoSense (Amplayo, Hwang, and Song, 2019)	Dynamic	61.7	7.96	22.16

Table 5.4 – Comparison of WSI-specific techniques on SemEval 2013 task 13. The (+MIM) label indicates that mutual information maximization is applied to obtain the clustered vectors. Otherwise, the vectors from the pre-trained language models are directly used.

SemEval 2013 task 13 performances are shown in Table 5.4. The best-performing baseline is PolyLM_{BASE} providing the highest F-BC and F-NMI scores. Although our approach did not outperform this baseline, it shows to be very competitive. In fact, the results on *SemEval 2013 task 13*, again show the positive contribution of our MIM approach, as we can observe a significant improvement whenever it is applied. For example, applying MIM to RoBERTa_{LARGE} with dynamic clustering led to an increase of 3.8 absolute points in terms of average score. On the other hand, our method has two main advantages over PolyLM_{BASE}: (1) It can use the dynamic number of clusters compared to eight fixed senses for all words in PolyLM. (2) It does not require a computational-heavy pre-training to apply WSI on other languages. Indeed our method can be applied on other languages using already pre-trained language models such as CamemBERT (Martin et al., 2020) or BART_{hez} (Kamal Eddine, Tixier, and Vazirgiannis, 2021)

for the French language, AraBERT (Antoun, Baly, and Hajj, 2020) for the Arabic language, etc..

In summary, (1) our proposed intermediate MIM phase led on average to an improved hierarchical clustering, and (2) the dynamic approach to choose the number of clusters maintained the stable and competitive performance of our different evaluated models.

5.5 TESTING VARIOUS PRE-TRAINED LANGUAGE MODELS

We conducted a study to assess the performance of our approach using different language models. We mainly tested RoBERTa_{LARGE}¹⁷, BERT_{LARGE}¹⁸, and DeBERTa_{XLARGE}⁴⁰^{mli}. We report the results of all models in table 5.5. This table further shows that our proposed dynamic approach to choose the number of clusters is effective and did not deteriorate the performance of our method and in some cases led to a better performance, mainly in the SemEval-2010 WSI task. Where, using the MIM approach on BERT_{LARGE}¹⁸ with the dynamic number of clusters leads to 53.1 average score compared to 51.26 average score using a fixed number of clusters. We argue that the good quality of the context in the SemEval-2010 training dataset leads to this improvement compared to short and incomplete context in the SemEval-2013 training dataset, as we can see in Appendix c. Since a short incomplete context will result in a (1) bad quality generated synthetic dataset using BART model and (2) suboptimal contextual word embedding for both original and synthetic contexts. Finally, we further validate the positive impact of our MIM approach that shows improvement in the average score on almost all language models and for both tasks SemEval-2010 task 14 and SemEval-2013 task 13. For example, using BERT_{LARGE}¹⁸ with MIM leads to almost one point of improvement in terms of F-BC, F-NMI and AVG scores of SemEval-2013 task 13. Here the average score for this task is 37.56, which outperforms the model of Amrami and Goldberg (2019), which uses the same language model as the backbone.

Model	#Clusters	SemEval-2010			SemEval-2013		
		V-measure	F-score	AVG	F-BC	F-NMI	AVG
RoBERTa _{LARGE} ¹⁷	7	39.8	67.18	51.71	64.1	19.28	35.16
RoBERTa _{LARGE} ¹⁷ (+MIM)	7	46.26	68.18	56.16	62.49	21.5	36.67
RoBERTa _{LARGE} ¹⁷	Dynamic	37	67.42	49.94	64.2	16.11	32.16
RoBERTa _{LARGE} ¹⁷ (+MIM)	Dynamic	45.06	68.79	55.67	64.8	19.95	35.95
BERT _{LARGE} ¹⁸	7	40.1	65.23	51.14	62.4	21.58	36.7
BERT _{LARGE} ¹⁸ (+MIM)	7	40.51	64.89	51.26	62.63	22.54	37.56
BERT _{LARGE} ¹⁸	Dynamic	41.2	67.17	52.6	64.81	20.86	36.77
BERT _{LARGE} ¹⁸ (+MIM)	Dynamic	41.8	67.43	53.1	64.42	21.22	36.97
DeBERTa _{XLARGE} ⁴⁰	7	40.5	66.64	51.95	63.16	18.57	34.25
DeBERTa _{XLARGE} ⁴⁰ (+MIM)	7	40.05	66.93	51.77	62.52	20.18	35.52
DeBERTa _{XLARGE} ⁴⁰	Dynamic	40.6	67.52	52.36	64.24	17.79	33.8
DeBERTa _{XLARGE} ⁴⁰ (+MIM)	Dynamic	40.58	67.89	52.48	64.44	19.27	35.26

Table 5.5 – Results of different pre-trained language models on *SemEval-2010 Task 14* and *SemEval-2013 Task 13*.

5.6 ABLATION STUDY

In the conducted ablation study in table 5.6, using the RoBERTa_{LARGE}¹⁷ model for SemEval 2010 Task 14, we explored the impact of individual and combined loss functions on model performance. Each reported result represents the mean average derived from eight independent runs. In particular, the AVG metric used corresponds to the geometric mean between the F-score and the V-measure, providing a balanced assessment of overall model performance. Our intuition is that the match loss encourages the network to learn embeddings where similar items are close together and dissimilar embeddings are far apart. In addition, the IIC loss encourages the model to capture the dependencies between variables in the latent space. Here, both losses are important for our architecture where we intend to maximize both abilities in our method. The employed loss functions: Match Loss and IIC Loss were systematically evaluated in different setups. The setup solely employing Match Loss demonstrated a V-Measure of 45.26, an F-score of 67.81, and an average score of 55.4. Conversely, using only IIC loss yielded slightly improved metrics with a V-Measure of 45.53, an F-score of 68.03, and an average score of 55.6. However, the most significant performance enhancement was observed when both match loss and IIC loss were used concurrently. This combined approach resulted in notable improvements with a V-Measure of 46.26, an F-score of 68.18, and the highest average score of 56.16. The effect of integrating both losses emphasizes the efficacy of capturing both similarity preservation and information dependency within the model learning process. These results underscore the effectiveness of the main approach, which leverages the combined utilization of IIC and Match Losses, showcasing improved performance compared to individual loss functions. The consistent performance across multiple runs supports the robustness of the combined loss strategy, suggesting its viability for tasks where both similarity and information dependency are pivotal aspects.

Model	V-Measure	F-score	AVG
RoBERTa _{LARGE} ¹⁷ (+MIM: Only Match Loss)	45.26	67.81	55.4
RoBERTa _{LARGE} ¹⁷ (+MIM: Only IIC Loss)	45.53	68.03	55.6
RoBERTa _{LARGE} ¹⁷ (+MIM)	46.26	68.18	56.16

Table 5.6 – Ablation study: Performance Comparison of Individual and Combined Loss Functions on SemEval-2010 task 14.

5.7 BEST LM LAYER

During the evaluation in section 5.4, we used the list provided by BERTScore (Zhang* et al., 2020) authors regarding the best performing layer. This choice is motivated by the fact that we are dealing with an unsupervised task, thus it is not possible to tune such a hyperparameter without access to gold annotations. However, Zhang* et al. (2020) chose the best layer based on how well it performs

in the machine translation evaluation task. When dealing with a WSI task, there is no guarantee that the best layer is the same. Thus, we carried out a study of the variation of the agglomerative clustering final score with respect to the layer used for the extraction of the vector representations. This study can help researchers in future work to choose the appropriate layer when dealing with a similar unsupervised task.

Figure 5.3 shows the variation of the agglomerative clustering performance as a function of the depth of the chosen layer. Interestingly, we see that the variation of performance follows a similar pattern in SemEval-2010 and SemEval-2013 which can suggest a generalizable pattern over word sense induction datasets. Also, we can see that the pattern changes across different models. Despite having a similar architecture, the best layer depth in RoBERTa_{LARGE} (layer 10) differs significantly from that of BERT_{LARGE} (layer 21). Future work should focus on this discrepancy and study the semantic information captured by each model's layers. Table 5.7 presents the results regarding the best layer

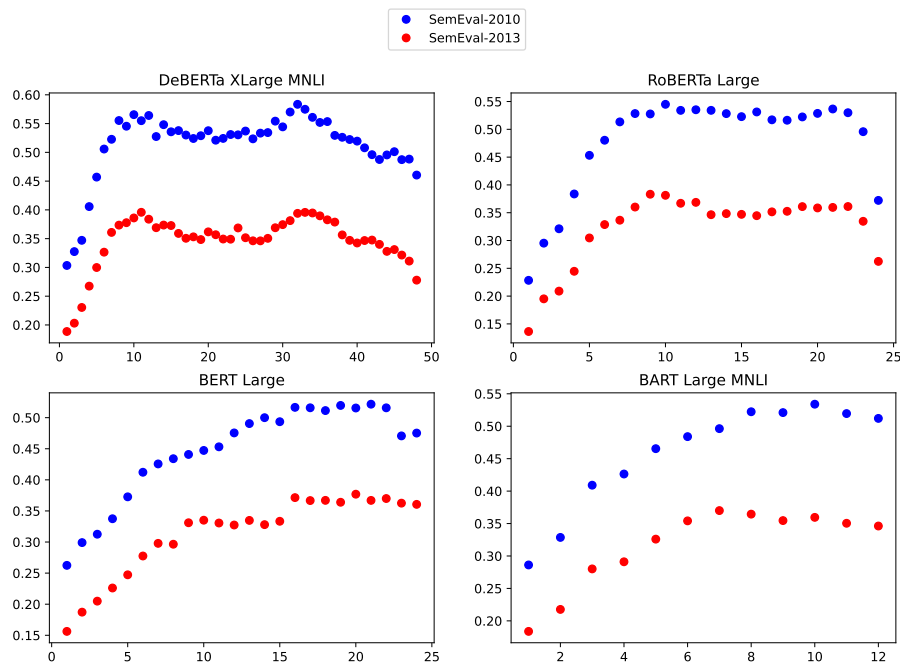


Figure 5.3 – The AVG scores of SemEval-2010 and SemEval-2013 WSI tasks using agglomerative clustering on all the layers of different pre-trained models.

of each pre-trained model on *SemEval 2010 task 14* and *SemEval 2013 task 13*. The best performing pre-trained contextual embeddings for both tasks is DeBERTa_{XLARGE} with a score that outperforms the state-of-the-art methods.

Due to the correlation between the two WSI tasks in terms of pre-trained models layers score, we emphasize on the importance in the future to have an annotated validation dataset to choose the best layer accordingly.

Model	#Clusters	SemEval-2010				SemEval-2013			
		Layer	V-measure	F-score	AVG	Layer	F-BC	F-NMI	AVG
RoBERTa _{LARGE}	7	10	43.6	68.12	54.5	9	63.87	23	38.32
RoBERTa _{LARGE}	dynamic	10	41.9	68.52	53.58	9	65.08	18.84	35.02
BERT _{LARGE}	7	21	40.8	66.7	52.17	20	63.16	22.07	37.34
BERT _{LARGE}	dynamic	21	41.3	67.65	52.85	20	65.54	21.26	37.32
DeBERTa _{XLARGE}	7	32	49	69.48	58.35	33	64.86	24.14	39.57
DeBERTa _{XLARGE}	dynamic	32	46.4	69.49	56.78	33	66.62	21.71	38.03

Table 5.7 – The best layers of different pre-trained language models on the word sense induction tasks: *SemEval-2010 Task 14* and *SemEval-2013 Task 13*.

5.8 CONCLUSION

In this work, we introduced an unsupervised method for the WSI task based on the tuning of contextual word embeddings extracted from a pre-trained language model. The method generates paraphrases of the input sentences. Therefore, both sentences belong to the same sense cluster. Next, it uses both sentences to train a MIM neural network that maximizes the mutual information between the two sentences’ outputs and minimizes the integrated match loss. The method improves the state-of-the-art in one of the two WSI tasks.

We also use the polysemy score to test the dynamic number of senses setup as it claims superiority over the fixed setting in two out of six experiments. The MIM method proves, in most cases, an improvement in score while it does not deteriorate the performance in the others.

The extraction of representations for the target word depends on the chosen layer from the used pre-trained language model. Thus, inspired by previous work, we conduct a comparison that helps the future studies in this choice.

5.9 LIMITATIONS

The aforementioned method presents an important improvement over some of the state-of-the-art solutions for WSI tasks. However, it suffers from some limitations that are worth highlighting:

(1) This method involves training an MIM model from scratch for each target word, proving a lack of generalizability. Thus, a further study can fulfill this task by training the MIM model starting from a pre-trained language model for all target words. Applying this might yield a general model that can give the sense embedding for all possible target words before applying agglomerative clustering.

(2) Using partially pre-trained language models in our pipeline makes our method costly in terms of computation time compared to *PolyLM*. As a consequence, our method suffers from a higher number of parameters, especially with models of larger size such as DeBERTa. Thus, a further approach is to test with smaller models (i.e., DitiBERT) that could maintain the same good performance with faster training and inference time. Finally, we must highlight the crucial role of the quality of the training data in determining the performance of our model on SemEval-2013 task 13. Unlike the comprehensive

and meticulously constructed training sentences utilized in SemEval-2010 task 14, the training sentences sourced from ukWac for SemEval-2013 task 13 are characterized by their brevity, incompleteness, and nonuniform extraction from the Web. To illustrate the disparities between the training sets for both tasks, we provide examples in Appendix c.

IN recent years, significant progress has been made in the field of protein function prediction with the development of various machine learning approaches. However, most existing methods formulate the task as a multiclassification problem, i.e. assigning predefined labels to proteins. In this chapter, we propose a novel approach, **Prot₂Text**, which predicts a protein's function in a free text style, moving beyond the conventional binary or categorical classifications. By combining graph neural networks (GNNs) and large language models (LLMs), in an encoder-decoder framework, our model effectively integrates diverse data types including protein sequence, structure, and textual annotation and description. This multimodal approach allows for a holistic representation of proteins' functions, enabling the generation of detailed and accurate functional descriptions. To evaluate our model, we extracted a multimodal protein dataset from SwissProt and empirically demonstrated the effectiveness of Prot₂Text. These results highlight the transformative impact of multimodal models, specifically the fusion of GNNs and LLMs, empowering researchers with powerful tools for more accurate function prediction of existing and first-to-see proteins.

6.1 INTRODUCTION

Understanding protein function is a central problem in biological sciences, as proteins are the fundamental elements of almost all biological functions. Accurate prediction of proteins' function is essential for understanding biological systems as well as for various applications, such as drug discovery, allowing researchers to identify and target specific proteins that play critical roles in disease pathways (Ha et al., 2021). Traditionally, protein function prediction has been approached through classification methods, assigning predefined labels to proteins based on their characteristics (Kulmanov and Hoehndorf, 2019). However, this approach often oversimplifies the complexity of the protein's functionality, limiting the depth of our understanding. To overcome this limitation, we propose a novel view on proteins' functions prediction based on reformulating the task using free-text proteins' descriptions instead of relying on predefined labels. Rapid progress in transformer-based models has brought about a massive revolution in the field of Natural Language Processing (NLP). These models have demonstrated impressive language generation capabilities, which allow them to perform a wide range of NLP tasks with remarkable performance, including text completion, translation, sentiment analysis, and question answering (Brown et al., 2020; Radford et al., 2019; Vaswani et al., 2017). On the other hand, graph neural networks (GNNs) have emerged as a powerful tool for modeling graph-structured data, capturing intricate relationships be-

tween different elements in a graph (Kipf and Welling, 2017; Reiser et al., 2022). However, the integration of GNNs and transformers faces various challenges, such as effectively handling the heterogeneity of data representations; therefore, the field is still in its early stages. Despite this, the potential benefits of leveraging both GNNs and transformers for graph-to-text applications, such as predicting the functional properties of proteins, are substantial. To this end, we develop a novel multimodal framework, **Prot2Text**, that can generate detailed and accurate descriptions of the functions of proteins in free text. We effectively integrate GNNs and Large Language Models (LLMs) to encompass both structural and sequential information of the protein’s 3D structure and amino acid sequence, respectively. The encoder-decoder architecture forms the backbone of our model, with the encoder component employing a Relational Graph Convolution Network (RGCN) (Schlichtkrull et al., 2018) to process the protein graphs and the ESM protein language model (Lin et al., 2023) to encode the protein sequence. The decoder component utilizes a pre-trained GPT2 model to generate detailed proteins’ descriptions. To train our multimodal model, we compile a dataset of proteins extracted from SwissProt, a comprehensive collection of protein annotations obtained from the UniProt database (Consortium, 2015). This dataset encompasses a vast number of proteins, each annotated with its corresponding function or description. In addition to the textual information, we obtain the 3D structure representation of the proteins from AlphaFold (Varadi et al., 2022). We also released this curated dataset to the public, allowing other researchers to use it for benchmarking and further advances in the field. Our main contributions can be summarized as follows:

- We introduce the **Prot2Text** framework, a novel multimodal approach for generating proteins’ functions in free text. Our model combines both GNNs and ESM to encode the protein in a fused representation, while a pretrained GPT2 decodes the protein’s text description.
- We propose various baselines for protein text generation and demonstrate that the integration of graph- and sequence-protein information leads to better generation capabilities.
- We further release a comprehensive multimodal protein dataset, which includes 256,690 protein structures, sequences, and textual function descriptions. Researchers can leverage this dataset to benchmark and compare their models, thereby driving advancements in the field and enabling for a more robust and standardized evaluation of protein functions prediction methods in free text format.

6.2 RELATED WORK

TRANSFORMERS. The transformer-based encoder-decoder model was first introduced by Vaswani et al. in their article ‘Attention is all you need’. Since then, this model architecture has become the de-facto standard encoder-decoder architecture in Natural Language Processing (NLP). Despite significant research on different pre-training objectives for transformer-based encoder-decoder models such as T5 (Raffel et al., 2019) and BART (Lewis et al., 2020), the model

architecture has remained largely the same. Radford et al. took advantage of the transformer architecture (Vaswani et al., 2017), which is superior and conceptually simpler than recurring neural networks to introduce the OpenAI GPT model. Specifically, they pre-trained a left-to-right transformer decoder as a general language model using the GPT architecture. Subsequently, they fine-tuned the model for 12 different language understanding tasks by applying various transformations to the input. Later, GPT-2 (Radford et al., 2019) was introduced, a more advanced version of GPT that has more trainable parameters. The authors showed that as long as general language models have very high capacities, they can reach reasonable performance on many specific natural language processing tasks. The use of the transformer architecture is later expanded to include modalities other than natural language, such as images (Dosovitskiy et al., 2021), protein amino acid sequences (Lin et al., 2023; Rives et al., 2021), and SMILES string molecules (Chithrananda, Grand, and Ramsundar, 2020; Fabian et al., 2020). All the models above are pre-trained with the Masked Language Modeling task (MLM) introduced in BERT (Devlin et al., 2019) and are able mostly to perform discriminative tasks.

MULTIMODAL MODELS. The success of the transformer’s unimodal tasks made this architecture broadly studied for multimodal representation learning. An example is the CLIP (Contrastive Language-Image Pre-training) model (Radford et al., 2021), which is a transformer model that facilitates cross-modal understanding between images and text. It combines a ViT vision encoder with a transformer-based language encoder to learn joint representations of images and their associated textual descriptions. Using transformers in both vision and text encoders, the CLIP model benefits from its ability to capture long-range dependencies. Another example is MolT5 (Edwards et al., 2022) which is a self-supervised learning framework based on the T5 model (Raffel et al., 2019) for pre-training models on a vast amount of unlabeled natural language text and molecule SMILES strings. MolT5 is capable of bidirectional translation between molecule representations and natural language, allowing molecule captioning and generation by providing text prompts. ProtST (Xu et al., 2023), enhances the classification and retrieval capabilities of protein language models by co-training it with biomedical text. While ProteinDT (Liu et al., 2023) uses protein language models and pre-trained language models to perform text-guided protein generation. Both of the aforementioned text-protein multimodal frameworks take only the protein sequence into consideration to encode the proteins.

GRAPH NEURAL NETWORKS. Graph neural networks (GNNs) have emerged as a powerful framework for modeling and analyzing graph-structured data (Kipf and Welling, 2017; Scarselli et al., 2009). By iteratively exchanging and integrating information among nodes, GNNs can propagate and refine features throughout the graph, ultimately encoding a comprehensive understanding of the graph’s structure and semantics. This ability to capture complex relationships within graphs has contributed to the success of GNNs in various

domains, including social network analysis, recommendation systems, and bioinformatics (Chatzianastasis, Vazirgiannis, and Zhang, 2023; Zhang et al., 2021; Zitnik, Agrawal, and Leskovec, 2018). Numerous studies have suggested various enhancements and expansions to the GNNs' models. Some notable contributions include the introduction of more expressive and adaptable aggregation functions, such as those proposed by Murphy et al. (2019), Seo, Loukas, and Perraudin (2019) and Chatzianastasis et al. (2023). Moreover, several schemes have been developed to incorporate different local structures or high-order neighborhoods, as explored by Morris, Rattan, and Mutzel (2020) and Nikolentzos, Dasoulas, and Vazirgiannis (2020). Furthermore, the domain of GNNs has expanded to include heterogeneous graphs, where nodes and edges can have different types and semantics, leading to the development of heterogeneous graph neural networks that effectively handle such complex graph structures (Schlichtkrull et al., 2018; Zhang et al., 2019a).

PROTEIN REPRESENTATION LEARNING. In the field of protein representation learning, various approaches have emerged over the years, aiming to capture meaningful information from proteins using different data modalities and computational techniques. A prominent avenue of research is the focus of sequence-based representations that extract features solely from the amino acid sequences of proteins. To achieve this, deep learning techniques such as recurrent neural networks (RNN) and convolutional neural networks (CNN) have been applied, allowing the direct learning of representations from protein sequences (Bileschi et al., 2019; Liu, 2017). Drawing inspiration from the remarkable achievements of language models in Natural Language Processing (NLP), researchers have also developed pre-trained language models tailored specifically for proteins (Brandes et al., 2022; Lin et al., 2023). These models take advantage of large-scale protein datasets to learn powerful representations that can be subsequently used for various prediction tasks. In addition to sequence-based approaches, graph-based representations leverage the three-dimensional (3D) structure of proteins to capture their functional properties. Zhang et al. (2022b) proposed a graph neural network model with a contrastive pertaining strategy for function prediction and fold classification tasks. Chen et al. (2023) proposed a 3D-equivariant graph neural network, specifically designed to estimate the accuracy of protein structural models. Wang et al. (2022) used a hierarchical graph network, which captures the hierarchical relations present in proteins and learns representations at different levels of granularity. Hybrid approaches integrate multiple data modalities, such as protein sequences, structures, and functional annotations, to create comprehensive representations. These methods combine the strengths of sequence-based and graph-based approaches to capture various aspects of protein function. Gligorijević et al. (2021) proposed DeepFRI which combines sequence features extracted from a pre-trained protein language model with protein structures. Our work aims to leverage protein sequence and structure models to generate free text annotations of proteins.

6.3 METHODOLOGY

In this section, we present our proposed multimodal framework, **Prot2Text**, for generating protein function descriptions in free text. An illustration of the proposed architecture can be found in Figure 6.1.

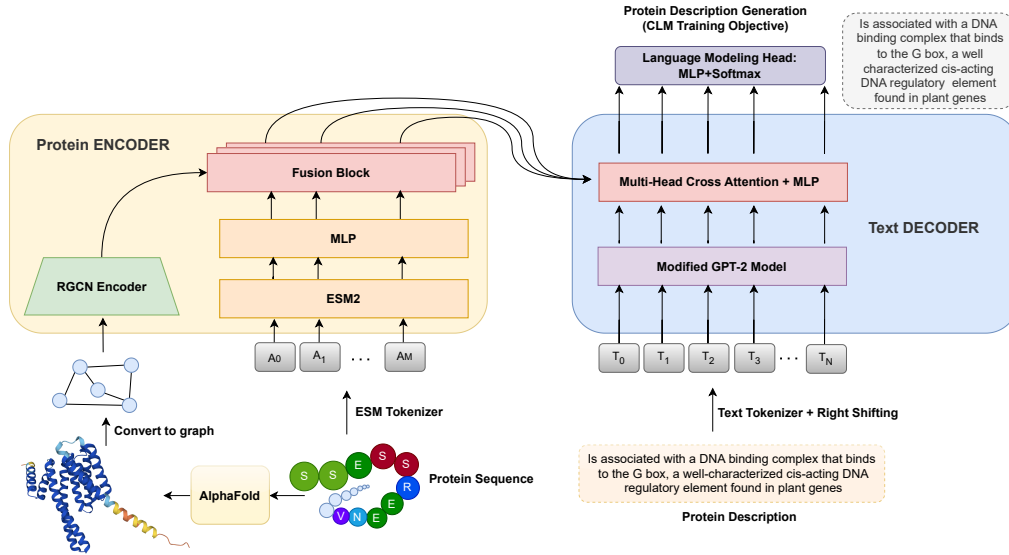


Figure 6.1 – Architecture of the proposed Prot2Text framework for predicting protein function descriptions in free text. The model leverages a multimodal approach that integrates protein sequence, structure, and textual annotations. The Encoder-Decoder framework forms the backbone of the model, with the encoder component utilizing an RGCN to process the protein graphs and an ESM model to process the protein sequence. A cross-attention mechanism facilitates the exchange of relevant information between the graph-encoded and the sequence-encoded vectors, creating a fused representation synthesizing the structural and textual aspects. The decoder component employs a pre-trained GPT-2 model, to generate detailed and accurate protein descriptions from the fused protein representation. By combining the power of GNN and LLM, Prot2Text enables a holistic representation of protein function, facilitating the generation of comprehensive protein descriptions.

GRAPH CONSTRUCTION. Upon obtaining the 3D proteins' structures using AlphaFold, we proceed to represent the proteins as a heterogeneous graph $G = (V, E, R)$ (inspired from (Qabel et al., 2023)), where $V = [N] := \{1, \dots, N\}$ is the set of vertices representing the amino-acids of the proteins, $E \subseteq V \times V$ is the set of edges representing various interactions between the nodes and R is a set of different edge interactions. Each node u is associated with a vector of characteristics $x_u \in \mathcal{R}^d$, which includes relevant information such as local structural characteristics and the physicochemical properties of the associated amino acids.

The included node features for each amino acid are the following: PHI and PSI angles, RSA (Relative Solvent Accessibility, it provides information about the

exposure of amino acid residues on the surface of a protein in relation to their accessibility to the solvent), ASA (Accessible Surface Area, it represents the accessible surface area of a molecule, including amino acids within a protein, refers to the surface area of the molecule that is exposed to the solvent and can interact with other molecules), and secondary structure (refers to the local folding patterns within a polypeptide chain and is primarily characterized by two common structures: alpha helices (α -helices) and beta sheets (β -sheets)). This enables the graph to retain fine-grained information critical to the protein’s structure and function. To model the diverse interactions and relationships between amino acids, we introduce different types of edges connecting the nodes. Therefore, each edge $i = (v, u)$ is associated with an edge type $e_i \in \mathcal{R}$. Sequential edges are employed to connect adjacent nodes in the protein sequence, effectively representing the sequential order of amino acids and capturing the linear arrangement of the protein’s primary structure. This sequential information is crucial for understanding the folding patterns and functional motifs within the protein. Additionally, we utilize spatial edges to establish connections between nodes that are in close spatial proximity within the 3D structure of the protein. These edges play a pivotal role in encoding the protein’s tertiary structure and folding patterns, enabling us to capture the intricate spatial arrangements of amino acids within the protein’s core. We further extend the graph construction to include hydrogen-bond interactions as an additional edge type. Hydrogen bonds are fundamental noncovalent interactions that are of paramount importance for stabilizing protein structures and enabling specific molecular recognition events. Through the integration of the different edge types, our comprehensive protein graph provides a more holistic and detailed representation of the protein structure while capturing both short and long-range interactions. In addition, we add the peptide bond, the k nearest neighbors, and the Delaunay triangulation (connectivity of amino acid residues in three-dimensional space) as edge types.

GRAPH ENCODING. To encode the protein graph G into a vector $h_G \in \mathcal{R}^{d_{out}}$, we employ a Relational Graph Convolutional Neural Network (RGCN) (Schlichtkrull et al., 2018), which effectively considers the various edge types present in the graph in the message passing mechanism. We denote the neighborhood of type r of a vertex u by $\mathcal{N}_r(u)$ such that $\mathcal{N}_r(u) = \{v : (v, u) \in E_r\}$, where E_r is the set of edges with the type of edge r . In layer k of the GNN, we update the node representations as follows:

$$x_i^k = \sigma \left(W_{\text{root}}^k \cdot x_i^{k-1} + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} \frac{1}{|\mathcal{N}_r(i)|} W_r^k \cdot x_j^{k-1} \right), \quad (6.1)$$

where W_{root}^k represents the learnable weight matrix for the root transformation in layer k , W_r^k denotes the learnable weight matrix of layer k for relation r and $\sigma(\cdot)$ is an element-wise activation function such as ReLU. This formulation allows nodes to update their representations by incorporating information from neighboring nodes based on the specific edge types, capturing the structural and relational dependencies within the protein graph. To obtain the graph

representation from the node representations of the last layer K of the GNN, we apply a mean-pooling layer as follows:

$$h_G = \frac{1}{N} \sum_{i=1}^N x_i^K \quad (6.2)$$

The resulting vector h_G serves as an informative encoded representation of the protein graph, capturing the essential structural and relational characteristics. This representation plays a crucial role in the subsequent text generation process, where it will be utilized to generate detailed and accurate protein functions.

SEQUENCE ENCODING. To encode the protein sequence P_S , we used ESM2-35M (Lin et al., 2023) as our base model. ESM2 is a protein language model that uses a transformer-based architecture to learn the interaction patterns between pairs of amino acids in the input sequence. This allows the ESM model to capture evolutionary information about amino acid sequences about proteins and their properties. To achieve uniform representation dimensions for all modalities within the spatial domain, a projection layer is applied after the last hidden layer of the ESM model. This layer functions as a projection layer that transforms the individual amino acid representations, derived from the ESM embedding dimension, into the graph embedding dimension d_{out} . As a result, a matrix denoted $H_S^0 \in \mathcal{R}^{N, d_{out}}$ is formed, which contains the amino acid representations:

$$H_S^0 = ESM(P_S)W_p \quad (6.3)$$

where W_p is a trainable matrix.

MULTIMODAL FUSION To obtain the final protein encoding, we utilize a fusion block that combines the representation of each amino acid inside the matrix H_S^0 with the graph representation vector h_G . The fusion process involves a simple element-wise addition of the two representations, followed by a projection layer. This fusion block enables the integration of information from both the sequence and the graph representations in a straightforward manner. Thus, each amino acid is contextually enriched with information from the graph representation. Additionally, a normalization layer is applied after each fusion block to maintain stable training and further enhance the learning process. Specifically, for each amino acid representation in H_S^k , and the graph representation h_G , the fusion block computes the combined representation H_S^{k+1} as follows:

$$H_S^{k+1} = \left(H_S^k + \mathbf{1}_n h_G W_V^k \right) W_O^k, \quad (6.4)$$

where W_V^k and W_O^k are trainable matrices in the fusion block k and $\mathbf{1}_n$ is a vector of ones of size n (length of the amino acid sequence).

By using this fusion block multiple times in the architecture (four times in this case), the model can capture complex interactions and dependencies between the sequence and graph representations, leading to an effective and contextually enriched encoding of the protein data. The fusion block could be seen as a special case of the transformer cross-attention block when the input from the encoder represents only one token.

TEXT GENERATION We employed the transformer decoder architecture to generate protein descriptions. We initialized the main components of the decoder, namely the text embedding matrix, self-attention, and language modeling head, with the pre-trained weights of GPT-2. In doing so, we leveraged the GPT-2 model’s capacity to grasp the underlying textual semantics. We forward the protein representation obtained from the protein encoder as input to the multi-head cross-attention module within the transformer decoder. This interaction enabled the model to effectively incorporate context from protein representation, contributing to the generation of coherent and meaningful protein descriptions. We adopted the identical vocabulary and tokenizer from the GPT-2 model, with the introduction of two unique special tokens. These additional tokens serve as essential markers that allow the model to discern the precise boundaries of each protein description within the input text. In the training phase, we employed Causal Language Modeling (CLM) as the training objective to optimize our model. Causal Language Modeling involves training the model to predict the next token in a sequence given the preceding tokens. This unidirectional prediction process ensures that the model generates text in a causal manner, without access to future tokens. The maximum length of each description is 256 tokens.

6.4 EXPERIMENTAL RESULTS

DATASET To train the Prot2Text framework using protein structures, sequences, and textual descriptions, we build a multimodal dataset with 256,690 proteins. For each protein, we have three information: the corresponding sequence, the AlphaFold accession ID, and the textual description. To build this dataset, we used the SwissProt database (Bairoch and Apweiler, 1996), including UniProtKB (Consortium, 2016). Release 2022_04. Swiss-Prot is a widely recognized and trusted resource in the field of protein sequence analysis as a larger and more rigorously curated dataset. Many researchers and studies also rely on Swiss-Prot as a primary source of high-quality protein data. Its extensive manual curation and annotation processes ensure a high level of data quality and accuracy. Initially, the SwissProt database in this release has 568,363 proteins on which we perform the following: (1) Select the following properties: name that gives the full name of the protein, sequence that gives the amino acid sequence of the protein, AlphaFoldDB that gives the accession ID of the protein in the AlphaFold database, taxon and text that gives the textual description of the protein. (2) Eliminate all samples that do not have all three information (modalities). (3) Remove all samples with a duplicate amino acid sequence. (4) Remove all the samples where the textual description contains “*By Similarity*”. (5) Apply the CD-HIT clustering algorithm (Li and Godzik, 2006) to create a training / validation / test scheme with 248,215 / 4,172 / 4,023 proteins, respectively. The maximum similarity threshold between the sets (train, validation test) used in the CD-HIT algorithm is 40%. (6) Preprocess the textual description to remove “*PubMed*” information. The AlphaFoldDB

accession is then used to download the protein structure in a ".PDB" file format using AlphaFoldDB version 4.

6.5 TOKENIZATION

PROTEINS TEXTUAL DESCRIPTION The training dataset consists of 256,690 proteins with a unique amino acid sequence. However, some proteins have the same description. In total, the training dataset has 48,251 unique function descriptions. The average number of tokens per description is 57.51. We chose to truncate all the descriptions during the tokenization to a maximum length of 256 since this number of tokens covers 98.7% of all the descriptions as we can see in figure 6.2.

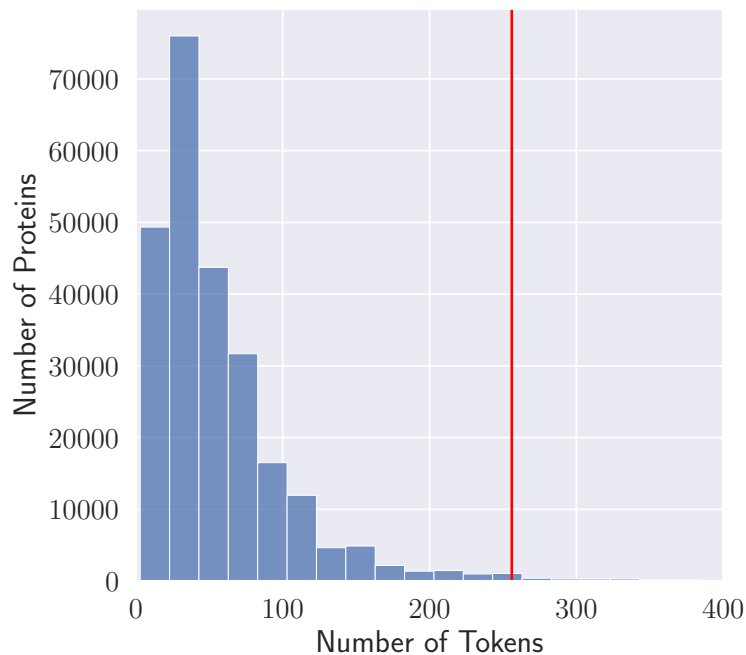


Figure 6.2 – Analyzing protein description lengths: Distribution of tokens per sample with threshold highlight at 256 tokens (in red).

TOKENIZER The Prot2Text tokenizer is an instance of the GPT2 tokenizer with two additional tokens. In GPT2 model, the pad token, the start of sequence token, and the end of sequence token share the same index. As the Prot2Text architecture is an encoder-decoder architecture, we chose to separate the three tokens by adding two extra tokens representing the start of the sequence and the end of sequence. For both added tokens, we equally need to add the corresponding embedding to the GPT2 word embedding matrix while keeping the rest of the matrix intact.

BASELINES. In our experimental evaluation, we used a comprehensive set of baselines to rigorously assess the performance of the text generation of the Prot2Text framework. Specifically, we compared our approach against

unimodal encoders, namely, RGCN, ESM, and a vanilla-transformer trained from scratch. These encoders focus exclusively on either the protein graph or the protein sequence representation. Furthermore, we compared it with a multimodal baseline, RGCN+ESM, that concatenates the graph and sequence representations without fusing the representation of each amino acid and the structure representation. Finally, we compare to the RGCN \times vanilla-transformer baseline, which has a similar architecture as Prot2Text, but instead uses a vanilla-transformer model from scratch instead of the pre-trained ESM2. In all ESM models, we use the last hidden state. The vanilla-transformer baseline follows the same configuration and number of parameters as the pre-trained ESM2-35M.

TRAINING DETAILS. We implemented all models using PyTorch and used 64 NVIDIA V100 GPUs for training. We used the AdamW optimizer (Kingma and Ba, 2015) with $\epsilon = 10^{-6}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, with a learning rate starting from 2.10^{-4} and decreasing to zero using a cosine scheduler. We used a warm-up of 6% of the total training steps. We fixed the batch size to four per GPU, and trained the models for 25 epochs. For the GNN encoder, we used six layers with a hidden size equal to the hidden size of GPT-2 (768 for the base model of GPT-2) in each layer. Regarding the tokenization of amino acid sequences, we used the same tokenizer and configuration of ESM2 including the hidden layer and hidden dimension. Training for each Base model lasted approximately 12 hours. All experiments were carried out using the Hugging Face transformers library (Wolf et al., 2020).

6.6 TEXT GENERATION

To generate the protein textual description during and after the training, we used the generation function implemented in the transformers library. We used the default generation parameters of `length_penalty=1.0`, `no_repeat_ngram_size` is set to 0 and `early_stopping=False`. The text generation was done during the training on the validation set, each 500 training steps using greedy search (number of beams equal to one) with a maximum length of 256 tokens per sample. However, a different configuration could be used, leading to multiple functions.

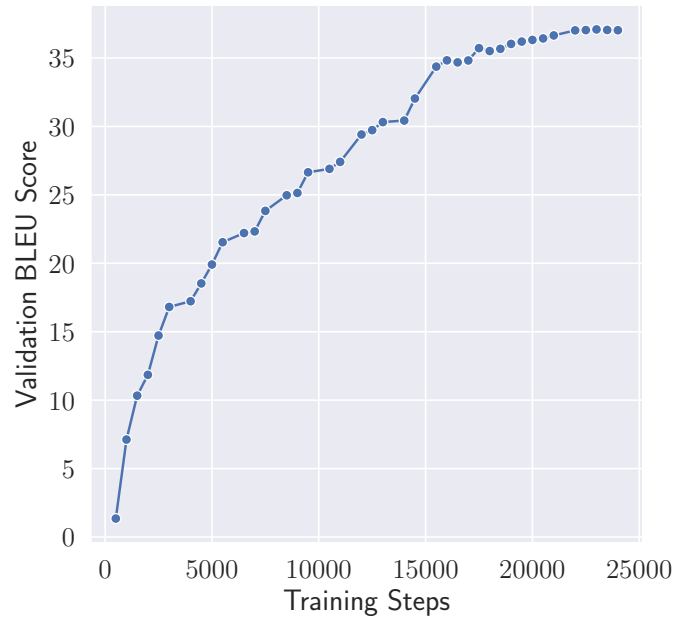


Figure 6.3 – Tracking Prot2Text_{BASE} BLEU score progression on validation set across training iterations. Higher is better.

Figure 6.3 shows the BLEU score validation throughout the training for the Prot2Text_{BASE} model. The validation BLEU score starts to stabilize after the 20th epoch, reaching the best validation BLEU score of 37.09 at step 23000. The checkpoint with the best validation score is used to compute the test scores.

CO₂ EMISSION RELATED TO EXPERIMENTS. Experiments were conducted using a private infrastructure, with a carbon efficiency of 0.057 kgCO₂eq/kWh. A cumulative of 23000 hours of computation was performed on hardware of type Tesla V100-SXM2-32GB (TDP of 300W). Total emissions are estimated to be 393.3 kgCO₂eq of which zero percent were directly offset. Estimations were conducted using the MachineLearning Impact calculator¹ presented in (Lacoste et al., 2019).

METRICS. In the experiments, we used several metrics to evaluate the performance of the model on the text generation task. Specifically, we used *BLEU Score* (Papineni et al., 2002), *ROUGE-1*, *ROUGE-2* and *ROUGE-L* scores (Lin, 2004), and *BERT Score* (Zhang et al., 2020) described in Chapter 2. In our experiments, we choose to use BioBERT_{LARGE}-cased v1.1 (Lee et al., 2020) to calculate the *BERT Score*.

RESULTS. We report the results in Table 6.1, for different encoder models, including unimodal encoders such as vanilla-transformer, ESM2-35M and RGCN, and multimodal encoders such as RGCN × vanilla-transformer and RGCN + ESM2-35. All models use a GPT-2 decoder. The unimodal vanilla-transformer baseline, based solely on the amino acid sequence of the protein, exhibits the

1. <https://mlco2.github.io/impact#compute>

Model	# Params	BLEU Score	ROUGE-1	ROUGE-2	ROUGE-L	BERT Score
vanilla-transformer	225M	15.75	27.80	19.44	26.07	75.58
ESM2-35M	225M	32.11	47.46	39.18	45.31	83.21
RGCN	220M	21.63	36.20	28.01	34.40	78.91
RGCN + ESM2-35M	255M	30.39	45.75	37.38	43.63	82.51
RGCN \times vanilla-transformer	283M	27.97	42.43	34.91	40.72	81.12
Prot2Text_{BASE}	283M	35.11	50.59	42.71	48.49	84.30

Table 6.1 – Prot2Text_{BASE} achieves the highest performance in all evaluation metrics, including the BLEU score, ROUGE scores, and BERT Score.

lowest performance in all evaluation metrics. However, we observe a significant improvement in performance when using the RGCN unimodal graph encoder. The RGCN outperforms the vanilla-transformer by more than five absolute points in terms of BLEU score and three points in terms of BERT score. This performance disparity highlights the importance of incorporating structural information through the RGCN encoder for protein function prediction. On the other hand, using the pre-trained protein language model ESM2-35M instead of initializing the vanilla-transformer randomly, results in a remarkable improvement in performance. The ESM2-35M encoder leads to a substantial increase of more than 16 BLEU score points and 18 ROUGE-L points compared to the standard vanilla-transformer configuration. This notable enhancement can be attributed to the pre-training of ESM2-35M using masked protein modeling, which enables the encoder to capture intricate relationships and patterns within protein sequences. In the context of multimodal protein representation, the evaluation results demonstrate that Prot2Text_{BASE} shows superior performance in all evaluation metrics. In particular, it achieves the highest BLEU score of 35.11, the highest ROUGE-1 score of 50.59, the highest ROUGE-2 score of 42.71, the highest ROUGE-L score of 48.49 and the highest BERT score of 84.3. These results highlight the effectiveness of combining protein structure and amino acid information in a multimodal manner. Incorporation of protein structure, facilitated by the Relational Graph Convolutional Network (RGCN) with sequential representations of ESM2-35 amino acids, significantly improves overall performance in all evaluation metrics. This improvement is attributed to the enriched understanding of proteins achieved through the synergy of these two modalities. Furthermore, the efficacy of the multimodal fusion approach is corroborated by the results obtained from the RGCN \times vanilla-transformer. Introducing structural information using RGCN into the randomly initialized vanilla-transformer yields a substantial improvement of more than 10 BLEU score points compared to using the vanilla-transformer alone, and an improvement of more than 6 BLEU score points over using RGCN in isolation. Finally, to show the importance of the fusion block in the Prot2Text framework, we compare it to RGCN + ESM2-25, which concatenates the representation of the protein structure with the representation of amino acids. In this case, the graph representation will simply be passed to the decoder along with the ESM output. We notice that using this strategy leads to slightly worse results than using ESM alone. This not only provides support for the selection of the fusion block employed in Prot2Text, but also suggests that indiscriminately increasing the

overall parameter count of the model could potentially lead to a degradation in its performance.

Model	# Params	BLEU Score	ROUGE-1	ROUGE-2	ROUGE-L	BERT Score	Inference Time
Prot2Text _{SMALL}	256M	30.01	45.78	38.08	43.97	82.60	1,225
Prot2Text _{BASE}	283M	35.11	50.59	42.71	48.49	84.30	1,379
Prot2Text _{MEDIUM}	398M	36.51	52.13	44.17	50.04	84.83	1,334
Prot2Text _{LARGE}	898M	36.29	53.68	45.60	51.40	85.20	1,667

Table 6.2 – Larger models outperform their smaller counterparts across most evaluation metrics, indicating the benefits of employing larger language models in the Prot2Text framework. The Prot2Text_{MEDIUM} model strikes an optimal balance between performance and computational efficiency. This configuration demonstrates improved performance compared to the smaller model while still maintaining reasonable computational costs. The inference time is in seconds for text generation for each model on the whole test set. The inference time here is computed during text generation using two NVIDIA RTX 6000 with 48GB memory in parallel and batch size of four per device.

SCALING TO LARGER MODELS. We conducted a study to assess the performance of our Prot2Text framework as we varied the number of parameters. The primary objective of this experiment was to evaluate the benefits of employing larger models in terms of generating more accurate and detailed textual representations of protein function. To conduct the ablation study, we systematically varied the size of the protein language model (ESM). Where Prot2Text_{SMALL}, Prot2Text_{BASE}, Prot2Text_{MEDIUM} and Prot2Text_{LARGE} use ESM2-8M, ESM2-35M, ESM2-150M and ESM2-650M, respectively. We evaluated each configuration on the same test set of proteins and used the same evaluation metrics as described above. The results of the ablation study, presented in Table 6.2, show a trend of performance improvement as we scale up the architecture of the model. Larger versions of the ESM outperformed their smaller counterparts in most evaluation metrics. The increase in model size led to more accurate and relevant descriptions, indicating the benefit of using larger language models in the Prot2Text framework. However, a complementary analysis that included the corresponding computation time showed an increase in the inference cost after the use of larger models (a higher number of parameters). Therefore, **Prot2Text_{MEDIUM}** (398M parameters) is a good trade-off that strikes the balance between performance and time cost. Furthermore, in Figure 6.4 we report the performance of all Prot2text models with respect to different similarity thresholds. The similarity represents the highest alignment score between the amino acid sequences of the test and training sets using the BLAST identity. We observe that for test proteins with low similarity scores with the train set (between 20% and 30%) and for proteins without a counterpart in the train set, Prot2Text_{MEDIUM} is the dominant, while for higher similarity scores, Prot2Text_{LARGE} performs better.

VISUALIZATION OF GENERATED DESCRIPTIONS. To gain deeper insights into the quality of the functions generated by our *Prot2Text* framework, we provide in Figure 6.5 a textual comparison of the predefined labels and gener-

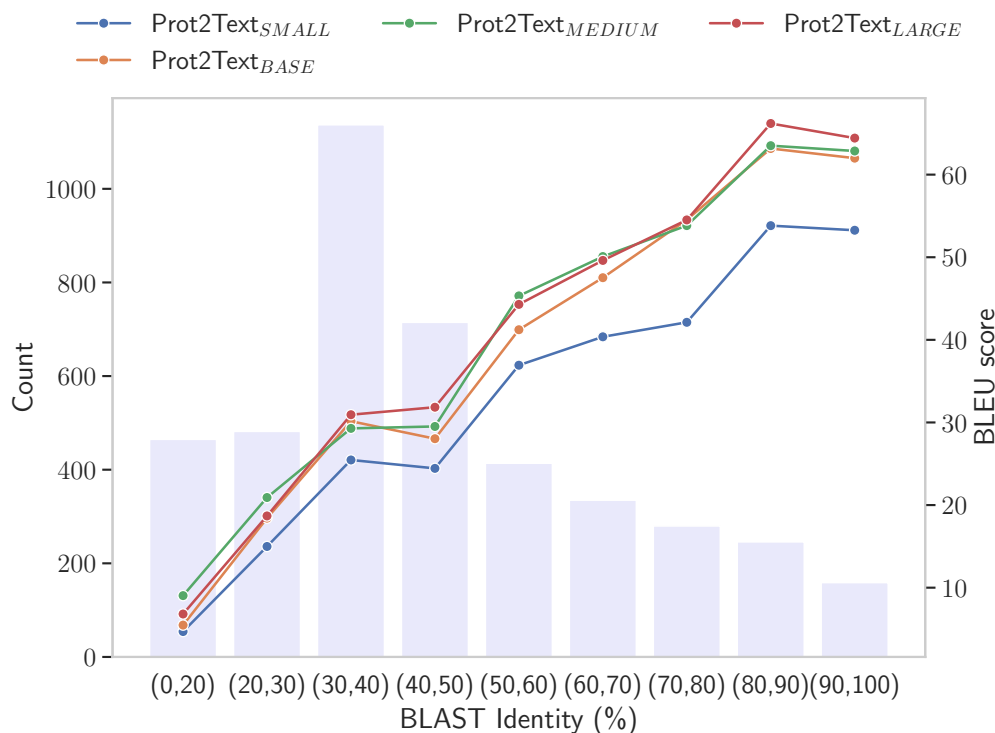


Figure 6.4 – The test BLEU score for Prot2Text models as a function of the percentage identity using BLAST hit between the test and the train sets.

ated text outputs for a selected set of proteins from the test set. It illustrates a comparison between ground truth and the corresponding descriptions generated by *Prot2Text* for three different proteins (*P36108*, *Q8NG08* and *P35713*) along with the name of each protein, amino acid sequence, and 3D structural representation. The results indicate a successful detailed reconstruction of the functions of the different proteins that includes richer information than the known description. Following, Fig. 6.5 showcases the model’s ability to generate coherent and informative free-text descriptions that align closely with the ground truth annotations.

6.7 CONCLUSION

In conclusion, our work introduces Prot2Text, a pioneering multimodal framework, for the accurate prediction of a protein’s function in free text format, from graph and sequential input. By reformulating the task as free-text prediction, we address the limitations of traditional classification-based methods, allowing for a more nuanced and in-depth understanding of a protein’s functionality. Leveraging the power of GNNs and LLMs, we integrated structural and textual protein information, resulting in highly detailed and coherent generated protein descriptions. The release of a comprehensive multimodal protein dataset further empowers the scientific community to benchmark and advance the field of protein function prediction in free text format. This innovative approach

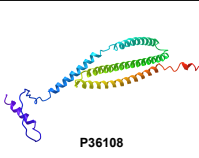

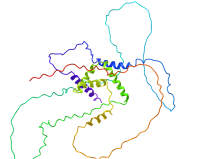
Protein	Reference Text	Generated Text
 <p>P36108</p>	<p>Required for the sorting and concentration of proteins resulting in the entry of these proteins into the invaginating vesicles of the multivesicular body (MVB). Acts as a component of the ESCRT-III complex, which appears to be critical for late steps in MVB sorting, such as membrane invagination and final cargo sorting and recruitment of late-acting components of the sorting machinery. The MVB pathway requires the sequential function of ESCRT-O, -I, -II and -III complex assemblies. Can directly stimulate VPS4 ATPase activity. The DID4/VPS2-VPS24 subcomplex is required for the VPS4-dependent dissociation of ESCRT-III.</p>	<p>Probable peripherally associated component of the endosomal sorting required for transport complex III (ESCRT-III) which is involved in multivesicular bodies (MVBs) formation and sorting of endosome-derived cargo proteins into MVBs. MVBs contain intraluminal vesicles (ILVs) that are generated by invagination and scission from the limiting membrane of the cell and mostly are delivered to lysosomes enabling degradation of membrane proteins, such as stimulated growth factor receptors, lysogenic enzymes and lipids. The MVB pathway appears to require the sequential function of ESCRT-O, -I, -II and -III complexes. ESCRT proteins mostly dissociate from the invaginating membrane before the ILV is released. The ESCRT machinery also functions in topologically equivalent membrane fission events, such the terminal stages of cytokinesis and the budding of enveloped viruses (HIV-1 and other lentiviruses). ESCRT I, -II and III proteins mostly mediate the sorting of proteins into lysophospholipids, primarily cytoplasmic membrane vesicle-derived proteins. ESCR-III proteins are believed to mediate a specialized role in the final stages</p>
 <p>Q8NG08</p>	<p>5'-3' DNA helicase involved in DNA damage response by acting as an inhibitor of DNA end resection. Recruitment to single-stranded DNA (ssDNA) following DNA damage leads to inhibit the nucleases catalyzing resection, such as EXO1, BLM and DNA2, possibly via the 5'-3' ssDNA translocase activity of HELB. As cells approach S phase, DNA end resection is promoted by the nuclear export of HELB following phosphorylation. Acts independently of TP53BP1. Unwinds duplex DNA with 5'-3' polarity. Has single-strand DNA-dependent ATPase and DNA helicase activities. Prefers ATP and dATP as substrates. During S phase, may facilitate cellular recovery from replication stress.</p>	<p>DNA-dependent ATPase and 5'-3' DNA helicase required for the maintenance of genome stability. Involved in various processes such as transcription-coupled nucleotide excision repair, mitotic spindle assembly, DNA damage response and DNA repair. During DNA double-strand break (DSB) repair, involved in limiting the number of single-stranded DSBs through DNA end resection by promoting the recruitment of RAD51 to sites of DNA damage. In addition to DNA end reannealing, also promotes the resumption of DNA synthesis from the new 3'-tail end of the existing DNA ends. Acts as a regulator of telomerase by inhibiting telomeric silencing; binds to telomeryosin, and is required for telomeronucleolytic silencing of endogenous small interfering RNAs (endo-siRNAs) and their targeting to the telomere. Involves also DNA damage signaling via the 5'-5' helicase and DNA helicases activities, preventing disintegration of the DNA duplex and subsequent resection of the damaged DNA strand. In vitro able to unwind 5'-overhanging flap DNA and catalyzes ATP-dependent unwinding of 5'-DNA ends. Also plays a role in</p>
 <p>P35713</p>	<p>Transcriptional activator that binds to the consensus sequence 5'-AACAAAG-3' in the promoter of target genes and plays an essential role in embryonic cardiovascular development and lymphangiogenesis. Activates transcription of PROX1 and other genes coding for lymphatic endothelial markers. Plays an essential role in triggering the differentiation of lymph vessels, but is not required for the maintenance of differentiated lymphatic endothelial cells. Plays an important role in postnatal angiogenesis, where it is functionally redundant with SOX17. Interaction with MEF2C enhances transcriptional activation. Besides, required for normal hair development.</p>	<p>Transcriptional activator that binds to the consensus sequence 5'-AACAAAG-3'. Plays an essential role in cellular differentiation, proliferation and survival. Plays a critical role in macrophage differentiation, migration and invasion, particularly in the gut. Required for normal gene expression in the macrophages, which are activated by biglycan-producing bacteria and fungi. Also required for normal chemotaxis. Plays important roles in the development of the central nervous system, where it is required for proper proliferation and migration of progenitor cells.</p>

Figure 6.5 – Ground-truth description vs text-free generated description: A textual comparison of the predefined labels and generated text output for three different proteins from the test set. The used generation configuration if these examples are the following: `length_penalty = 2.0`, `no_repeat_ngram_size=3` and `early_stopping=True`.

opens new horizons for research and applications in drug discovery, protein engineering, and various biological sciences, with the potential to revolutionize our understanding of protein functions.

6.8 LIMITATIONS AND FUTURE WORK

One limitation of our proposed Prot2Text model is that the RGCN encoder is not pretrained. Unlike the ESM encoder, which benefits from pretraining on a large corpus, the RGCN encoder lacks this initial knowledge. As a result, the RGCN encoder may struggle to capture complex patterns and may not fully leverage the underlying protein data, potentially leading to suboptimal performance. To address this limitation, we aim to explore pretraining techniques specifically tailored for graph neural networks. This could involve pre-training the RGCN encoder on auxiliary graph-related tasks, leveraging graph-level or node-level information to build a foundational understanding of protein structures.

CONCLUSION

IN this chapter, we conclude the dissertation by providing a brief summary of our primary contributions, which have been discussed in detail in the previous chapters. Additionally, we discuss the limitations and highlight several interesting areas for future research that require further investigation.

7.1 CONTRIBUTIONS AND LIMITATIONS DISCUSSION

Pretrained Language Models

We contributed two different pre-trained language models. In the field of the French legal domain, we have made a contribution by developing a pre-trained language model called JuriBERT. Similarly, we have also developed a pre-trained language model called GreekBART, the first seq2seq model specifically for the Greek language. In addition, we contributed the first Greek seq2seq dataset for news summarization, GreekSUM, which contains two natural language generation tasks, title generation, and abstract generation.

Both JuriBERT and GreekBART have very competitive performances in their respective fields despite being smaller models than the competitors in most cases.

Different challenges arise in the area of pretrained language models for low-resource languages and domains. Mostly, we mention the lack of well-annotated generation datasets. For example, despite the effort to collect and create GreekSUM, it is constrained by its limitation to news articles from a single website. In addition, for some specialized domains, there is even a lack of raw-text pretraining corpora for the non-English language, which we can see clearly in the JuriBERT case, where the existing corpus was able to train only the SMALL model variant efficiently. Thus, the importance of research focusing on datasets and corpora creation is highlighted.

Word Sense Induction

In Chapter 5, we proposed a new method to address some of the problems faced by the task of word sense induction. our contributions encompass a novel unsupervised methodology that combines pre-trained language models, hierarchical clustering, and mutual information maximization. This approach not only overcomes certain limitations present in previous endeavors but also demonstrates competitive performance. For instance, we introduced a new technique for estimating the dynamic number of senses in target words, leveraging the quantification of word polysemy instead of using a fixed number of clusters. Furthermore, our investigation of performance variation concerning

the depth of the selected layer, as presented in Section 5.7, in four different models, offers valuable insights for researchers involved in future work on word sense induction.

However, multiple problems still exist for the word sense induction task. For example, the choice of the dynamic cluster number still requires lower and upper limits, which are not totally automated. In addition, we argue that the evaluation of word sense induction misses the ability to detect new senses or to divide an old sense into multiple ones. Since language is always evolving, the aforementioned changes can easily happen and the language model can capture the new senses during pretraining, however, still be considered as a wrong clustering solution by the current ground truth. Thus, the importance of finding new evaluation systems for unsupervised NLP tasks.

Furthermore, different investigations highlight the best-performing layer of a transformer-based pre-trained model. We noticed that this layer is not fixed for different tasks, indicating that the transformer can learn different patterns from different layers, which is also validated by our experiments in Section 5.7, where the layer for word sense representation was different from the one for the WMT16 task. This is worth further investigation to find an unsupervised method to detect the best layer or even a combination of layers. This could dramatically improve performance in different tasks upon model finetuning.

Prot2Text

In Chapter 6, we introduced Prot2Text, the first multimodal architecture that combines graph-based and text-based data modalities. Prot2Text leverages Graph Neural Networks (GNNs) and Evolutionary Scale Modeling (ESM) to encode proteins in a fused representation, with a pre-trained GPT-2 model decoding detailed textual descriptions of protein functions. Beyond the framework, we introduced and explored various baselines, highlighting the superiority of integrating graph- and sequence-based protein information. Additionally, a comprehensive multimodal protein dataset, including 256,690 instances, has been released to the research community. This dataset, comprising protein structures, sequences, and textual function descriptions, serves as a valuable resource for benchmarking and comparison, fostering advancements, and standardizing the evaluation of protein function prediction methods in free text format. Overall, these contributions not only expand our understanding of protein functions, but also equip researchers with powerful tools to propel further developments in this critical field.

However, multiple challenges and limitations remain. We mainly mention the evaluation process, which is for now only based on trivial natural language generation metrics such as BLEU, ROUGE, and BERT scores. Although they are well-studied and used metrics in language generation, they cannot guarantee the factuality of the generated protein function, which requires biologist intervention to judge its quality and usefulness, which cannot be done for billions of unstudied proteins. Additionally, Prot2Text uses the structure of the protein predicted by AlphaFold, which is for some cases (especially for

unknown proteins) not confident.

Furthermore, in terms of architecture, the weights of the GNN model and the cross-attention layers are randomly initialized, as opposed to the self-attention weights of the text decoder initialized from GPT2 and the protein sequence encoder initialized from ESM2, which can make training suboptimal. In Section 7.2, we propose some future research directions to overcome some of the limitations described.

7.2 FUTURE RESEARCH DIRECTIONS

The exponential growth and transformative capabilities demonstrated by large language models (LLMs) in the previous year present a compelling avenue for future research endeavors. Leveraging the remarkable abilities of LLMs, particularly their proficiency in capturing intricate patterns and semantic nuances within vast datasets, holds immense promise for diverse applications. As a next step, exploring the adaptability and performance of LLMs in specialized domains and applications emerges as a key focus.

LOW RESOURCES LANGUAGES APPLICATIONS USING LLMS As we discuss in Chapter 4, mBART50 performs very well despite not seeing the Greek language during pre-training. This somehow implicates a correlation between some languages that can improve the performance on unrepresented languages in the pre-training corpus. With the power of multilingual LLMs, this idea can be investigated, providing a well-annotated dataset with few-shot learning. For example, GreekSUM tasks can be tested on different monolingual and multilingual LLMs to study how the Greek language and other low-resource languages interact with other languages. This can help later, building a larger pre-training corpus for low-resource languages.

DECODER-ONLY PROTEIN TO TEXT MULTIMODAL MODEL In order to reduce the number of randomly initialized parameters and to make use of the potential of LLMs, in the future, we would like to explore the incorporation of a pre-trained GNN architecture in a decoder-only large language model without any additional cross-attention layer. To do so, we first explore pre-training techniques specifically tailored for graph neural networks, such as masked edge prediction, to build a foundational understanding of protein structures. In addition, we would also like to explore the protein structure pre-training by applying contrastive learning with the larger ESM models. Second, we aim to use the graph and node representations as a combined input with the token embeddings to the self-attention layer of an LLM such as Llama 2 (Touvron et al., 2023b).

7.3 EPILOGUE

In recent times, we see Natural Language Processing (NLP) all around us, such as when we chat with our smart home assistant, use social media, or

translate text online. Big models are behind the scenes doing amazing things to make our lives easier. However, there are challenges and risks that come with this convenience. This dissertation is an important step toward dealing with some of these challenges and testing the ability of language models in different applications. We hope that in the future we can contribute even more and find better solutions. By exploring how protein functions can be predicted and coming up with new ideas like **Prot2Text**, we have started to make progress in using language models for specific areas such as biomedicine and legal (e.g. **JuriBERT**) domains. As we continue, we aim to learn more and find better ways to make the most of these large language models.

BIBLIOGRAPHY

- Abadji, Julien, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot (June 2022). « Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. » In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4344–4355. URL: <https://aclanthology.org/2022.lrec-1.463> (cit. on p. 40).
- Abdine, Hadi, Christos Xypolopoulos, Moussa Kamal Eddine, and Michalis Vazirgiannis (2021). « Evaluation Of Word Embeddings From Large-Scale French Web Content. » In: *CoRR abs/2105.01990*. arXiv: 2105.01990. URL: <https://arxiv.org/abs/2105.01990> (cit. on p. 25).
- Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa (2009). « A Study on Similarity and Relatedness Using Distributional and WordNet-Based Approaches. » In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL '09. Boulder, Colorado: Association for Computational Linguistics, 19–27. ISBN: 9781932432411 (cit. on p. 51).
- Agirre, Eneko and Aitor Soroa (June 2007). « SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. » In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 7–12. URL: <https://aclanthology.org/S07-1002> (cit. on p. 58).
- Ahuir, Vicent, Lluís-F. Hurtado, José Ángel González, and Encarna Segarra (2021). « NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish. » In: *Applied Sciences* 11.21. ISSN: 2076-3417. DOI: 10.3390/app11219872. URL: <https://www.mdpi.com/2076-3417/11/21/9872> (cit. on p. 38).
- Alayrac, Jean-Baptiste et al. (2022). « Flamingo: a Visual Language Model for Few-Shot Learning. » In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. URL: <https://openreview.net/forum?id=EbMuimAbPbs> (cit. on p. 5).
- Amigó, Enrique, Julio Gonzalo, Javier Artiles, and M. Verdejo (Oct. 2009). « Amigó E, Gonzalo J, Artiles J et alA comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inform Retrieval* 12:461-486. » In: *Information Retrieval* 12, pp. 461–486. DOI: 10.1007/s10791-008-9066-8 (cit. on p. 23).
- Amplayo, Reinald Kim, Seung-won Hwang, and Min Song (2019). « AutoSense Model for Word Sense Induction. » In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 6212–6219. DOI: 10.1609/aaai.v33i01.33016212. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/4580> (cit. on pp. 53, 61).

- Amrami, Asaf and Yoav Goldberg (2018). « Word Sense Induction with Neural biLM and Symmetric Patterns. » In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4860–4867. DOI: [10.18653/v1/D18-1523](https://doi.org/10.18653/v1/D18-1523). URL: <https://aclanthology.org/D18-1523> (cit. on pp. 53, 54, 58, 61).
- (2019). *Towards better substitution-based word sense induction*. arXiv: [1905.12598](https://arxiv.org/abs/1905.12598) [cs.CL] (cit. on pp. 52–54, 58, 60–62).
- Ansell, Alan, Felipe Bravo-Marquez, and Bernhard Pfahringer (Apr. 2021). « PolyLM: Learning about Polysemy through Language Modeling. » In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 563–574. DOI: [10.18653/v1/2021.eacl-main.45](https://doi.org/10.18653/v1/2021.eacl-main.45). URL: <https://aclanthology.org/2021.eacl-main.45> (cit. on pp. 52, 53, 56, 58, 61).
- Antoun, Wissam, Fady Baly, and Hazem Hajj (May 2020). « AraBERT: Transformer-based Model for Arabic Language Understanding. » English. In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. Marseille, France: European Language Resource Association, pp. 9–15. ISBN: 979-10-95546-51-1. URL: <https://aclanthology.org/2020.osact-1.2> (cit. on p. 62).
- Bagga, Amit and Breck Baldwin (1998). « Entity-Based Cross-Document Coreferencing Using the Vector Space Model. » In: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. URL: <https://aclanthology.org/C98-1012> (cit. on pp. 23, 59).
- Bairoch, Amos and Rolf Apweiler (Jan. 1996). « The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL. » In: *Nucleic Acids Research* 24.1, pp. 21–25. ISSN: 0305-1048. DOI: [10.1093/nar/24.1.21](https://doi.org/10.1093/nar/24.1.21) (cit. on pp. 9, 74).
- Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency (2019). « Multimodal Machine Learning: A Survey and Taxonomy. » In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.2, 423–443. ISSN: 0162-8828. DOI: [10.1109/TPAMI.2018.2798607](https://doi.org/10.1109/TPAMI.2018.2798607). URL: <https://doi.org/10.1109/TPAMI.2018.2798607> (cit. on p. 5).
- Bambroo, Purbid and Aditi Awasthi (2021). « LegalDB: Long DistilBERT for Legal Document Classification. » In: *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1–4. DOI: [10.1109/ICAECT49130.2021.9392558](https://doi.org/10.1109/ICAECT49130.2021.9392558) (cit. on p. 26).
- Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta (Sept. 2009). « The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. » In: *Language Resources and Evaluation* 43, pp. 209–226. DOI: [10.1007/s10579-009-9081-4](https://doi.org/10.1007/s10579-009-9081-4) (cit. on p. 59).
- Beltagy, Iz, Kyle Lo, and Arman Cohan (Nov. 2019). « SciBERT: A Pretrained Language Model for Scientific Text. » In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3615–3620. DOI: [10.18653/v1/D19-1170](https://doi.org/10.18653/v1/D19-1170).

- 18653/v1/D19-1371. URL: <https://aclanthology.org/D19-1371> (cit. on p. 26).
- Bileschi, Maxwell L, David Belanger, Drew Bryant, Theo Sanderson, Brandon Carter, D Sculley, Mark A DePristo, and Lucy J Colwell (2019). « Using deep learning to annotate the protein universe. » In: *BioRxiv*, p. 626507 (cit. on p. 70).
- Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2016). « Enriching Word Vectors with Subword Information. » In: *arXiv preprint arXiv:1607.04606* (cit. on p. 4).
- Brandes, Nadav, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial (2022). « ProteinBERT: a universal deep-learning model of protein sequence and function. » In: *Bioinformatics* 38.8, pp. 2102–2110 (cit. on pp. 1, 6, 18, 70).
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). « Language models are few-shot learners. » In: *Advances in neural information processing systems* 33, pp. 1877–1901 (cit. on pp. 15, 38, 67).
- Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez (2020). « Spanish Pre-Trained BERT Model and Evaluation Data. » In: *PML4DC at ICLR 2020* (cit. on p. 38).
- Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos (Nov. 2020). « LEGAL-BERT: The Muppets straight out of Law School. » In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2898–2904. DOI: 10.18653/v1/2020.findings-emnlp.261. URL: <https://aclanthology.org/2020.findings-emnlp.261> (cit. on pp. 26, 28).
- Chang, Baobao, Wenzhe Pei, and Miaohong Chen (Aug. 2014). « Inducing Word Sense with Automatically Learned Hidden Concepts. » In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 355–364. URL: <https://aclanthology.org/C14-1035> (cit. on p. 53).
- Chatzianastasis, Michail, Johannes Lutzeyer, George Dasoulas, and Michalis Vazirgiannis (2023). « Graph ordering attention networks. » In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pp. 7006–7014 (cit. on p. 70).
- Chatzianastasis, Michail, Michalis Vazirgiannis, and Zijun Zhang (2023). « Explainable Multilayer Graph Neural Network for Cancer Gene Prediction. » In: *arXiv preprint arXiv:2301.08831* (cit. on p. 70).
- Chen, Chen, Xiao Chen, Alex Morehead, Tianqi Wu, and Jianlin Cheng (2023). « 3D-equivariant graph neural networks for protein model quality assessment. » In: *Bioinformatics* 39.1, btado30 (cit. on pp. 19, 70).
- Chithrananda, Seyone, Gabriel Grand, and Bharath Ramsundar (2020). « ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. » In: *arXiv preprint arXiv:2010.09885* (cit. on p. 69).
- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke

- Zettlemoyer, and Veselin Stoyanov (July 2020). « Unsupervised Cross-lingual Representation Learning at Scale. » In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: [10.18653/v1/2020.acl-main.747](https://doi.org/10.18653/v1/2020.acl-main.747). URL: <https://aclanthology.org/2020.acl-main.747> (cit. on pp. 38, 39, 44, 46).
- Conneau, Alexis and Guillaume Lample (2019). « Cross-lingual language model pretraining. » In: *Advances in neural information processing systems* 32 (cit. on p. 38).
- Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov (2018). « XNLI: Evaluating Cross-lingual Sentence Representations. » In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2475–2485. DOI: [10.18653/v1/D18-1269](https://doi.org/10.18653/v1/D18-1269). URL: <https://aclanthology.org/D18-1269> (cit. on p. 45).
- Consortium, The UniProt (Nov. 2016). « UniProt: the universal protein knowledgebase. » In: *Nucleic Acids Research* 45.D1, pp. D158–D169. ISSN: 0305-1048. DOI: [10.1093/nar/gkw1099](https://doi.org/10.1093/nar/gkw1099) (cit. on p. 74).
- Consortium, UniProt (2015). « UniProt: a hub for protein information. » In: *Nucleic acids research* 43.D1, pp. D204–D212 (cit. on p. 68).
- Corrêa, Jr and Diego Amancio (Mar. 2018). « Word sense induction using word embeddings and community detection in complex networks. » In: *Physica A: Statistical Mechanics and its Applications* 523. DOI: [10.1016/j.physa.2019.02.032](https://doi.org/10.1016/j.physa.2019.02.032) (cit. on p. 53).
- Danon, Leon, Jordi Duch, Albert Diaz-Guilera, and Alex Arenas (June 2005). « Comparing community structure identification. » In: *Journal of Statistical Mechanics: Theory and Experiment* 2005. DOI: [10.1088/1742-5468/2005/09/P09008](https://doi.org/10.1088/1742-5468/2005/09/P09008) (cit. on p. 23).
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. » In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL: <https://aclanthology.org/N19-1423> (cit. on pp. xii, 4, 5, 7, 16, 26, 31, 37, 38, 46, 47, 53, 69).
- Dosovitskiy, Alexey et al. (2021). « An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. » In: *International Conference on Learning Representations* (cit. on pp. 1, 5, 69).
- Douka, Stella, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles (Nov. 2021). « JuriBERT: A Masked-Language Model Adaptation for French Legal Text. » In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 95–101. DOI: [10.18653/v1/2021.nllp-1.9](https://doi.org/10.18653/v1/2021.nllp-1.9). URL: <https://aclanthology.org/2021.nllp-1.9> (cit. on pp. 3, 25).

- Eddine, Moussa Kamal, Antoine J.-P. Tixier, and Michalis Vazirgiannis (2020). « BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. » In: *CoRR* abs/2010.12321. arXiv: 2010.12321. URL: <https://arxiv.org/abs/2010.12321> (cit. on p. 26).
- Eddine, Moussa Kamal, Antoine J. P. Tixier, and Michalis Vazirgiannis (2021). *BARThez: a Skilled Pretrained French Sequence-to-Sequence Model*. arXiv: 2010.12321 [cs.CL] (cit. on pp. 2, 5).
- Eddine, Moussa Kamal, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis (2022). *AraBART: a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization*. DOI: 10.48550/ARXIV.2203.10945. URL: <https://arxiv.org/abs/2203.10945> (cit. on p. 38).
- Edwards, Carl, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji (2022). « Translation between Molecules and Natural Language. » In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305 (cit. on pp. 6, 69).
- Elwany, Emad, Dave Moore, and Gaurav Oberoi (2019). « BERT Goes to Law School: Quantifying the Competitive Advantage of Access to Large Legal Corpora in Contract Understanding. » In: *CoRR* abs/1911.00473. arXiv: 1911.00473. URL: <http://arxiv.org/abs/1911.00473> (cit. on p. 26).
- Fabian, Benedek, Thomas Edlich, Hélène Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed (2020). *Molecular representation learning with language models and domain-relevant auxiliary tasks*. arXiv: 2011.13230 [cs.LG] (cit. on pp. 6, 69).
- Feinerer, Ingo and Kurt Hornik (2020). *wordnet: WordNet Interface*. R package version 0.1-15. URL: <https://CRAN.R-project.org/package=wordnet> (cit. on p. 51).
- Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*. Bradford Books (cit. on pp. 51, 58).
- Ferruz, Nicolás, Steffen Schmidt, and Birte Höcker (2022). « ProtGPT2 is a deep unsupervised language model for protein design. » In: *Nature Communications* 13.1, p. 4348. DOI: 10.1038/s41467-022-32007-7 (cit. on p. 18).
- Fey, Matthias and Jan E. Lenssen (2019). « Fast Graph Representation Learning with PyTorch Geometric. » In: *ICLR Workshop on Representation Learning on Graphs and Manifolds* (cit. on p. 10).
- Garneau, Nicolas, Eve Gaumont, Luc Lamontagne, and Pierre-Luc Déziel (2021). « CriminelBART: A French Canadian Legal Language Model Specialized in Criminal Law. » In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ICAIL '21. São Paulo, Brazil: Association for Computing Machinery, 256–257. ISBN: 9781450385268. DOI: 10.1145/3462757.3466147. URL: <https://doi.org/10.1145/3462757.3466147> (cit. on p. 26).
- Gligorijević, Vladimir, P Douglas Renfrew, Tomasz Kosciolk, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. (2021). « Structure-based protein function prediction using graph convolutional networks. » In: *Nature communications* 12.1, p. 3168 (cit. on p. 70).

- Goyal, Kartik and Eduard Hovy (Aug. 2014). « Unsupervised Word Sense Induction using Distributional Statistics. » In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 1302–1310. URL: <https://aclanthology.org/C14-1123> (cit. on p. 53).
- Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov (2018). « Learning Word Vectors for 157 Languages. » In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)* (cit. on p. 25).
- Ha, Jaeyoung, Hankum Park, Jongmin Park, and Seung Bum Park (2021). « Recent advances in identifying protein targets in drug discovery. » In: *Cell Chemical Biology* 28.3, pp. 394–423. ISSN: 2451-9456. DOI: <https://doi.org/10.1016/j.chembiol.2020.12.001> (cit. on pp. 6, 67).
- Harris, Charles R. et al. (Sept. 2020). « Array programming with NumPy. » In: *Nature* 585.7825, pp. 357–362. DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). URL: <https://doi.org/10.1038/s41586-020-2649-2> (cit. on p. 10).
- Harris, Zellig S. (1954). « Distributional Structure. » In: *WORD* 10.2-3, pp. 146–162. DOI: [10.1080/00437956.1954.11659520](https://doi.org/10.1080/00437956.1954.11659520). eprint: <https://doi.org/10.1080/00437956.1954.11659520>. URL: <https://doi.org/10.1080/00437956.1954.11659520> (cit. on p. 3).
- He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen (2021). « DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. » In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=XPZiaotutsD> (cit. on p. 4).
- Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom (2015). « Teaching Machines to Read and Comprehend. » In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc. URL: <https://proceedings.neurips.cc/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf> (cit. on p. 43).
- Hoffmann, Jordan et al. (2022). « Training Compute-Optimal Large Language Models. » In: *CoRR* abs/2203.15556 (cit. on p. 32).
- Hovy, Eduard H., Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel (2006). « OntoNotes: The 90% Solution. » In: *HLT-NAACL*. Ed. by Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson. The Association for Computational Linguistics. URL: <http://dblp.uni-trier.de/db/conf/naacl/naacl2006.html#HovyMPRW06> (cit. on pp. 4, 58).
- Huang, Haoyang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei (Dec. 2023). « Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. » In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 12365–12394. DOI: [10.18653/v1/2023](https://doi.org/10.18653/v1/2023).

- findings - emnlp.826. URL: <https://aclanthology.org/2023.findings-emnlp.826> (cit. on p. 2).
- Hunter, J. D. (2007). « Matplotlib: A 2D graphics environment. » In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: [10.1109/MCSE.2007.55](https://doi.org/10.1109/MCSE.2007.55) (cit. on p. 10).
- Iakovos Evdaimon, Hadi Abdine, Christos Xypolopoulos, Stamatis Outsios, Michalis Vazirgiannis, and Giorgos Stamou (2024). « GreekBART: The First Pretrained Greek Sequence-to-Sequence Model. » In: *Proceedings of the LREC-COLING 2024 conference*. Torino, Italy (cit. on p. 2).
- Ide, Nancy and Keith Suderman (May 2004). « The American National Corpus First Release. » In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA). URL: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/518.pdf> (cit. on p. 59).
- Inoue, Go, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash (Apr. 2021). « The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. » In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, pp. 92–104. URL: <https://aclanthology.org/2021.wanlp-1.10> (cit. on p. 38).
- Ji, Xu, João F Henriques, and Andrea Vedaldi (2019). « Invariant information clustering for unsupervised image classification and segmentation. » In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9865–9874 (cit. on p. 56).
- Jiang, Albert Q. et al. (2023). *Mistral 7B*. arXiv: [2310.06825 \[cs.CL\]](https://arxiv.org/abs/2310.06825) (cit. on p. 15).
- Jurgens, David and Ioannis Klapaftis (June 2013). « SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. » In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 290–299. URL: <https://aclanthology.org/S13-2049> (cit. on pp. 5, 20, 58).
- Kamal Eddine, Moussa, Antoine Tixier, and Michalis Vazirgiannis (Nov. 2021). « BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. » In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 9369–9390. DOI: [10.18653/v1/2021.emnlp-main.740](https://doi.org/10.18653/v1/2021.emnlp-main.740). URL: <https://aclanthology.org/2021.emnlp-main.740> (cit. on pp. 38, 40, 42, 61).
- Kilgarriff, Adam (1997). « I Don't Believe in Word Senses. » In: *Computers and the Humanities* 31.2, pp. 91–113. ISSN: 00104817. URL: <http://www.jstor.org/stable/30204773> (visited on 02/12/2024) (cit. on p. 4).
- Kim, Hwihan and Mamoru Komachi (Aug. 2021). « TMU NMT System with Japanese BART for the Patent task of WAT 2021. » In: *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. Online: Association for Compu-

- tational Linguistics, pp. 133–137. DOI: [10.18653/v1/2021.wat-1.13](https://doi.org/10.18653/v1/2021.wat-1.13). URL: <https://aclanthology.org/2021.wat-1.13> (cit. on p. 38).
- Kingma, Diederik P. and Jimmy Ba (2015). « Adam: A Method for Stochastic Optimization. » In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: <http://arxiv.org/abs/1412.6980> (cit. on pp. 41, 57, 76).
- Kipf, Thomas N and Max Welling (2017). « Semi-Supervised Classification with Graph Convolutional Networks. » In: *5th International Conference on Learning Representations* (cit. on pp. 19, 68, 69).
- Koehn, Philipp (2005). « Europarl: A Parallel Corpus for Statistical Machine Translation. » In: *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, pp. 79–86. URL: <https://aclanthology.org/2005.mtsummit-papers.11> (cit. on p. 40).
- Komninos, Alexandros and Suresh Manandhar (June 2016). « Dependency Based Embeddings for Sentence Classification Tasks. » In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 1490–1500. DOI: [10.18653/v1/N16-1175](https://doi.org/10.18653/v1/N16-1175). URL: <https://aclanthology.org/N16-1175> (cit. on p. 53).
- Kothapalli, Vignesh, Tom Tirer, and Joan Bruna (2023). « A Neural Collapse Perspective on Feature Evolution in Graph Neural Networks. » In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: <https://openreview.net/forum?id=sxao2udWXi> (cit. on p. 56).
- Koutsikakis, John, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos (2020). « GREEK-BERT: The Greeks Visiting Sesame Street. » In: *11th Hellenic Conference on Artificial Intelligence*. SETN 2020. Athens, Greece: Association for Computing Machinery, 110–117. ISBN: 9781450388788. DOI: [10.1145/3411408.3411440](https://doi.org/10.1145/3411408.3411440). URL: <https://doi.org/10.1145/3411408.3411440> (cit. on pp. xii, 39, 40, 44, 46).
- Krishna, Ranjay et al. (2017). « Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. » In: *International Journal of Computer Vision* 123.1, pp. 32–73. DOI: [10.1007/s11263-016-0981-7](https://doi.org/10.1007/s11263-016-0981-7). URL: <https://doi.org/10.1007/s11263-016-0981-7> (cit. on p. 5).
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2017). « ImageNet Classification with Deep Convolutional Neural Networks. » In: *Commun. ACM* 60.6, 84–90. ISSN: 0001-0782. DOI: [10.1145/3065386](https://doi.org/10.1145/3065386). URL: <https://doi.org/10.1145/3065386> (cit. on p. 37).
- Kudo, Taku and John Richardson (Nov. 2018). « SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. » In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71. DOI: [10.18653/v1/D18-2012](https://doi.org/10.18653/v1/D18-2012). URL: <https://aclanthology.org/D18-2012> (cit. on pp. 38, 41).

- Kulmanov, Maxat and Robert Hoehndorf (July 2019). « DeepGOPlus: improved protein function prediction from sequence. » In: *Bioinformatics* 36.2, pp. 422–429 (cit. on p. 67).
- Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres (2019). « Quantifying the Carbon Emissions of Machine Learning. » In: *arXiv preprint arXiv:1910.09700* (cit. on p. 77).
- Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin (Apr. 2012). « Word Sense Induction for Novel Sense Detection. » In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, pp. 591–601. URL: <https://aclanthology.org/E12-1060> (cit. on p. 53).
- Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang (2019). « BioBERT: a pre-trained biomedical language representation model for biomedical text mining. » In: *CoRR* abs/1901.08746. arXiv: 1901.08746. URL: <http://arxiv.org/abs/1901.08746> (cit. on p. 26).
- (2020). « BioBERT: a pre-trained biomedical language representation model for biomedical text mining. » In: *Bioinformatics* 36.4, pp. 1234–1240 (cit. on p. 77).
- Lewis, Armanda (July 2022). « Multimodal large language models for inclusive collaboration learning tasks. » In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, pp. 202–210. DOI: 10.18653/v1/2022.naacl-srw.26. URL: <https://aclanthology.org/2022.naacl-srw.26> (cit. on p. 2).
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (July 2020). « BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. » In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: <https://aclanthology.org/2020.acl-main.703> (cit. on pp. 2, 5, 7, 16, 37, 39, 40, 44, 46, 55, 68).
- Lhoest, Quentin et al. (Nov. 2021). « Datasets: A Community Library for Natural Language Processing. » In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 175–184. DOI: 10.18653/v1/2021.emnlp-demo.21. URL: <https://aclanthology.org/2021.emnlp-demo.21> (cit. on p. 9).
- Li, Weizhong and Adam Godzik (May 2006). « Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. » In: *Bioinformatics* 22.13, pp. 1658–1659. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl158 (cit. on pp. 9, 74).

- Lin, Chin-Yew (2004). « ROUGE: A package for automatic evaluation of summaries. » In: *Text summarization branches out*, pp. 74–81 (cit. on pp. 22, 47, 77).
- Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014). « Microsoft COCO: Common Objects in Context. » In: *CoRR abs/1405.0312*. arXiv: 1405.0312. URL: <http://arxiv.org/abs/1405.0312> (cit. on p. 5).
- Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. (2023). « Evolutionary-scale prediction of atomic-level protein structure with a language model. » In: *Science* 379.6637, pp. 1123–1130 (cit. on pp. 1, 6, 18, 68–70, 73).
- Ling, Chen et al. (2023). *Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey*. arXiv: 2305.18703 [cs.CL] (cit. on p. 2).
- Lioudakis, Michalis, Stamatis Outsios, and Michalis Vazirgiannis (Dec. 2020). « An Ensemble Method for Producing Word Representations focusing on the Greek Language. » In: *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*. Suzhou, China: Association for Computational Linguistics, pp. 99–107. URL: <https://aclanthology.org/2020.loresmt-1.13> (cit. on pp. 39, 45, 46).
- Liu, Shengchao et al. (2023). « A Text-guided Protein Design Framework. » In: *arXiv preprint arXiv:2302.04611* (cit. on p. 69).
- Liu, Xueliang (2017). « Deep recurrent neural network for protein function prediction from sequence. » In: *arXiv preprint arXiv:1701.08318* (cit. on p. 70).
- Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer (2020). « Multilingual Denoising Pre-training for Neural Machine Translation. » In: *Transactions of the Association for Computational Linguistics* 8, pp. 726–742. DOI: 10.1162/tacl_a_00343. URL: <https://aclanthology.org/2020.tacl-1.47> (cit. on pp. 38–40, 44).
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL] (cit. on pp. 4, 16, 51).
- Louviere, Jordan J., Terry N. Flynn, and A. A. J. Marley (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press. DOI: 10.1017/CB09781107337855 (cit. on p. 49).
- Luo, Yizhen, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie (2023). *BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine*. arXiv: 2308.09442 [cs.CE] (cit. on p. 1).
- Manandhar, Suresh and Ioannis Klapaftis (June 2009). « SemEval-2010 Task 14: Evaluation Setting for Word Sense Induction & Disambiguation Systems. » In: *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*. Boulder, Colorado: Association for Computa-

- tional Linguistics, pp. 117–122. URL: <https://aclanthology.org/W09-2419> (cit. on pp. 5, 20, 58).
- Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot (July 2020). « CamemBERT: a Tasty French Language Model. » In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7203–7219. DOI: [10.18653/v1/2020.acl-main.645](https://doi.org/10.18653/v1/2020.acl-main.645). URL: <https://aclanthology.org/2020.acl-main.645> (cit. on pp. 5, 25, 28, 38, 61).
- McCrae, John P., Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa (Jan. 2021). « The Global-WordNet Formats: Updates for 2020. » In: *Proceedings of the 11th Global Wordnet Conference*. University of South Africa (UNISA): Global Wordnet Association, pp. 91–99. URL: <https://aclanthology.org/2021.gwc-1.11> (cit. on p. 4).
- Micikevicius, Paulius et al. (2017). « Mixed Precision Training. » In: *CoRR abs/1710.03740*. arXiv: [1710.03740](https://arxiv.org/abs/1710.03740). URL: <http://arxiv.org/abs/1710.03740> (cit. on p. 40).
- Mikolov, Tomáš, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). « Efficient Estimation of Word Representations in Vector Space. » In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. URL: <http://arxiv.org/abs/1301.3781> (cit. on pp. 3, 53).
- Morris, Christopher, Gaurav Rattan, and Petra Mutzel (2020). « Weisfeiler and Leman go sparse: Towards scalable higher-order graph embeddings. » In: *Advances in Neural Information Processing Systems*. Vol. 34 (cit. on p. 70).
- Murphy, Ryan, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro (2019). « Relational Pooling for Graph Representations. » In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 4663–4673 (cit. on p. 70).
- Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (2018). « Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. » In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1797–1807. DOI: [10.18653/v1/D18-1206](https://doi.org/10.18653/v1/D18-1206). URL: <https://aclanthology.org/D18-1206> (cit. on pp. 7, 42, 47, 49).
- Nguyen, Xuan-Phi, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing (2023). *Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts*. arXiv: [2306.11372](https://arxiv.org/abs/2306.11372) [cs.CL] (cit. on p. 2).
- Nikolentzos, Giannis, George Dasoulas, and Michalis Vazirgiannis (2020). « k-hop graph neural networks. » In: *Neural Networks* 130, pp. 195–205 (cit. on p. 70).
- Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli (June 2019). « fairseq: A Fast, Extensible Toolkit for Sequence Modeling. » In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics,

- pp. 48–53. DOI: [10.18653/v1/N19-4009](https://doi.org/10.18653/v1/N19-4009). URL: <https://aclanthology.org/N19-4009> (cit. on pp. 9, 42).
- Outsios, Stamatis, Konstantinos Skianis, Polykarpos Meladianos, Christos Xypolopoulos, and Michalis Vazirgiannis (2018). « Word Embeddings from Large-Scale Greek Web content. » In: *arXiv preprint arXiv:1810.06694* (cit. on pp. 39, 40).
- Pantel, Patrick and Dekang Lin (2002). « Discovering Word Senses from Text. » In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. Edmonton, Alberta, Canada: Association for Computing Machinery, 613–619. ISBN: 158113567X. DOI: [10.1145/775047.775138](https://doi.org/10.1145/775047.775138). URL: <https://doi.org/10.1145/775047.775138> (cit. on p. 51).
- Papantoniou, Katerina and Yannis Tzitzikas (2020). « NLP for the Greek Language: A Brief Survey. » In: *11th Hellenic Conference on Artificial Intelligence*. SETN 2020. Athens, Greece: Association for Computing Machinery, 101–109. ISBN: 9781450388788. DOI: [10.1145/3411408.3411410](https://doi.org/10.1145/3411408.3411410). URL: <https://doi.org/10.1145/3411408.3411410> (cit. on p. 39).
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). « Bleu: a method for automatic evaluation of machine translation. » In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318 (cit. on pp. 21, 77).
- Paszke, Adam et al. (2019). « PyTorch: An Imperative Style, High-Performance Deep Learning Library. » In: *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., pp. 8024–8035. URL: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf> (cit. on p. 9).
- Pedregosa, F. et al. (2011). « Scikit-learn: Machine Learning in Python. » In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 9).
- Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). « GloVe: Global Vectors for Word Representation. » In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: [10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162). URL: <https://aclanthology.org/D14-1162> (cit. on p. 4).
- Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). « Deep Contextualized Word Representations. » In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: [10.18653/v1/N18-1202](https://doi.org/10.18653/v1/N18-1202). URL: <https://aclanthology.org/N18-1202> (cit. on pp. 4, 53).
- Qabel, Aymen, Sofiane Ennadir, Giannis Nikolentzos, Johannes F. Lutzeyer, Michail Chatzianastasis, Henrik Boström, and Michalis Vazirgiannis (2023). « Advancing Antibiotic Resistance Classification with Deep Learning Using Protein Sequence and Structure. » In: *bioRxiv*. DOI: [10.1101/2022.10.06.511103](https://doi.org/10.1101/2022.10.06.511103). eprint: <https://www.biorxiv.org/content/early/2023/04/06/>

- 2022.10.06.511103.full.pdf. URL: <https://www.biorxiv.org/content/early/2023/04/06/2022.10.06.511103> (cit. on p. 71).
- Radford, Alec and Karthik Narasimhan (2018). « Improving Language Understanding by Generative Pre-Training. » In: (cit. on p. 55).
- Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). « Improving language understanding by generative pre-training. » In: (cit. on pp. 15, 37, 69).
- Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). « Language Models are Unsupervised Multitask Learners. » In: (cit. on pp. 15, 37, 67, 69).
- Radford, Alec et al. (2021). « CLIP: Learning Transferable Visual Models From Natural Language Supervision. » In: *International Conference on Learning Representations* (cit. on pp. 6, 69).
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2019). « Exploring the limits of transfer learning with a unified text-to-text transformer. » In: *arXiv preprint arXiv:1910.10683* (cit. on pp. 68, 69).
- Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*. arXiv: 2204.06125 [cs.CV] (cit. on p. 6).
- Reiser, Patrick, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, et al. (2022). « Graph neural networks for materials science and chemistry. » In: *Communications Materials* 3.1, p. 93 (cit. on p. 68).
- Rives, Alexander, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. (2021). « Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. » In: *Proceedings of the National Academy of Sciences* 118.15. bioRxiv 10.1101/622803, e2016239118. DOI: 10.1073/pnas.2016239118 (cit. on pp. 1, 6, 18, 69).
- Rosenberg, Andrew and Julia Hirschberg (June 2007). « V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. » In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, pp. 410–420. URL: <https://aclanthology.org/D07-1043> (cit. on pp. 22, 58).
- Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). « DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. » In: *CoRR abs/1910.01108*. arXiv: 1910.01108. URL: <http://arxiv.org/abs/1910.01108> (cit. on p. 26).
- Scarselli, Franco, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini (2009). « The Graph Neural Network Model. » In: *IEEE Transactions on Neural Networks* 20.1, pp. 61–80 (cit. on pp. 19, 69).
- Schlichtkrull, Michael, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling (2018). « Modeling relational data with graph convolutional networks. » In: *The Semantic Web: 15th International Conference,*

- ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, *Proceedings* 15. Springer, pp. 593–607 (cit. on pp. 20, 68, 70, 72).
- Senior, Andrew W. et al. (2020). « Improved protein structure prediction using potentials from deep learning. » In: *Nature* 577.7792, pp. 706–710. DOI: 10.1038/s41586-019-1923-7. URL: <https://doi.org/10.1038/s41586-019-1923-7> (cit. on pp. 6, 9, 18).
- Sennrich, Rico, Barry Haddow, and Alexandra Birch (Aug. 2016). « Neural Machine Translation of Rare Words with Subword Units. » In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: <https://aclanthology.org/P16-1162> (cit. on p. 41).
- Seo, Younjoo, Andreas Loukas, and Nathanaël Perraudin (2019). « Discriminative structural graph classification. » In: *arXiv preprint arXiv: 1905.13422* (cit. on p. 70).
- Song, Linfeng, Zhiguo Wang, Haitao Mi, and Daniel Gildea (Aug. 2016). « Sense Embedding Learning for Word Sense Induction. » In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Berlin, Germany: Association for Computational Linguistics, pp. 85–90. DOI: 10.18653/v1/S16-2009. URL: <https://aclanthology.org/S16-2009> (cit. on p. 53).
- Strubell, Emma, Ananya Ganesh, and Andrew McCallum (July 2019). « Energy and Policy Considerations for Deep Learning in NLP. » In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: <https://aclanthology.org/P19-1355> (cit. on p. 50).
- Sulea, Octavia-Maria, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith (2017). « Exploring the Use of Text Classification in the Legal Domain. » In: *CoRR* abs/1710.09306. arXiv: 1710.09306. URL: <http://arxiv.org/abs/1710.09306> (cit. on pp. 25, 26).
- Tang, Yuqing, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan (2020). *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning*. DOI: 10.48550/ARXIV.2008.00401. URL: <https://arxiv.org/abs/2008.00401> (cit. on pp. 38, 39, 44).
- Touvron, Hugo et al. (2023a). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL] (cit. on pp. 1, 15).
- Touvron, Hugo et al. (2023b). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs.CL] (cit. on pp. 1, 15, 85).
- Tran, Nguyen Luong, Duong Minh Le, and Dat Quoc Nguyen (2021). « BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. » In: *CoRR* abs/2109.09701. arXiv: 2109.09701. URL: <https://arxiv.org/abs/2109.09701> (cit. on p. 38).
- Varadi, Mihaly, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. (2022). « AlphaFold Protein Structure Database: massively ex-

- panding the structural coverage of protein-sequence space with high-accuracy models. » In: *Nucleic acids research* 50.D1, pp. D439–D444 (cit. on p. 68).
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). « Attention is all you need. » In: *Advances in neural information processing systems* 30 (cit. on pp. 1, 4, 11, 37, 53, 67–69).
- Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio (2018). « Graph Attention Networks. » In: *ICLR*. arXiv: [1710.10903 \[stat.ML\]](https://arxiv.org/abs/1710.10903) (cit. on p. 20).
- Vinh, Nguyen, Julien Epps, and James Bailey (Jan. 2009). « Information theoretic measures for clusterings comparison: Is a correction for chance necessary? » In: p. 135. DOI: [10.1145/1553374.1553511](https://doi.org/10.1145/1553374.1553511) (cit. on p. 23).
- Véronis, Jean (July 2004). « HyperLex: Lexical cartography for information retrieval. » In: *Computer Speech & Language* 18, pp. 223–252. DOI: [10.1016/j.csl.2004.05.002](https://doi.org/10.1016/j.csl.2004.05.002) (cit. on p. 51).
- Wallace, Mike (2007). *Jawbone Java WordNet API*. URL: <https://sites.google.com/site/mfwallace/jawbone> (cit. on p. 51).
- Wang, Jing, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart, and Clement T. Yu (2015). « A Sense-Topic Model for Word Sense Induction with Unsupervised Data Enrichment. » In: *Transactions of the Association for Computational Linguistics* 3, pp. 59–71. DOI: [10.1162/tacl_a_00122](https://doi.org/10.1162/tacl_a_00122). URL: <https://aclanthology.org/Q15-1005> (cit. on p. 53).
- Wang, Limei, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji (2022). « Learning protein representations via complete 3d graph networks. » In: *arXiv preprint arXiv:2207.12600* (cit. on pp. 19, 70).
- Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave (May 2020). « CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. » English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4003–4012. ISBN: 979-10-95546-34-4. URL: <https://aclanthology.org/2020.lrec-1.494> (cit. on p. 38).
- Whittington, James C. R., Joseph Warren, and Timothy E. J. Behrens (2022). *Relating transformers to models and neural representations of the hippocampal formation*. arXiv: [2112.04035 \[cs.NE\]](https://arxiv.org/abs/2112.04035) (cit. on p. 1).
- Williams, Adina, Nikita Nangia, and Samuel Bowman (June 2018). « A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. » In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. DOI: [10.18653/v1/N18-1101](https://doi.org/10.18653/v1/N18-1101). URL: <https://aclanthology.org/N18-1101> (cit. on p. 45).
- Wolf, Thomas et al. (Oct. 2020). « Transformers: State-of-the-Art Natural Language Processing. » In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: [10.18653/v1/2020.emnlp-](https://doi.org/10.18653/v1/2020.emnlp-)

- demos.6. URL: <https://aclanthology.org/2020.emnlp-demos.6> (cit. on pp. 9, 76).
- Workshop, BigScience et al. (2023). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. arXiv: 2211.05100 [cs.CL] (cit. on p. 3).
- Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann (2023). *BloombergGPT: A Large Language Model for Finance*. arXiv: 2303.17564 [cs.LG] (cit. on p. 3).
- Xu, Keyulu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka (2019). « How Powerful are Graph Neural Networks? » In: *7th International Conference on Learning Representations* (cit. on p. 20).
- Xu, Minghao, Xinyu Yuan, Santiago Miret, and Jian Tang (2023). « ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts. » In: *arXiv preprint arXiv:2301.12040* (cit. on p. 69).
- Xu, P., X. Zhu, and D. A. Clifton (2023). « Multimodal Learning With Transformers: A Survey. » In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10, pp. 12113–12132. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2023.3275156 (cit. on p. 5).
- Xypolopoulos, Christos, Antoine Tixier, and Michalis Vazirgiannis (Apr. 2021). « Unsupervised Word Polysemy Quantification with Multiresolution Grids of Contextual Embeddings. » In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 3391–3401. DOI: 10.18653/v1/2021.eacl-main.297. URL: <https://aclanthology.org/2021.eacl-main.297> (cit. on pp. 8, 52, 54, 58).
- Yang, Jingfeng, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu (2023). *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*. arXiv: 2304.13712 [cs.CL] (cit. on p. 1).
- Yang, Yi, Mark Christopher Siy Uy, and Allen Huang (2020). « FinBERT: A Pretrained Language Model for Financial Communications. » In: *CoRR abs/2006.08097*. arXiv: 2006.08097. URL: <https://arxiv.org/abs/2006.08097> (cit. on p. 26).
- Yi-Lin Sung Jaemin Cho, Mohit Bansal (2022). « VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks. » In: *CVPR* (cit. on p. 5).
- Zhang, Chuxu, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla (2019a). « Heterogeneous graph neural network. » In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 793–803 (cit. on p. 70).
- Zhang, Susan et al. (2022a). *OPT: Open Pre-trained Transformer Language Models*. arXiv: 2205.01068 [cs.CL] (cit. on p. 3).
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2019b). « BERTScore: Evaluating Text Generation with BERT. » In: *CoRR abs/1904.09675*. arXiv: 1904.09675. URL: <http://arxiv.org/abs/1904.09675> (cit. on p. 47).
- Zhang*, Tianyi, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi (2020). « BERTScore: Evaluating Text Generation with BERT. » In: *In-*

- ternational Conference on Learning Representations*. URL: <https://openreview.net/forum?id=SkeHuCVFDr> (cit. on pp. 22, 55, 63).
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020). « BERTScore: Evaluating Text Generation with BERT. » In: *International Conference on Learning Representations* (cit. on p. 77).
- Zhang, Xiao-Meng, Li Liang, Lin Liu, and Ming-Jing Tang (2021). « Graph neural networks and their current applications in bioinformatics. » In: *Frontiers in genetics* 12, p. 690049 (cit. on p. 70).
- Zhang, Zuobai, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang (2022b). « Protein representation learning by geometric structure pretraining. » In: *arXiv preprint arXiv:2203.06125* (cit. on pp. 19, 70).
- Zitnik, Marinka, Monica Agrawal, and Jure Leskovec (2018). « Modeling polypharmacy side effects with graph convolutional networks. » In: *Bioinformatics* 34.13, pp. i457–i466 (cit. on p. 70).
- team, The pandas development (Feb. 2020). *pandas-dev/pandas: Pandas*. Version latest. DOI: [10.5281/zenodo.3509134](https://doi.org/10.5281/zenodo.3509134). URL: <https://doi.org/10.5281/zenodo.3509134> (cit. on p. 10).
- Hadi Abdine**, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis (2024). « Prot2Text: Multimodal Protein’s Function Generation with GNNs and Transformers. » In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence* (cit. on p. 6).
- Hadi Abdine**, Moussa Kamal Eddine, Davide Buscaldi, and Michalis Vazirgiannis (2023). « Word sense induction with agglomerative clustering and mutual information maximization. » In: *AI Open* 4, pp. 193–201. ISSN: 2666-6510. DOI: <https://doi.org/10.1016/j.aiopen.2023.12.001>. URL: <https://www.sciencedirect.com/science/article/pii/S2666651023000232> (cit. on p. 54).

APPENDIX

APPENDIX : JURIBERT

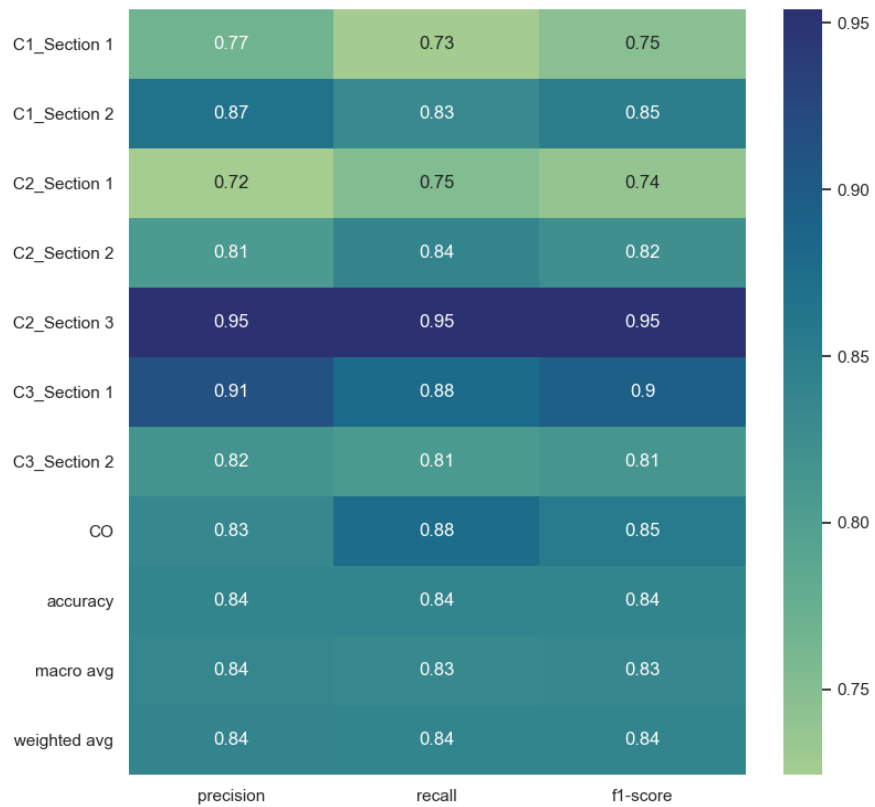


Figure a.1 – Accuracy, Precision, Recall and F1-Score of JuriBERT_{SMALL} on the chambers and sections classification task on the test dataset. The graph contains all eight classes.

Confusion matrix

Predicted \ Actual	C1_Section 1	C1_Section 2	C2_Section 1	C2_Section 2	C2_Section 3	C3_Section 1	C3_Section 2	CO	sum_lin
C1_Section 1	1717 8.70%	130 0.66%	50 0.25%	43 0.22%	10 0.05%	27 0.14%	66 0.33%	189 0.96%	2232 76.93% 23.07%
C1_Section 2	104 0.53%	2273 11.52%	72 0.36%	21 0.11%	8 0.04%	49 0.25%	27 0.14%	66 0.33%	2620 86.76% 13.24%
C2_Section 1	86 0.44%	69 0.35%	1391 7.05%	50 0.25%	28 0.14%	96 0.49%	74 0.37%	126 0.64%	1920 72.45% 27.55%
C2_Section 2	60 0.30%	41 0.21%	43 0.22%	1338 6.78%	39 0.20%	41 0.21%	58 0.29%	40 0.20%	1660 80.60% 19.40%
C2_Section 3	11 0.06%	9 0.05%	35 0.18%	24 0.12%	2012 10.19%	1 0.01%	0	17 0.09%	2109 95.40% 4.60%
C3_Section 1	27 0.14%	26 0.13%	55 0.28%	25 0.13%	3 0.02%	2308 11.69%	57 0.29%	27 0.14%	2528 91.30% 8.70%
C3_Section 2	63 0.32%	53 0.27%	34 0.17%	38 0.19%	0	71 0.36%	1583 8.02%	92 0.47%	1934 81.85% 18.15%
CO	297 1.50%	130 0.66%	168 0.85%	52 0.26%	9 0.05%	36 0.18%	93 0.47%	3950 20.01%	4735 83.42% 16.58%
sum_col	2365 72.60% 27.40%	2731 83.23% 16.77%	1848 75.27% 24.73%	1591 84.10% 15.90%	2109 95.40% 4.60%	2629 87.79% 12.21%	1958 80.85% 19.15%	4507 87.64% 12.36%	19738 83.96% 16.04%

Figure a.2 – Confusion Matrix of JuriBERT_{SMALL} on the chambers and sections classification task on the test dataset. The graph includes accuracy and error rate for each class.

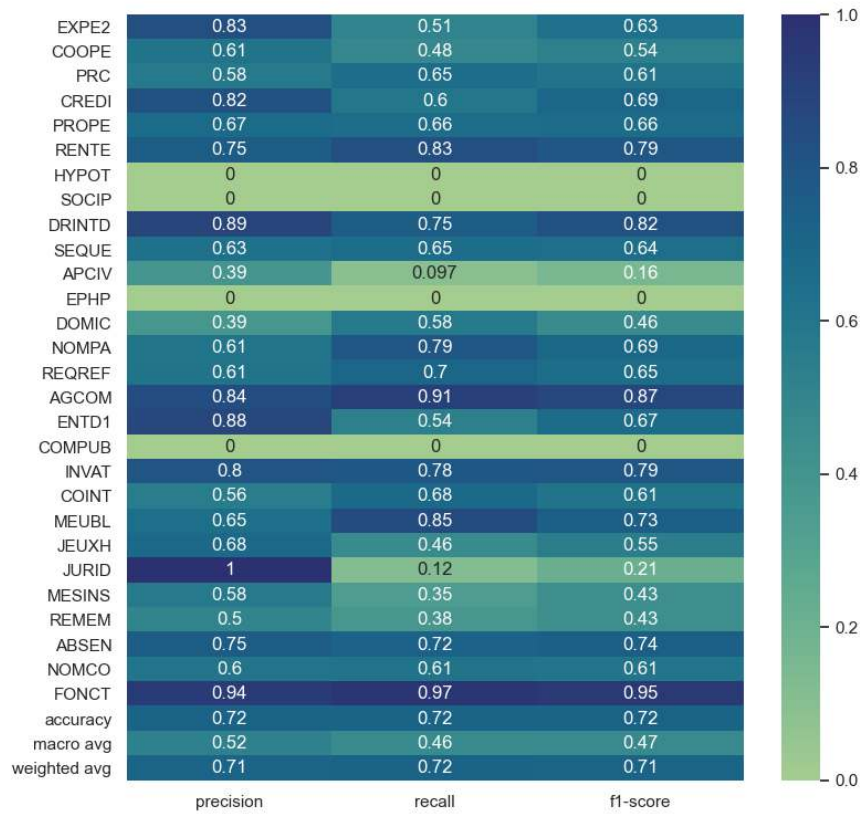


Figure a.3 – Sample of Accuracy, Precision, Recall and F1-Score of JuriBERT_{SMALL} on the *matières* classification task on the test dataset. The graph contains 28 classes and the overall accuracy of all 148 classes.

APPENDIX : GREEKBART EXAMPLES

B.1 GREEKSUM ABSTRACT

In this Appendix section, we present the reference and model summaries of 5 randomly selected documents from the test set of the GreekSUM Abstract.

Document	«Ο κύβος ερρίφθη. Ο Άρμιν Λάσετ θα είναι ο υποψήφιος Καγκελάριος για την Χριστιανική Ένωση», δήλωσε πριν από λίγο ο Αρχηγός της Χριστιανοκοινωνικής Ένωσης (CSU) και Πρωθυπουργός της Βαυαρίας Μάρκους Ζέντερ, αναγνωρίζοντας το αποτέλεσμα της ψηφοφορίας του προεδρείου του Χριστιανοδημοκρατικού Κόμματος (CDU), το οποίο σε ποσοστό 77,5% τάχθηκε υπέρ της υποψηφιότητας του κ. Λάσετ. Πριν από την συνεδρίαση του προεδρείου του CDU, ο κ. Ζέντερ είχε δηλώσει ότι παραχωρεί στο CDU το προβάδισμα στην επιλογή του υποψήφιου Καγκελάριου της Χριστιανικής Ένωσης (CDU/CSU) και σήμερα επανέλαβε ότι δέχεται το αποτέλεσμα «χωρίς μνησικακία» και ότι τάσσεται υπέρ της ενότητας της Χριστιανικής Ένωσης.
ABSTRACT	<p>Gold</p> <p>Ο Άρμιν Λάσετ θα είναι ο υποψήφιος των CDU και CSU για την καγκελαρία της Γερμανίας στις εκλογές του Σεπτεμβρίου.</p> <p>BART-random</p> <p>Ο Άρμιν Λάσετ θα είναι ο υποψήφιος πρωθυπουργός της Χριστιανικής Ένωσης, μετά από σχετική συνεδρίαση.</p> <p>mBART₂₅</p> <p>Ο πρωθυπουργός της Βαυαρίας δέχθηκε το αποτέλεσμα της ψηφοφορίας του προεδρείου του CDU, το οποίο σε ποσοστό 77,5% τάχθηκε υπέρ της υποψηφιότητας του Άρμιν Λάσετ.</p> <p>mBART₅₀</p> <p>Σε ποσοστό 77,5% τάχθηκε υπέρ της υποψηφιότητας του Άρμιν Λάσετ στο προεδρείο του CDU, ο Πρωθυπουργός της Βαυαρίας Μάρκους Ζέντερ.</p> <p>GreekBART</p> <p>Υπέρ του Άρμιν Λάσετ τάσσεται ο Μάρκους Ζέντερ, αναγνωρίζοντας το αποτέλεσμα της ψηφοφορίας του προεδρείου του CDU.</p>

Table b.1 – Example 1-GreekSUM Abstract

Document	Κλειστή είναι η λεωφόρος Βασ. Κωνσταντίνου στο ύψος του Παναθηναϊκού Σταδίου, με αποτέλεσμα να έχει δημιουργηθεί κυκλοφοριακό πρόβλημα, καθώς έχει χυθεί μεγάλη ποσότητα λαδιού από φορτηγό, στην συμβολή με την λεωφόρο Βασ. Όλγας και είναι επικίνδυνη η διέλευση των οχημάτων. Η Τροχαία έχει διακόψει την κυκλοφορία στο κατερχόμενο ρεύμα στο ύψος της οδού Ριζάρη και στο ανερχόμενο από την αρχή της Αρδητού και κάνει εκτροπή, αλλά έχει δημιουργηθεί μποτιλιάρισμα. Στο σημείο που έχουν χυθεί τα λάδια βρίσκονται συνεργεία του Δήμου, που ρίχνουν πριονίδι και άλλα υλικά για να αντιμετωπίσουν την ολοσθιρότητα του οδοστρώματος και να αποκατασταθεί η κυκλοφορία.	
ABSTRACT	Gold	Η Τροχαία έχει διακόψει την κυκλοφορία στο κατερχόμενο ρεύμα στο ύψος της οδού Ριζάρη και στο ανερχόμενο από την αρχή της Αρδητού και κάνει εκτροπή - Μποτιλιάρισμα στο σημείο.
	BART-random	Η Τροχαία, που έχει δημιουργηθεί στο ύψος του Παναθηναϊκού, έχει διακοπεί την κυκλοφορία των οχημάτων στην λεωφόρο Βασ.Α.
	mBART25	Κλειστή είναι η λεωφόρος Βασ. Κωνσταντίνου στο ύψος του Παναθηναϊκού Σταδίου, με αποτέλεσμα να έχει δημιουργηθεί κυκλοφοριακό πρόβλημα.
	mBART50	Κυκλοφοριακό πρόβλημα στο ύψος του Παναθηναϊκού Σταδίου, καθώς έχει χυθεί μεγάλη ποσότητα λαδιού από φορτηγό σε λεωφόρο Βασ. Κωνσταντίνου.
	GreekBART	Κυκλοφοριακό πρόβλημα έχει δημιουργηθεί στην Λεωφόρο Βασ. Κωνσταντίνου στο ύψος του Παναθηναϊκού Σταδίου, με αποτέλεσμα να έχει δημιουργηθεί μποτιλιάρισμα.

Table b.2 – Example 2-GreekSUM Abstract

Document	<p>Η Καγκελάριος Άνγκελα Μέρκελ δεν θα παραστεί στην επίσημη δεξίωση που θα παραθέσει την Παρασκευή ο Ομοσπονδιακός Πρόεδρος Φρανκ-Βάλτερ Στάϊνμάιερ προς τιμήν του Προέδρου της Τουρκίας Ρετζέπ Ταγίπ Ερντογάν, σύμφωνα με κυβερνητικές πηγές τις οποίες επικαλείται το περιοδικό «Der Spiegel». Η δεξίωση αλλά και οι στρατιωτικές τιμές με τις οποίες θα υποδεχθεί τον προσκεκλημένο του ο Γερμανός Πρόεδρος προκαλούν σοβαρές αντιδράσεις στον πολιτικό κόσμο της χώρας. Η Μέρκελ είναι πάντα προσκεκλημένη του Ομοσπονδιακού Προέδρου σε δεξιώσεις ή επίσημα δείπνα που παρατίθενται προς τιμήν υψηλών προσκεκλημένων. Η ίδια ωστόσο συνηθίζει να παρευρίσκεται μόνο σε εξαιρετικές περιπτώσεις. Η τελευταία φορά που παρέστη σε κάτι ανάλογο ήταν το επίσημο δείπνο που είχε παρατεθεί το 2015 προς τιμήν της Βασίλισσας Ελισάβετ, ενώ την προηγούμενη χρονιά είχε παρευρεθεί στο δείπνο με τον Εμίρη του Κατάρ. Αντιθέτως, δεν είχε παρευρεθεί στην δεξίωση προς τιμήν του Κινέζου Προέδρου Σι Τζινπίνγκ το 2017. Η Καγκελάριος όμως δεν θα είναι η μόνη που θα απορρίψει την πρόσκληση του Στάϊνμάιερ. Ο Πρόεδρος των Φιλελευθέρων (FDP) Κρίστιαν Λίντντερ ανακοίνωσε ότι δεν σκοπεύει να παραστεί, καθώς δεν επιθυμεί «να συμμετάσχει στην προπαγάνδα του Ερντογάν». Την ίδια στάση θα τηρήσει και η εκπρόσωπος του κόμματος για την εξωτερική πολιτική, Μπιτζάν Ντζιρ-Σαράι, ενώ σύσσωμη η ηγετική ομάδα των Πρασίνων, οι συμπρόεδροι Αναλένα Μπέρμποκ και Ρόμπερτ Χάμπεκ και οι επικεφαλής της Κοινοβουλευτικής Ομάδας Κάτριν Γκέρνινγκ-Έκαρτ και Άντον Χοφράιτερ, δήλωσαν ότι θα απέχουν από την δεξίωση. Το ίδιο ισχύει και για τους επικεφαλής της Εναλλακτικής για την Γερμανία (AfD) Άλεξάντερ Γκάουλαντ και Αλίσ Βαϊντέλ και για την επικεφαλής της Κ. Ο. της Αριστεράς Σεβίμ Νταγκντελέν. Αντιθέτως, την πρόθεσή του να παραστεί στην δεξίωση στο Προεδρικό Ανάκτορο Bellevue εξέφρασε ο πρώην Πρόεδρος των Πρασίνων Τζεμ Έζντεμιρ, διευκρινίζοντας ταυτόχρονα ότι ο Τούρκος Πρόεδρος «δεν είναι κανονικός Πρόεδρος και δεν αξίζει» να παρατεθεί δεξίωση προς τιμήν του. Με την παρουσία του, δήλωσε ο κ. Έζντεμιρ στην «Tagesspiegel», ελπίζει να στείλει ένα μήνυμα τόσο προς την Τουρκία όσο και προς την τουρκογερμανική κοινότητα: «Η αντιπολίτευση στην Γερμανία είναι μέρος της πολιτικής αυτής της χώρας, είμαστε ένα σταθερό και απαραίτητο συστατικό στοιχείο της δημοκρατίας μας. Ο κ. Ερντογάν θα πρέπει να με ανεχθεί».</p>
<p>Gold</p> <p>ABSTRACT</p> <p>BART-random</p> <p>mBART₂₅</p> <p>mBART₅₀</p> <p>GreekBART</p>	<p>Η καγκελάριος είναι πάντα προσκεκλημένη του ομοσπονδιακού προέδρου σε δεξιώσεις ή δείπνα προς τιμήν υψηλών προσκεκλημένων, ωστόσο δίνει το παρών μόνο σε εξαιρετικές περιπτώσεις.</p> <p>Δεν θα παραστεί στην επίσημη δεξίωση που θα παραθέσει την Τουρκία προς τιμήν του Ρετζέπ Ταγίπ Ερντογάν ο εκπρόσωπος της Γερμανίας Άνγκελα Μέρκελ.</p> <p>Αντιδράσεις από τον πολιτικό κόσμο της χώρας προκαλούν η δεξίωση που θα παραθέσει ο Φρανκ-Βάλτερ Στάϊνμάιερ προς τιμήν του Προέδρου της Τουρκίας - Δεν θα είναι η μόνη που θα απορρίψει την πρόσκληση του Στάϊνμάιερ.</p> <p>Η Μέρκελ είναι πάντα προσκεκλημένη του Ομοσπονδιακού Προέδρου σε δεξιώσεις ή επίσημα δείπνα που παρατίθενται προς τιμήν υψηλών προσκεκλημένων. Η ίδια ωστόσο συνηθίζει να παρευρίσκεται μόνο σε εξαιρετικές περιπτώσεις.</p> <p>Από τον πολιτικό κόσμο της Γερμανίας. Η Άνγκελα Μέρκελ δεν θα παραστεί στην επίσημη δεξίωση προς τιμήν του Γερμανού Προέδρου Φρανκ-Βάλτερ Στάϊνμάιερ.</p>

Table b.3 – Example 3-GreekSUM Abstract

Document	<p>Από το 2011 και μετά αρκετοί εκατοντάδες άνθρωποι έχουν πεθάνει στην προσπάθειά τους να βγάλουν την τέλεια selfie. Οι περισσότεροι θάνατοι έχουν λάβει χώρα στην Ινδία. Ακολουθεί η Ρωσία, οι Ηνωμένες Πολιτείες και ύστερα το Πακιστάν με τους νεκρούς συνολικά να φτάνουν τους 259. Βέβαια υπάρχουν κάποια σημεία, τα οποία σύμφωνα με έρευνες, παρουσιάζουν μεγαλύτερη επικινδυνότητα, όπως το νερό και οι ψηλές κυλιόμενες σχάλες. Οι πιο «συνηθισμένες» αιτίες θανάτου από selfie συμπεριλαμβάνουν τον πνιγμό, την πτώση, τη σύγκρουση με κινούμενο όχημα και τις φωτιές. Όσον αφορά τα στατιστικά στοιχεία τα 3/4 των θυμάτων είναι άνδρες και κάτω από την ηλικία των 30. Αν και οι γυναίκες βγάζουν περισσότερες selfie σύμφωνα με τις μελέτες, οι άνδρες είναι πιο επιρρεπείς στον κίνδυνο. Ακόμα, οι τουρίστες είναι αυτοί που πλήττονται πιο συχνά στην προσπάθεια να βγάλουν μια φωτογραφία που θα εντυπωσιάσει τους ακολούθους τους. Οι αρχές ψάχνουν τρόπους προκειμένου να αποτρέψουν τους θανάτους. Για παράδειγμα η ρωσική αστυνομία μοίρασε φυλλάδια, τα οποία εμπεριείχαν προειδοποιήσεις σχετικά με τους κινδύνους που «καραδοκούν» πίσω από μια selfie. Στις Ηνωμένες Πολιτείες, τα εθνικά πάρκα έχουν εκδώσει οδηγούς για το πώς να βγάζεις «ασφαλείς» selfies, ενώ στην Ινδία υπάρχουν επίσημα σχεδιασμένες πινακίδες που προειδοποιούν για υψηλού κινδύνου περιοχές ή αλλιώς “No selfie zones”. Αν και η εμμονή με τις selfie δεν φαίνεται να περνάει οι αρχές κάνουν ότι μπορούν για να περιορίσουν την επικινδυνότητα και τους θανάτους.</p>
<p>ABSTRACT</p> <p>Gold</p> <p>BART-random</p> <p>mBART25</p> <p>mBART50</p> <p>GreekBART</p>	<p>Οι πιο «συνηθισμένες» αιτίες θανάτου από selfie συμπεριλαμβάνουν πνιγμό, πτώση, και τη σύγκρουση με κινούμενο όχημα - Άνδρες κάτω των 30 τα περισσότερα θύματα.</p> <p>Οι Ηνωμένες Πολιτείες, Ινδία, Αν. και Πακιστάν και Αν. Ινδία αναζητούν αναζητούν στοιχεία για να βγάλουν την τέλεια selfie τους στην προσπάθειά τους.</p> <p>Η Ινδία μετράει τους 259 θανάτους από selfie, τα οποία συμπεριλαμβάνουν τον πνιγμό, την πτώση, τη σύγκρουση με κινούμενο όχημα και τις φωτιές. Οι αρχές ψάχνουν τρόπους προκειμένου να αποτρέψουν τους θανάτους.</p> <p>Στην Ινδία, τα εθνικά πάρκα έχουν εκδώσει οδηγούς για το πώς να βγάζεις «ασφαλείς» selfies, ενώ στην Ινδία υπάρχουν επίσημα σχεδιασμένες πινακίδες που προειδοποιούν για υψηλού κινδύνου περιοχές.</p> <p>Πολλοί άνθρωποι έχουν πεθάνει στην προσπάθειά τους να βγάλουν μια selfie, με τις «συνηθισμένες» αιτίες να συμπεριλαμβάνουν τον πνιγμό, την πτώση, τη σύγκρουση με κινούμενο όχημα και τις φωτιές.</p>

Table b.4 – Example 4-GreekSUM Abstract

Document	<p>Στην απώλεια του Μίκη Θεοδωράκη αναφέρθηκε ο πρωθυπουργός Κυριάκος Μητσοτάκης στην έναρξη της συνεδρίασης του Υπουργικού Συμβουλίου, κηρύσσοντας τριήμερο εθνικό πένθος. Ο πρωθυπουργός ειδικότερα δήλωσε: “Τη σημερινή μας συνεδρίαση σιαάζει δυστυχώς μία πολύ θλιβερή είδηση: Ο Μίκης Θεοδωράκης περνά πια στην αιωνιότητα. Η φωνή του σίγησε και μαζί του σίγησε και ολόκληρος ο Έλληνας. Όπως είχε γραφτεί και για τον Παλαμά, «όλοι είχαμε ξεχάσει πως είναι θνητός». Όμως, μας αφήνει παρακαταθήκη τα τραγούδια του, την πολιτική του δράση, αλλά και την εθνική του προσφορά σε κρίσιμες στιγμές. Η Ρωμιοσύνη σήμερα κλαίει. Και γι’ αυτό και με απόφαση της κυβέρνησης από σήμερα κηρύσσεται τριήμερο εθνικό πένθος. Όπως ξέρετε, είχα την τιμή να τον γνωρίζω για πολλά χρόνια και σχετικά πρόσφατα μάλιστα τον είχα επισκεφτεί. Οι συμβουλές του ήταν πάντα πολύτιμες για μένα, κυρίως αυτές που αφορούσαν στην ενότητα του λαού μας και στην υπέρβαση των διαχωριστικών γραμμών. Πιστεύω πως η καλύτερη τιμή προς αυτόν τον παγκόσμιο Έλληνα θα είναι εμείς, με το καθημερινό μας έργο, να κάνουμε πράξη αυτό ακριβώς το μήνυμά του. Ο Μίκης είναι η Ιστορία μας και πρέπει να τη συνεχίσουμε όπως θα ήθελε και εκείνος.” Πέθανε ο Μίκης Θεοδωράκης - Ορφάνεψε η Ρωμιοσύνη Έλενα Ακρίτα - Ο Ήλιος (που κρύφτηκε) και ο Χρόνος (που χάθηκε), Μίκη Μίκης Θεοδωράκης: Τα 5 τραγούδια του σπουδαίου μουσικού που «μιλούν» στην ψυχή της Ελλάδας Ο πολιτικός Μίκης Θεοδωράκης: Πάντα στο πλευρό των απλών ανθρώπων.</p>
<p style="writing-mode: vertical-rl; transform: rotate(180deg);">ABSTRACT</p> <p>Gold</p> <p>BART-random</p> <p>mBART25</p> <p>mBART50</p> <p>GreekBART</p>	<p>Η Ρωμιοσύνη σήμερα κλαίει δήλωσε ο πρωθυπουργός στην έναρξη της συνεδρίασης του υπουργικού συμβουλίου αναφερόμενος στο θάνατο του Μίκη Θεοδωράκη.</p> <p>Ο πρωθυπουργός κατά την έναρξη της συνεδρίασης του Υπουργικού Συμβουλίου κηρύσσοντας την απώλεια του Μίκη Θεοδωράκη.</p> <p>Ο πρωθυπουργός Κυριάκος Μητσοτάκης απο το υπουργικό συμβούλιο για τον θάνατο του Μίκη Θεοδωράκη.</p> <p>Τριήμερο εθνικό πένθος κηρύχθηκε στη συνεδρίαση του υπουργικού συμβουλίου, με τον πρωθυπουργό να σημειώνει ότι ο Μίκης Θεοδωράκης περνά πια στην αιωνιότητα.</p> <p>Το δικό του μήνυμα για την απώλεια του Μίκη Θεοδωράκη έστειλε ο πρωθυπουργός Κυριάκος Μητσοτάκης κατά τη συνεδρίαση του Υπουργικού Συμβουλίου.</p>

Table b.5 – Example 5-GreekSUM Abstract

B.2 GREEKSUM TITLE

In the second section of the appendices, we present the reference and model titles of 5 randomly selected documents from the test set of the GreekSUM Title.

Document	Ένας 33χρονος έχασε τη ζωή του, ύστερα από σύγκρουση δύο αυτοκινήτων, έξω από τη Θεσσαλονίκη. Όπως έγινε γνωστό, το θανατηφόρο τροχαίο συνέβη στις 2.15 μετά τα μεσάνυχτα σε παράδρομο της Εγνατίας Οδού, στο ύψος του Ωραιοκάστρου. Σύμφωνα με την Αστυνομία, ο 33χρονος, οδηγός του ενός οχήματος, διακομίστηκε στο νοσοκομείο Παπαγεωργίου, όπου όμως λίγη αργότερα υπέκυψε στα τραύματά του, ενώ η οδηγός του άλλου οχήματος υπέστη ελαφρά τραύματα. Οι ακριβείς συνθήκες υπό τις οποίες προκλήθηκε η σύγκρουση ερευνώνται από το αρμόδιο τμήμα τροχαίας.
Gold	Τροχαίο δυστύχημα στη Θεσσαλονίκη με έναν νεκρό
BART-random	Τροχαίο έξω από τη Θεσσαλονίκη - Δύο τραυματίες
mBART25	Θεσσαλονίκη: Νεκρός 33χρονος ύστερα από σύγκρουση δύο αυτοκινήτων
mBART50	Θεσσαλονίκη: Νεκρός 33χρονος ύστερα από σύγκρουση δύο αυτοκινήτων
GreekBART	Τροχαίο στη Θεσσαλονίκη: Νεκρός 33χρονος σε παράδρομο

Table b.6 – Example 1-GreekSUM Title

Document	<p>Όλες οι χώρες της Ευρωπαϊκής Ένωσης συμφωνούν ότι δεν θα πληρώσουν τη Ρωσία απευθείας σε ρούβλια για τις εισαγωγές ρωσικού φυσικού αερίου, δήλωσαν υψηλόβαθμοι ευρωπαϊκοί αξιωματούχοι, σημειώνοντας ότι οι επόμενες πληρωμές είναι προγραμματισμένες για τις 20 Μαΐου. «Αυτό που γνωρίζουμε, και υπάρχει συναίνεση επ'αυτού μεταξύ όλων των κρατών μελών, είναι ότι κανείς δεν είναι πρόθυμος να πληρώσει σε ρούβλια», δήλωσε ο ένας αξιωματούχος κατά την διάρκεια ενημέρωσης των δημοσιογράφων και προσθέτοντας ότι η Ευρωπαϊκή Επιτροπή δεν γνωρίζει πόσοι αγοραστές έχουν ανοίξει λογαριασμούς για πληρωμές προμήθειας φυσικού αερίου μέσω της Gazprombank. Στο μεταξύ, ανώτερος αξιωματούχος της Ευρωπαϊκής Ένωσης δήλωσε πως και μόνο το άνοιγμα τραπεζικού λογαριασμού σε ρούβλια στην Gazprombank ενδέχεται να αποτελεί παραβίαση των κυρώσεων που έχει επιβάλει η ΕΕ σε βάρος της Ρωσίας, όμως η ΕΕ δεν έχει ένδειξη πως κάποια εταιρεία φυσικού αερίου της ΕΕ έχει κάνει κάτι τέτοιο. Ο αξιωματούχος δήλωσε πως «εκ πρώτης όψεως» το άνοιγμα τραπεζικών λογαριασμών σε ρούβλια από εισαγωγείς φυσικού αερίου φαίνεται ότι παραβιάζει τις κυρώσεις. Ο αξιωματούχος πρόσθεσε πως η Ευρωπαϊκή Επιτροπή δεν έχει κάποια επίσημη ένδειξη ότι εταιρείες της ΕΕ έχουν δημιουργήσει στην Gazprombank λογαριασμούς σε ρούβλια για την πληρωμή του φυσικού αερίου. Επίσης διευκρίνισε πως η Πολωνία και η Βουλγαρία χρησιμοποίησαν τις υφιστάμενες μεθόδους πληρωμής για το ρωσικό αέριο, πριν η Μόσχα αναστείλει χθες, Τετάρτη, τις προμήθειες των χωρών αυτών με αέριο, και πως δεν χρησιμοποίησαν τον μηχανισμό που προτείνει η Μόσχα για να πληρώσουν σε ρούβλια. «Σύμφωνα με τις πληροφορίες μας, αμφότερες οι χώρες επέμειναν στην αρχική μορφή πληρωμής», δήλωσε ο αξιωματούχος σε δημοσιογράφους. Ωστόσο δύο πηγές είπαν σήμερα στο Ρόιτερς ότι λίγες ευρωπαϊκές εταιρείες έχουν αρχίσει να πληρώνουν σε ρούβλια τη Ρωσία για το φυσικό αέριο, αν και μεγάλοι πελάτες της δεν το έχουν κάνει ακόμη. «Μερικές εμπορικές εταιρείες, ίσως περισσότερες από πέντε, έχουν αρχίσει τις πληρωμές», είπε μία πηγή, ζητώντας να μην κατονομαστεί, επειδή δεν είχε εξουσιοδοτηθεί να μιλήσει στα μέσα ενημέρωσης.</p>
<p>Gold</p> <p>BART-random</p> <p>mBART25</p> <p>mBART50</p> <p>GreekBART</p>	<p>Φυσικό αέριο: Όλες οι χώρες της ΕΕ συμφωνούν ότι δεν θα πληρώσουν τη Ρωσία σε ρούβλια</p> <p>E.E.: «Δεν θα πληρώσουν» οι χώρες της ΕΕ για το φυσικό αέριο σε ρούβλια</p> <p>EE: Οι χώρες δεν πληρώνουν σε ρούβλια τη Ρωσία για το φυσικό αέριο</p> <p>EE: Οι χώρες δεν πληρώνουν σε ρούβλια τη Ρωσία για το φυσικό αέριο</p> <p>EE: Δεν θα πληρώσουμε τη Ρωσία σε ρούβλια για το φυσικό αέριο</p>

Table b.7 – Example 2-GreekSUM Title

Document	<p>Στις ημέρες του Πάσχα έχει προσαρμοστεί το πρόγραμμα λειτουργίας λεωφορείων, τρόλεϊ, ηλεκτρικού και μετρό. Ειδικότερα, τα λεωφορεία και τα τρόλεϊ σήμερα, Μεγάλη Παρασκευή, θα κινούνται με πρόγραμμα Σαββάτου. Οι συρμοί στο μετρό θα διέρχονται από τους σταθμούς ανά 7 λεπτά από τις 09.00 έως τις 17.00 και ανά 10 λεπτά τις υπόλοιπες ώρες. Υπενθυμίζεται πως δεν θα ισχύσει η δίωρη παράταση λειτουργίας που εφαρμόζεται τις Παρασκευές. Στον ηλεκτρικό οι συρμοί θα διέρχονται από τους σταθμούς ανά 10,5 λεπτά. Τα λεωφορεία και τα τρόλεϊ θα κινηθούν με πρόγραμμα Κυριακής, ενώ θα αποσυρθούν νωρίτερα, ώστε να βρίσκονται στα αμαξοστάσια στις 23.00. Τα λεωφορεία θα κινηθούν με πρόγραμμα Κυριακής και τα τρόλεϊ με ειδικό πρόγραμμα Κυριακής. Τόσο στα δρομολόγια των λεωφορείων όσο και σ' αυτά των τρόλεϊ θα εφαρμοστεί ειδικό πρόγραμμα Σαββάτου. Ακινήτοποιημένοι θα μείνουν την Τετάρτη 1η Μαΐου οι συρμοί του ηλεκτρικού (πρώην ΗΣΑΠ), τα λεωφορεία, τα τρόλεϊ, αλλά και ο σιδηρόδρομος, λόγω 24ωρης απεργίας των εργαζομένων, που θα συμμετάσχουν στις απεργιακές συγκεντρώσεις για την Πρωτομαγιά. Όπως αναφέρουν σε ανακοίνωσή τους οι εργαζόμενοι στον πρώην ΗΣΑΠ, «είναι μέρα αγώνα, τιμής και μνήμης. Θυμόμαστε και τιμάμε τους πρωτοπόρους αγωνιστές και τα θύματα των εργατικών αγώνων για βελτίωση των συνθηκών δουλειάς για αξιοπρεπείς αμοιβές και την κατοχύρωση των δικαιωμάτων μας. Ανασυγκροτούμαστε, θέτουμε τους στόχους μας και προχωράμε σε νέους αγώνες. Διεκδικούμε και παλεύουμε για την αναπλήρωση απωλειών από τις μνημονιακές πολιτικές λιτότητας, για πραγματικές αυξήσεις στους μισθούς και στις κοινωνικές παροχές». Και προσθέτουν «υπερασπιζόμαστε τον δημόσιο χαρακτήρα των συγκοινωνιών. Διεκδικούμε την υπογραφή νέας Συλλογικής Σύμβασης Εργασίας. Αγωνιζόμαστε για ασφαλείς, φθηνές συγκοινωνίες. Με αγώνες κατακτάμε τα δικαιώματά μας».</p>
<p>Gold</p> <p>BART-random</p> <p>mBART25</p> <p>mBART50</p> <p>GreekBART</p>	<p>Πάσχα 2019: Πώς θα κινηθούν λεωφορεία, τρόλεϊ, ηλεκτρικός και μετρό</p> <p>Μέσα Μαζικής Μεταφοράς: Πώς θα κινηθούν σήμερα τα Μέσα Μεταφοράς</p> <p>Μέσα Πάσχα: Πώς θα κινηθούν σήμερα λεωφορεία, τρόλεϊ, ηλεκτρικό και μετρό</p> <p>Μέσα Πάσχα: Πώς θα κινηθούν σήμερα λεωφορεία, τρόλεϊ, ηλεκτρικό και μετρό</p> <p>Πάσχα: Πώς θα κινηθούν σήμερα λεωφορεία, τρόλεϊ, ηλεκτρικού και μετρό</p>

Table b.8 – Example 3-GreekSUM Title

Document	<p>Συνάντηση με οικονομικούς παράγοντες από το Σίτι του Λονδίνου έχει αυτή την ώρα ο Αλέξης Τσίπρας στο κέντρο της βρετανικής πρωτεύουσας. Τον Έλληνα πρωθυπουργό υποδέχθηκε ο αντιπρόεδρος της Επιτροπής Πολιτικής του Σίτι, Τομ Σλέι (Tom Sleigh). Επισημαίνεται ότι η Επιτροπή υπέχει θέση Διοίκησης του Σίτι του Λονδίνου. Από την αίθουσα της «Παλιάς Βιβλιοθήκης», ο πρωθυπουργός θα απευθυνθεί σε έναν κύκλο περισσότερων από εκατό σημαίνοντων στελεχών της επενδυτικής/χρηματοπιστωτικής κοινότητας του Σίτι και, σύμφωνα με πληροφορίες, στη συνέχεια θα ακολουθήσει συνάντηση σε πιο στενό κύκλο συμμετεχόντων. Στον απόηχο της απόφασης του Eurogroup για την ελάφρυνση του χρέους, οι επαφές του Αλέξη Τσίπρα με σημαντικούς εκπροσώπους της επενδυτικής/χρηματοπιστωτικής της κοινότητας του οικονομικού κέντρου της Ευρώπης, σηματοδοτούν ένα ευκρινές διεθνές μήνυμα για τις προοπτικές της ελληνικής οικονομίας και της «επόμενης μέρας», στην περίοδο μετά την ολοκλήρωση των μνημονίων. Όπως ανέφερε κυβερνητικός αξιωματούχος, οι σημερινές συναντήσεις είναι ένας σημαντικός σταθμός σε μια «αλυσίδα» επαφών και συνομιλιών που θα συνεχιστούν στο αμέσως επόμενο διάστημα των καλοκαιρινών μηνών και το φθινόπωρο. Ενδεικτική της ευνοϊκής συγκυρίας για την ελληνική οικονομία και το στοίχημα της ανάκαμψης, η χθεσινολογική αναβάθμιση, από τον αμερικανικό οίκο αξιολόγησης Standard & Poor's της μακροπρόθεσμης πιστοληπτικής ικανότητας της χώρας σε B+, χαιρετίζοντας την απόφαση του Eurogroup . Στις 18:00 το απόγευμα ώρα Ελλάδας, ο πρωθυπουργός θα περάσει το κατώφλι της Downing Street 10 προκειμένου να συναντηθεί με την πρωθυπουργό της Βρετανίας, Τερέζα Μέι. Στη συνέχεια θα έχει συνάντηση με τον αρχηγό του Εργατικού Κόμματος, Τζέρεμι Κόρμπιν.</p>
<p>Gold</p> <p>BART-random</p> <p>mBART25</p> <p>mBART50</p> <p>GreekBART</p>	<p>Συνάντηση με οικονομικούς παράγοντες από το Σίτι του Λονδίνου έχει ο Αλέξης Τσίπρας</p> <p>Μήνυμα Τσίπρα στο Λονδίνο για το χρέος</p> <p>Συνάντηση Τσίπρα με οικονομικούς παράγοντες στο Σίτι</p> <p>Συνάντηση Τσίπρα με οικονομικούς παράγοντες στο Σίτι</p> <p>Βλέμματα στο Λονδίνο για την ελληνική οικονομία</p>

Table b.9 – Example 4-GreekSUM Title

Document	<p>Επιβατικό τρένο εκτροχιάστηκε σήμερα περίπου 20 χλμ. βόρεια της Ραμπάτ, με αποτέλεσμα να σκοτωθούν έξι άνθρωποι και άλλοι 86 να τραυματιστούν, σύμφωνα με επίσημο απολογισμό που ανακοινώθηκε στον τόπο του δυστυχήματος. «Ο εκτροχιασμός προκάλεσε έξι θανάτους, σύμφωνα με τον τρέχοντα απολογισμό, και 86 τραυματίες σε σοβαρή κατάσταση», δήλωσε ο Μοχάμεντ Ραμπί Ραχίλ, γενικός διευθυντής της εταιρίας σιδηροδρόμων ONCF, ο οποίος μετέβη επί τόπου. «Ξεκίνησε έρευνα για τον προσδιορισμό των αιτιών του δυστυχήματος», πρόσθεσε, σε βίντεο που αναρτήθηκε στα μέσα κοινωνικής δικτύωσης. Θεαματικές εικόνες του δυστυχήματος, που σημειώθηκε γύρω στις 13:00 ώρα Ελλάδας, περίπου 20 χλμ. βόρεια της πρωτεύουσας Ραμπάτ, στο ύψος της κοινότητας Σιντί Μπουκναντέλ, κάνουν τον γύρο των μέσων κοινωνικής δικτύωσης, που είναι πολύ επικριτικά εναντίον της ONCF. Οι εικόνες δείχνουν πολλά βαγόνια εκτροχιασμένα κοντά σε μια γέφυρα στους αγρούς, ενώ η μηχανή είναι πλήρως κατεστραμμένη. Ο οδηγός της αμαξοστοιχίας είναι νεκρός, σύμφωνα με πολλά τοπικά ΜΜΕ. Ο βασιλιάς αποφάσισε να αναλάβει τα έξοδα της κηδείας των θυμάτων και οι τραυματίες θα διακομιστούν στο στρατιωτικό νοσοκομείο της Ραμπάτ με βασιλικές οδηγίες, αναφέρεται σε ανακοίνωση του γραφείου του βασιλιά.</p>
Gold	Εκτροχιασμός τρένου στο Μαρόκο: Στους 6 οι νεκροί - 86 τραυματίες
BART-random	Ραμπάτ: 20 νεκροί από εκτροχιασμό τρένου
mBART25	ΗΠΑ: Επιβατικό τρένο εκτροχιάστηκε - Έξι νεκροί και 86 τραυματίες
mBART50	ΗΠΑ: Επιβατικό τρένο εκτροχιάστηκε - Έξι νεκροί και 86 τραυματίες
GreekBART	Εκτροχιασμός τρένου στη Ραμπάτ: Έξι νεκροί και 86 τραυματίες

Table b.10 – Example 5-GreekSUM Title

APPENDIX : VARIOUS EXAMPLES FROM THE WORD SENSE INDUCTION DATASETS

In what follows, we provide four examples from each SemEval WSI train and test datasets. Table c.2 shows the short and incomplete context in the SemEval-2013 training set compared to the counterpart examples from SemEval-2010 shown in Table c.1. The short context in the training set will mainly affect badly the dynamic number of clusters and our main approach since the representation of the target words will be sub-optimal.

<p>The Commission seeks comment on whether the analytical framework that was used to streamline AT &T 's services should be applied to incumbent LEC access services. In particular , the Commission seeks comment on which of the factors that it used in examining AT &T 's pricing behavior could be used to determine when to remove incumbent LEC access services from price cap regulation. It cites demand elasticity , supply elasticity , market share , and the pricing of services under price cap regulation as relevant factors .</p>
<p>This works fine if AudioPlayer is n't going to be subclassed. But what if you were going to create a class called StereoAudioPlayer that is a subclass of AudioPlayer ? This class would want access to the openSpeaker () method so that it can override it and provide stereo-specific speaker initialization. You still do n't want the method generally available to random objects (and so it should n't be public) , but you want the subclass to have access to it-so protected is just the solution .</p>
<p>502.4 Floor or Ground Surfaces. Parking spaces and access aisles serving them shall comply with 302. Access aisles shall be at the same level as the parking spaces they serve. Changes in level are not permitted .</p>
<p>When developing kernel code , it is usually important to consider constraints and requirements of architectures other than your own. Otherwise , your code may not be portable to other architectures , as I recently discovered when an unaligned memory access bug was reported in a driver which I develop. Not having much familiarity with the concepts of unaligned memory access , I set out to research the topic and complete my understanding of the issues .</p>

Table c.1 – Random examples for the target word 'Access' from SemEval-2010 task 14 training set

Baby Welcome to my eBay Shop. Please add me to your list of favourite sellers and
digital jesters guys said they would NEVER add collision detection to TM , as this is
Also in the Spanish version, but more were added especially for the Japanese Complete Editions
destination that you have entered . You can add any number of intermediate waypoints to

Table c.2 – Random examples for the target word 'Add' from SemEval-2013 task 13 training set

In more than four years , 2.2 billion yuan has been invested in the construction of harbors and docks , storage fields , support facilities and infrastructure of the ports and city , creating good conditions for building access to the sea for the Great Southwest .
The FDA is expected to approve today a program granting access free of charge to the drug AZT for children with AIDS .
Federal health officials are expected today to approve a program granting long - deferred access to the drug AZT for children with acquired immune deficiency syndrome .
The dispute stems from pretrial maneuvering in the pending court case , in which prosecutors have been demanding access to a host of internal company memos , reports and documents .

Table c.3 – Random examples for the target word 'Access' from SemEval-2010 task 14 test set

Lewinsky wrote "Return to Sender" on the envelope, adding , "You must be morons to send me this letter!"
For instance, the Post also has the story about the woman meeting with Clinton just days before his first Inaugural, but adds the detail that she says all the encounters were innocent.
if you add the um uh people of various sexual persuasions and those who never intend to marry and those who are retired and those who are um just looking for fun they people with families turn out to be such a small minority that they can't get the tax bill passed no matter what happens
The tripe with onions and garlic is cooked for several hours, posole or hominy is added , along with red chile.

Table c.4 – Random examples for the target word 'Add' from SemEval-2013 task 13 test set

Titre : Exploiter les Modèles de Langage Basés sur les Transformers pour Comblent le Fossé entre le Langage et les Domaines Spécialisés

Mots clés : transformers, modèles de langage, génération du langage naturel, induction du sens des mots, apprentissage non-supervisé, apprentissage profond, multi-modale

Résumé : L'ère des modèles de langage basés sur des 'transformers' a ouvert la voie à un nouveau paradigme dans le traitement du langage naturel (NLP), permettant des performances remarquables dans un large éventail de tâches dans les domaines de la compréhension du langage naturel (NLU) et de la génération du langage naturel (NLG). Cette thèse se penche sur le potentiel de transformation des modèles de langage basés sur les 'transformers' lorsqu'ils sont appliqués à des domaines et des langues spécialisés. Elle comprend quatre projets de recherche, chacun contribuant à l'objectif global d'amélioration de la compréhension et de la génération du langage dans des contextes spécialisés. Pour répondre à la rareté des modèles de langue non anglophones pré-entraînés dans les domaines généraux et spécialisés, nous explorons la création de deux modèles de langue : JuriBERT et GreekBART. JuriBERT est un ensemble de modèles BERT spécifiques au domaine juridique français, et qui répondent aux besoins des professionnels juridiques. JuriBERT est évalué sur deux tâches juridiques françaises provenant de la cour de cassation en France. Les résultats soulignent que certaines tâches spécialisées peuvent être mieux traitées avec de petits modèles spécifiques à un domaine qu'avec leurs homologues génériques de plus grande taille. Nous présentons également GreekBART, le premier modèle Seq2Seq grec. Basés sur BART, ces modèles sont particulièrement bien adaptés aux tâches génératives. Nous évaluons les performances de GreekBART par rapport à d'autres modèles sur diverses tâches discriminatives et évaluons ses ca-

pacités en NLG en utilisant deux tâches génératives grecques de GreekSUM, un nouvel ensemble de données introduit dans cette recherche. Nous montrons que GreekBART est très compétitif par rapport aux modèles linguistiques multilingues et monolingues basés sur BERT, tels que GreekBERT et XLM-R.

Nous examinons ensuite le domaine de la sémantique en tirant parti des représentations vectorielles contextuelles basées sur les 'transformers' pour résoudre le problème de l'induction du sens des mots (WSI). Nous proposons une nouvelle méthode non supervisée qui utilise le regroupement d'informations invariantes (IIC) et le regroupement agglomératif pour enrichir et regrouper les représentations des mots cibles. Une évaluation approfondie sur deux tâches WSI et de multiples modèles de langage pré-entraînés démontre la compétitivité de notre approche par rapport à l'état de l'art.

Enfin, nous présentons Prot2Text, une approche multimodale permettant de générer des fonctions de protéines en texte brut en combinant trois modalités : la structure des protéines, la séquence des protéines et le langage naturel. Prot2Text fait progresser la prédiction des fonctions des protéines au-delà des classifications traditionnelles. Prot2Text intègre des réseaux neuronaux graphiques (GNN) et des grands modèles de langage (LLM) dans un cadre codeur-décodeur. Une évaluation empirique sur un ensemble de données protéiques multimodales montre l'efficacité de Prot2Text, qui offre des outils puissants pour la prédiction de la fonction d'une large gamme de protéines.

Titre : Leveraging Transformer-Based Language Models to Bridge the Gap Between Language and Specialized Domains

Keywords : transformers, language models, natural language generation, word sense induction, unsupervised learning, deep learning, multi-modal

Abstract : The era of transformer-based language models has led the way in a new paradigm in Natural Language Processing (NLP), enabling remarkable performance across a wide range of tasks from both fields Natural Language Understanding (NLU) and Natural Language Generation (NLG). This dissertation delves into the transformative potential of transformer-based language models when applied to specialized domains and languages. It comprises four distinct research endeavors, each contributing to the overarching goal of enhancing language understanding and generation in specialized contexts.

To address the scarcity of non-English pretrained language models in both general and specialized domains, we explore the creation of two language models JuriBERT and GreekBART. JuriBERT is a set of French legal domain-specific BERT models tailored to French text, catering to the needs of legal professionals. JuriBERT is evaluated on two French legal tasks from the court of cassation in France. The findings underscore that certain specialized tasks can be better addressed with smaller domain-specific models compared to their larger generic counterparts. We equally introduce GreekBART, the first Greek Seq2Seq model. Being based on BART, these models are particularly well-suited for generative tasks. We evaluate GreekBART's performance against other mo-

odels on various discriminative tasks and assess its capabilities in NLG using two Greek generative tasks from GreekSUM, a novel dataset introduced in this research. We show GreekBART to be very competitive with state-of-the-art BERT-based multi-lingual and mono-lingual language models such as GreekBERT and XLM-R.

We dive next into the domain of semantics by leveraging the transformer-based contextual embeddings to solve the challenging problem of Word Sense Induction (WSI). We propose a novel unsupervised method that utilizes invariant information clustering (IIC) and agglomerative clustering to enrich and cluster the target word representations. Extensive evaluation on two WSI tasks and multiple pretrained language models demonstrates the competitiveness of our approach compared to state-of-the-art baselines.

Finally, we introduce Prot2Text framework, a multi-modal approach for generating proteins' functions in free text by combining three modalities: protein structure, protein sequence and natural language. Prot2Text advances protein function prediction beyond traditional classifications. Integrating Graph Neural Networks (GNNs) and Large Language Models (LLMs) in an encoder-decoder framework. Empirical evaluation on a multi-modal protein dataset showcases the effectiveness of Prot2Text, offering powerful tools for function prediction in a wide range of proteins.