

Towards Information Retrieval for Scholarly Document Processing

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in

Electronics and Communication Engineering by Research

by

Amit Pandey

2020702009

amit.pandey@research.iiit.ac.in



International Institute of Information Technology

Hyderabad - 500 032, INDIA

May 2024

Copyright © Amit Pandey, 2024
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “**Towards Information Retrieval for Scholarly Document Processing**” by **Amit Pandey**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Advisor: Prof. Vikram Pudi

Dedicated with profound gratitude to the pillars of my life
*my Mother, whose boundless love nourished my dreams, my Father, whose sacrifice
paved the way, my Sister, a constant source of inspiration and shared dreams, my
cherished Wife, whose enduring strength, unwavering support and encouragement turned
aspirations into achievements, and to Almighty God, the ultimate source of wisdom.*

Acknowledgments

As I tender my MS thesis, my heart overflows with gratitude for the remarkable individuals who have been indispensable companions on my IIIT-Hyderabad journey.

Foremost among them, I extend a sincere thanks to Prof. Vikram Pudi, my guide. His unwavering support and guidance have been the driving force propelling this endeavor to fruition. Words fall short of conveying the depth of gratitude for his consistent encouragement and support throughout this transformative journey. Beyond academic and financial aid, I deeply value his serene composure, kindness, and profound perspective on life. In moments of uncertainty and challenges, his guidance was a source of enduring hope.

My gratitude extends to Prof. Makarand Tapaswi, Mr. Vengupoyal Gundimeda, Prof. P K Reddy, and Prof. Naresh Manwani. Their passion and dedication to their work stand as a luminous example, inspiring me throughout. Collaborating with them has offered invaluable lessons in addressing real-world problems, conducting comprehensive research, and elegantly presenting ideas.

I reserve special appreciation for Dr. Narendra Babu, my senior, co-author, and friend, whose guidance helped shape my research ideas. He assumed the role of an elder brother, offering support and wisdom.

A profound debt of gratitude is owed to my cherished friends Omama, Krishna, Sai Teja, Hozaifa, Waqas, and Udit. They not only navigated me through countless queries and doubts but also enriched my journey with meaningful and unforgettable moments, making this experience truly special. Heartfelt thanks to Aarathy Rose, Kinal, Shantanu, Swayatta, Yash, Saideep, Harshit, Neel, Prateek, Abhinaba, Ashish, Zeeshan, Jigyasu, Nikhil, Sagar, Sanjay, Harnadh, Aravind, Ritam, Saurav, and Vilal, whose camaraderie transformed my IIIT experience into cherished memories. Our shared moments were a source of joy and learning. These individuals collectively formed a supportive family, offering solace and strength.

I express my gratitude to the IIIT faculty and staff for fostering an environment that is vibrant and enriching, making IIIT an outstanding institute.

Finally, my heartfelt gratitude goes to my family for unwaveringly believing in me, even during their times of hardship. Moreover, I extend sincere thanks to the almighty for guiding and gracing my journey, playing a pivotal role in helping me achieve this significant milestone.

Abstract

The relentless growth of scholarly publications, exemplified by the annual publication rate exceeding 5 million articles, has posed a formidable challenge for researchers seeking efficient literature review methodologies. Systematic Literature Reviews (SLRs), crucial for understanding existing knowledge and identifying research gaps, are hindered by the manual extraction of information, contributing to extended timelines and potential obsolescence. This thesis addresses the urgent need for improved literature review methodologies by focusing on two challenges: Cited Text Span Retrieval (CTSR) and Named Entity Recognition (NER). CTSR involves identifying cited text spans, facilitating the tracing of information origin, while NER identifies and categorizes entities within the text.

In this thesis, we introduce CitRet, a hybrid model for CTSR, leveraging semantic and syntactic characteristics of scientific documents and outperforming existing methods on the CLSciSumm shared tasks. Using only 1040 documents for finetuning, CitRet achieves a remarkable over 15% improvement in the F1 score evaluation.

Further, we explore Complex NER for English, a non-trivial task of identifying rare and semantically ambiguous entities. Utilizing pre-trained language models, our models consistently outperform the baseline, with the best model advancing the baseline F1-score by over 9%.

Expanding our scope to complex NER for low-resource languages, we leverage pre-trained language models for Chinese and Spanish. Employing Whole Word Masking (WWM) to enhance the Masked Language Modeling objective, our models, incorporating CRF, BiLSTMs, and Linear Classifiers, outperform the baseline by a significant margin. The best-performing model attains a competitive position on the evaluation leaderboard for the blind test set. This work aims to catalyze further research in the challenging domain of ambiguous, low-resource, complex NER.

By addressing CTSR and NER, our thesis contributes significantly to the broader goal of enhancing systematic literature reviews (SLRs). Integration of these tasks provides a structured and comprehensive approach to navigating the vast scientific publication landscape, easing the burden on researchers and promoting a more efficient dissemination of knowledge.

Contents

Chapter	Page
1 Introduction	1
1.1 Motivation	1
1.2 Research Problem: Cited Text Span Retrieval	2
1.2.1 The Significance of Citations and the Evolution of Citation Analysis	2
1.2.2 Exploring Citation Linkage in Scholarly Communication	3
1.3 Research Problem: Named Entity Recognition	5
1.4 Applications of CTSR and NER	9
1.5 Challenges	10
1.6 Key Contributions and Thesis Outline	11
2 Related Work	13
2.1 Word and Word Sequence Representation	13
2.2 Cited Text Span Retrieval	14
2.2.1 History of Cited Text Span Retrieval Tasks	14
2.2.2 Prior Approaches of CTSR	15
2.3 Named Entity Recognition	16
2.3.1 Evolution of Named Entity Recognition	16
2.3.2 Prior Approaches of NER	17
3 A Hybrid Model for Cited Text Span Retrieval	19
3.1 Overview	19
3.2 Introduction	19
3.3 Method	21
3.3.1 Background	21
3.3.1.1 SBERT	21
3.3.1.2 Word Mover’s Distance	22
3.3.2 Contextual Distance	23
3.3.2.1 Finetuning the SBERT	23
3.3.2.2 Weighted contextual embeddings	23
3.3.2.3 Denoising	24
3.3.3 Non-contextual Distance	24
3.4 Detailed Experimental Setup and Analysis	25
3.5 Results	26
3.6 Discussion	27
3.7 Implementation details	28

4	Transformer Based Architectures for Complex NER in English	29
4.1	Overview	29
4.2	Introduction	29
4.3	Task Description	30
4.4	Dataset	30
4.5	Models	31
4.6	Implementation Details	34
4.7	Results	34
4.8	Error Analysis	35
5	Complex NER in Semantically Ambiguous Settings for Low Resource Languages	37
5.1	Overview	37
5.2	Introduction	37
5.3	Task Description	38
5.4	Dataset	39
5.5	System Overview	39
5.6	Implementation Details	42
5.7	Results	42
5.8	Error Analysis	43
6	Conclusion and Future Work	45
6.1	Conclusion	45
6.2	Future Work	46
	Bibliography	49

List of Figures

Figure	Page
1.1 Example of a citation. The text on the left (highlighted in green) is referred to as the Citance text, while the text on the right (highlighted in red) is referred to as the Cited/Reference text.	3
1.2 Example of Epistemic Value Drift. The claim in (Voorhoeve et al., 2006) becomes fact in (Okada et al., 2011).	4
1.3 Labelled Sequence and Explanation for English NER task.	8
1.4 Labelled Sequence and Explanation for Spanish NER task.	9
3.1 Illustration of the CitRet model. WMD and weighted contextual embeddings (WCEs) are calculated for an input pair. The WCEs are then denoised using the common component removal technique. These denoised WCEs are used to find cosine similarity between the sentences of the input pair. Finally, WMD and cosine scores are added, and top k similar sentences in an RP for a citance are retrieved.	21
3.2 n-gram intersection of two sentences	23
3.3 Flow between 2 sentences S_0 and S_1 using WMD	25
4.1 BERT-based architecture	31

List of Tables

Table	Page
3.1 Performance comparison of our model with the baseline models. The last four rows show the ablation study of our model marked with †. D denotes the denoising step.	27
4.1 Total sentences in English monolingual track	33
4.2 Entity types in the label space	33
4.3 Results of our models on validation dataset	33
4.4 Comparison of model performances with the baseline on the validation dataset .	34
4.5 Performance of model on test dataset	35
5.1 Total sentences in Chinese and Spanish monolingual track	41
5.2 Entity Types in the label space	41
5.3 Results of our models on validation dataset for the Spanish language.	41
5.4 Results of our models on validation dataset for the Chinese language.	42
5.5 Performance of the Spanish model on the test dataset.	43
5.6 Performance of the Chinese model on the test dataset.	44

Chapter 1

Introduction

1.1 Motivation

Scientific documents and publications are the cornerstone of academic knowledge dissemination, providing a repository of critical findings and insights. The rapid proliferation of scholarly content, exemplified by an exponential increase in the annual publication rate, poses a significant challenge for researchers. In 2022, over 5 million academic articles were published, contributing to a cumulative total surpassing 200 million articles. Navigating this extensive literature becomes increasingly daunting, leading to the inadvertent oversight of crucial scientific discoveries [76].

The traditional method of conducting literature reviews, notably the Systematic Literature Review (SLR), has become a formidable task in this voluminous landscape. The importance of such reviews cannot be overstated, as they provide a comprehensive understanding of existing knowledge, identify research gaps, and guide future investigations. However, the sheer volume of publications makes the manual extraction of relevant information a time-consuming and error-prone process. Researchers may need to sift through dozens or even hundreds of papers to gain a rudimentary grasp of the state-of-the-art in a particular domain, increasing the likelihood of overlooking significant results.

SLRs usually demand a multidisciplinary team with specialized skills, including subject matter experts. Despite assembling such teams, SLRs remain time-consuming endeavors because of the intricate nature of SLRs, involving time-consuming and labor-intensive steps of data extraction, analysis, summarization, and synthesis, and the duration to complete an SLR may range from six months to 1.5 years [3][1]. The extended timeline has substantial implications for accuracy and relevance, as 23% of SLRs are considered outdated within two years of publication due to emerging evidence.

The comprehensive manual effort required in SLRs poses a scalability challenge, and addressing this challenge necessitates innovative approaches. Tools and methodologies such as Question Answering (QA) models, Retrieval-Augmented Generation (RAG), and summariza-

tion present promising avenues to streamline the literature review process. These techniques aim to automate the extraction of relevant information, thereby facilitating quicker and more precise comprehension of scientific content.

Our motivation aligns with this urgent need for improved literature review methodologies, echoing the efforts of SciAssist by WING@NUS [2] and PaperQA [76]. In this context, our thesis focuses on two specific challenges: Cited Text Span Retrieval (CTSR) and Named Entity Recognition (NER). These tasks, while seemingly narrow in scope, play a crucial role in addressing the broader issue of facilitating effective literature review.

Cited text span retrieval involves identifying and extracting portions of text referenced in a document, offering a means to trace the origin and context of cited information. Simultaneously, named entity recognition focuses on identifying and categorizing entities (e.g., authors, concepts) within the text. These tasks collectively contribute to a more automated and efficient literature review process, forming the foundation for enhanced tools to alleviate the burden on researchers.

By addressing CTSR and NER, our thesis contributes to the broader goal of making systematic literature reviews more accessible and effective. These tasks, when integrated, provide a structured and comprehensive approach to navigating the vast landscape of scientific publications, ultimately easing the burden on researchers and promoting a more efficient dissemination of knowledge.

In the subsequent parts of this section, we explore the formal definitions, motivations, and challenges of the tasks undertaken, while also delineating the contributions of our research work.

1.2 Research Problem: Cited Text Span Retrieval

1.2.1 The Significance of Citations and the Evolution of Citation Analysis

In the realm of academic communication, citations function as crucial markers, symbolizing the interconnected network of ideas, knowledge, and scholarly progress. The act of citation goes beyond mere formality and a simple acknowledgment; it constitutes a nuanced dialogue where scholarly works validate their arguments and engage thoughtfully with prior research through precise bibliographic references. Citations construct a contextual bridge to previous works and assist readers in navigating the chronological progression of knowledge, aiding in a comprehensive understanding of the intellectual landscape and laying the groundwork for new contributions.[112]

The inception of citation analysis can be traced back to Eugene Garfield's pioneering work on citation analysis as a tool for journal evaluation, as elaborated upon in his 1972 publication [41]. Garfield's introduction of citation indexing marked a significant development in the evaluation

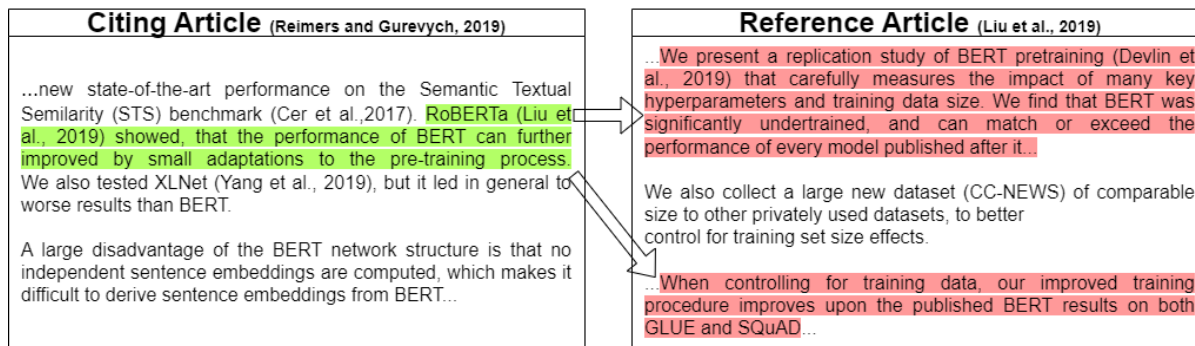


Figure 1.1: Example of a citation. The text on the left (highlighted in green) is referred to as the Citance text, while the text on the right (highlighted in red) is referred to as the Cited/Reference text.

of scholarly impact. This innovative approach involved creating indexes that encompassed all references in a research document. By examining citation patterns, Garfield proposed a quantitative measure for assessing the influence and contribution of academic journals. The frequency of citations to a journal’s articles emerged as a tangible metric, reflecting its influence within the scientific community.

1.2.2 Exploring Citation Linkage in Scholarly Communication

In the landscape of scientific publications, a fundamental practice involves citing referenced documents. However, a notable limitation exists within this practice- a citation, while indicating a connection to another document, remains silent on the specific span of text within that document to which it refers. The impediments posed by this limited citation specificity are listed below.

- **Compelled Comprehensive Reading Due to Unspecified Citation Context:** As a consequence of this inherent limitation, interested readers encountering a citation encounter an obstacle. The citation, devoid of information regarding the exact span of text in the referenced document, necessitates a comprehensive study of the entire cited document for those seeking a deeper understanding of the issue at hand. This requirement not only imposes a time-intensive task on the reader but also poses efficiency challenges in extracting relevant information. Consequently, the scholarly discourse is hindered by an inherent lack of precision and accessibility.
- **Subjectivity and Bias in Cited Information:** Adding to the complexity, the information ascertained about a referenced paper is susceptible to subjectivity and potential bias introduced by the citing author(s) [139]. As shown in figure 1.2 (directly picked from

[27]), this subjectivity can significantly alter the epistemic value of claims presented in the cited work [27]. The citing author’s intentions, opinions, and biases may influence the interpretation of the referenced paper, leading to variations in the perceived reliability and objectivity of the cited information. Such subjective interpretations might result in reporting preliminary results or claims as definite facts, introducing an additional layer of complexity to the reliability of scholarly information.

Reference Article

(Voorhoeve et al., 2006): “*These miRNAs neutralize p53-mediated CDK inhibition, possibly through direct inhibition of the expression of the tumor suppressor LATS2.*”

Citing Article

(Okada et al., 2011): “*Two oncogenic miRNAs, miR-372 and miR-373, directly inhibit the expression of Lats2, thereby allowing tumorigenic growth in the presence of p53 (Voorhoeve et al., 2006).*”

Figure 1.2: Example of Epistemic Value Drift. The claim in (Voorhoeve et al., 2006) becomes fact in (Okada et al., 2011).

Limitations of Traditional Citation Metrics: The traditional metric of a paper’s impact based solely on the count of citations is not immune to these challenges. While often considered a measure of a paper’s influence, the count of citations can be influenced by the subjective interpretations and biases of citing authors. This reliance on quantitative metrics alone may not accurately capture the true impact or significance of a paper, necessitating a shift towards more nuanced and qualitative approaches capable of overcoming inherent biases in scholarly discourse.

This establishes the foundation for delving into the analysis of **citation context** and **citation linkage analysis** as pivotal augmentations to scholarly communication. In the context of a research article, a citation typically pertains to a segment within the referenced paper, which is defined as the citation context. The citation context, in practice, may encompass anywhere from a single sentence to several paragraphs. The process of associating a citation with a specific span of text is termed citation linkage [112].

Citation Linkage, by identifying precise text spans referenced by citations within a paper, can help us improve the assessment of a referenced publication’s impact, highlights, and weak-

nesses, contributing to refined citation-based metrics, enhanced information retrieval, and improved literature navigation tools, thereby elevating the qualitative aspect of citation evaluation. Furthermore, in the context of automated related work generation, grounding outputs in the content of cited papers becomes imperative to avoid inaccuracies and ensure factual precision [70]. The formal details of our approach to this Cited Text Span Retrieval task are covered in Chapter 3.

1.3 Research Problem: Named Entity Recognition

Named Entity Recognition (NER) stands as a foundational task in Natural Language Processing (NLP), pivotal in transforming unstructured text into structured, interpretable information. This process involves identifying and classifying entities within a text, such as persons, organizations, locations, dates, and more. The representation of named entities provides crucial contextual information, enabling machines to comprehend and process language more effectively.

Named Entity Recognition (NER) involves several key steps, including tokenization, span detection, and classification, to extract and understand entities in text. The steps are briefly described below.

1. **Tokenization and Contextual Representation:** The first step in NER is to tokenize the input text into individual words or tokens. Tokenization is crucial as it breaks down the text into smaller units for analysis. The tokens are then embedded into a contextual representation, capturing the relationships and meanings between words in the sentence. Contextual embeddings, often generated by pre-trained language models like BERT or GPT, enhance the model’s understanding of the text.
2. **Span Detection and Classification:** The span detection step involves identifying the beginning and end of a sequence of tokens that form a named entity. This is crucial for delineating the boundaries of the entity within the text. Once the span of the named entity is detected, the next step is to classify the type of entity it represents. This involves assigning a predefined label to the identified span, such as person, organization, location, date, etc. In standard NER tasks, the BIO (Begin, Inside, Outside) tagging scheme is commonly employed for classifying entity spans within a text. Entities are labeled with B-<entity_type> for the beginning, I-<entity_type> for inside, and O for outside the entity span. This tagging system allows for the explicit delineation of entities and their boundaries.

NER finds extensive use in various domains, offering substantial benefits in applications such as information retrieval, question answering, and text summarization. In the general domain,

NER has been a well-studied problem, with models achieving state-of-the-art results on datasets like CoNLL 2002 [117] and CoNLL 2003 [118]. However, it has been noted that these impressive results can be attributed to their training on well-structured news text, the prevalence of "easy" entities like person names, and the potential for memorization due to entity overlap between training and testing sets [9].

As NER technology matures, contemporary challenges emerge, particularly in recognizing Complex Named Entities (Complex NEs). Traditional NER systems, trained on general domain data, encounter difficulties when confronted with these Complex NEs, and it results in a significant decline in performance [88, 37]. The syntactic diversity and ambiguity of Complex NEs pose challenges in contextual recognition, as they may not conform to the conventional structure of named entities such as persons or locations.

The complexities are exemplified in domains involving creative works and scientific literature, where named entities are not only diverse but also exhibit long-tail distributions. Scientific entities, often described in technical language, pose challenges in NER due to their complex, rare, and domain-specific nature[83].

Complex NEs, particularly those associated with Creative Works (CW), present additional intricacies with fine-grained classifications such as SCIENTIST AND ATHELETE [38]. Recognizing these entities requires nuanced understanding and the ability to discern subtleties in the linguistic composition.

Recent studies underscore the difficulties in processing complex and long-tail named entities. Even state-of-the-art pre-trained Transformers face limitations without external knowledge, necessitating the infusion of transformers with knowledge bases and gazetteers[38]. However, such solutions prove brittle against out-of-knowledge-base entities and scenarios involving spelling mistakes or typos. The emergence of new entities compounds these difficulties, creating a scenario where entity types are open classes, continually evolving with faster growth rates in specific categories. This evolving landscape demands test sets with a multitude of unseen entities to simulate an open-world setting. Moreover, these challenges are amplified for low-resource languages, which lack foundational work compared to well-established languages like English. Addressing NER complexities in such linguistic landscapes requires tailored solutions and underscores the need for broader research efforts in these domains. A list summarizing the complexities we've discussed is provided below.

- **Noisy NER:** A user input contains a typo: "somy xpria" instead of "Sony Xperia." Gazetteer-based models, relying on exact matches, struggle to recognize the entity due to the misspelling, significantly degrading their performance.
- **Ambiguous Entities and Contexts:** The term "Inside Out" may refer to a movie in some contexts but could have a different meaning in other contexts, such as describing an emotional state. Ambiguous entities like "Inside Out," "Among Us," and "Bonanza" pose

challenges as they resemble regular syntactic constituents and may or may not be entities based on the context.

- **Complex Entities - General Domains:** In the context of creative works, entities like "Eternal Sunshine of the Spotless Mind" present linguistic complexity as they are expressed as complex noun phrases. Current parsers/NER systems often struggle to recognize such entities, and syntactic parsing becomes challenging.
- **Ambiguous Entities and Contexts - Voice and Search Domains:** In voice or search domains, ambiguity arises with short inputs. For instance, the entity "Bonanza" could refer to a TV show in some contexts, but in the context of a search query, its intended meaning may differ. The lack of surface features like capitalization/punctuation in short inputs complicates the NER task.
- **Emerging Entities:** In domains with growing entities, such as the release of new books, songs, or movies, there's a continuous influx of entity types. For instance, a recently released book, not present in the training data, poses a challenge for NER systems. True generalization requires test sets with numerous unseen entities to mimic an open-world setting.

This thesis aims to explore, analyze, and propose solutions for the complexities inherent in NER tasks. By delving into the nuances of these challenges, we seek to contribute to the advancement of NER methodologies, fostering a deeper understanding of language representation in intricate and specialized contexts.

We use the MultiCoNER dataset [81] for our task, selected for its capacity to mirror real-world scenarios. The MultiCoNER dataset stands out as an extensive and intricate resource designed for Multilingual Complex Named Entity Recognition (NER). Encompassing various domains such as Wiki, question, and search queries across 11 languages, it offers a comprehensive representation of the challenges inherent in NER research. This dataset deliberately incorporates entities that are low-context, structurally complex, semantically ambiguous, and emerging in nature, making it particularly pertinent for tackling the complexities associated with recognizing entities like titles of creative works.

The deliberate inclusion of such entities ensures that a substantial number of pre-trained language models lack prior exposure to them, thereby enhancing the realism of the task. Additionally, the dataset's strategic design results in a test data quantity that surpasses that of the training data by over 100 times, effectively mitigating the issue of potential memorization due to entity overlap between training and testing sets, which has been a concern in previous tasks. This thoughtful approach addresses and rectifies problems encountered in earlier tasks and ensures a more robust evaluation of the NER model's generalization capabilities. Figure 1.3 and Figure 1.4 consist of labelled sequence and the explanation of the assigned NER tags.

Labelled Sequence	Comments
<p>1. the Hubble B-PROD Space I-PROD Telescope I-PROD is a space telescope that was launched into low B-LOC earth I-LOC orbit I-LOC.</p> <p>2. her debut album was named eye B-CW to I-CW the I-CW telescope I-CW</p> <p>3. Richwoods is southeast of Palestine B-LOC.</p> <p>4. The film also suggests that he worked behind the lines for Wadie B-PER Haddad I-PER, a branch of the Popular B-GRP Front I-GRP for I-GRP the I-GRP Liberation I-GRP of I-GRP Palestine I-GRP</p>	<p>In sentence 1, the tokens "Space" and "Telescope" appear twice, and are classified differently. Also, since the token "Low" appears with "Earth Orbit", it is classified as LOC.</p> <p>Similarly, "Palestine" has been tagged as LOC and GRP.</p> <p>This illustrates that NER depends on the context as well as the span of the entity.</p>
<p>5. all isolated topologies include a transformer B-PROD and thus can produce an output of higher or lower voltage.</p> <p>6. a list of characters who appeared in the 1988 anime series Transformers B-CW Supergod I-CW Masterforce I-CW</p> <p>7. the car turns out to be a transformer B-PROD.</p>	<p>In Sentence 5, "transformer" is correctly tagged as a PROD (electrical device). In Sentence 6 "Transformers" is correctly tagged as CW (fictional series).</p> <p>However, in Sentence 7, since seen in limited context along with the term "car", "transformer" is incorrectly tagged as a PROD, instead of CW. This illustrates the difficulty in discerning between different entity categories based on context.</p> <p>Note: sentence 7 is an incorrectly labeled training sample.</p>
<p>8. thousand B-CW splendid I-CW suns I-CW</p> <p>9. ford B-PROD cargo I-PROD price</p> <p>10. let us play among B-CW us I-CW</p>	<p>Very short context and ambiguous entities</p>
<p>Labels: Beginning (B), Inside (I), Outside (O), Product (Prod), Creative Work (CW), Person (PER), Location (LOC), Group (GRP)</p>	

Figure 1.3: Labelled Sequence and Explanation for English NER task.

This thesis aims to explore, analyze, and propose solutions for the complexities inherent in NER tasks, with a specific focus on Creative Works and Scientific Domains. By delving into the nuances of these challenges, we seek to contribute to the advancement of NER methodologies, fostering a deeper understanding of language representation in intricate and specialized contexts. The formal details of our approach to Complex NER in English and Complex NER for Low Resource Languages (Chinese and Spanish) tasks are covered in Chapter 4 and Chapter 5 respectively.

Labelled Sequence	Comments
<p>1. ahora usando python B-CW para sumar dos números con el objeto calculador.</p> <p>English: Now use Python to add two numbers with the calculator object.</p> <p>2. Optimus aparece de nuevo en la película, Transformers B-CW: el I-CW lado I-CW oscuro I-CW de I-CW la I-CW luna I-CW.</p> <p>English: Optimus appears again in the movie Transformers: The Dark Side of the Moon.</p> <p>3. Este diseño ha pasado a denominarse telescopio B-PROD herscheliano I-PROD.</p> <p>English: This design has come to be called the Herschelian telescope.</p> <p>4. Medios como el B-GRP país I-GRP consideraron estas declaraciones discriminatorias hacia la comunidad árabe de origen palestino.</p> <p>English: Media outlets, like the country, considered these statements discriminatory towards the Arab community of Palestinian origin.</p> <p>5. El viaje contó con tres etapas: Amán B-LOC (Jordania B-LOC), Belén B-LOC (Palestina B-LOC), y Jerusalén B-LOC (Israel B-LOC).</p> <p>English: The journey had three stages: Amman (Jordan), Bethlem (Palestine), and Jerusalem (Israel).</p>	<p>In sentence 1, "Python" a programming language is correctly tagged as CW. Sometimes it is also labeled as PROD.</p> <p>In sentence 2, especially when seen in comparison with sentence 1, it is interesting to note the challenge posed by both the span of the entity as well as general domain words such as "luna".</p> <p>In sentence 3, "telescopio" is labeled as PROD because it is seen in the context of "herscheliano" and refers to a unique product.</p> <p>In sentence 4, "palestino" is not labelled as LOC since it translates to "Palestinian", but in sentence 5, since it translates to "Palestine" it has been labelled as LOC</p>
<p>Labels: Beginning (B), Inside (I), Outside (O), Product (Prod), Creative Work (CW), Person (PER), Location (LOC), Group (GRP)</p>	

Figure 1.4: Labelled Sequence and Explanation for Spanish NER task.

1.4 Applications of CTSR and NER

Joint applications of Cited Text Span Retrieval (CTSR) and Named Entity Recognition (NER) can enhance the overall understanding of the literature and improve information retrieval in various domains. Here are some potential joint applications:

- Academic Literature Mining:** Combined CTSR with NER can facilitate the extraction of specific spans related to citations, aiding researchers in comprehending the context of citations and the entities mentioned. This enhances the literature review processes. In domains like Biomedical Text Mining, this integration can accelerate the otherwise time-consuming extraction of information from cited research papers, clinical trials, and named entities such as genes, proteins, and diseases.

- **Automated Summarization:** Jointly identifying cited text spans and named entities contributes to more informative and contextually rich document summaries, enabling a swift grasp of key concepts and entities in academic papers or lengthy documents.
- **Semantic Search and Information Retrieval:** The integration of CTSR with NER can enhance semantic search capabilities, enabling users to search for specific entities in the cited context. This can result in more precise and relevant search outcomes, further contributing to improved content recommendation.
- **Knowledge Graph Construction and Database Creation:** Named entities within cited text spans can be leveraged to construct or enhance knowledge graphs, establishing connections between entities in the literature. This process not only improves relationship representation in specific domains but can also contribute to the creation of structured and searchable databases.
- **Legal Document Analysis:** In legal texts, the combination of CTSR with NER can aid legal professionals in efficiently extracting pertinent information. This process assists in locating references to legal cases, statutes, and entities, ultimately establishing the legal basis for arguments.
- **Media Analysis:** Combining CTSR with NER in news articles or media content can enhance the extraction and comprehension of references to individuals, organizations, locations, and events. This integration, significant in the context of extensive data, including social media, can notably contribute to advanced media analysis and is particularly valuable for efficient claim and fact-checking processes.

1.5 Challenges

- **Low Annotated Data and Deep Learning Requirements:** Most existing efforts are centered on common language corpora, like news articles, Twitter posts, and online product reviews. The applicability of models trained on such data to specialized domains, especially scientific literature, poses significant challenges. Deep learning models, widely utilized in both cited text span retrieval and NER, demand substantial annotated data, and annotating data for scholarly literature requires domain expertise.
- **Biases in Annotated Data for Cited Text Span Retrieval:** The challenge intensifies for cited text span retrieval, where annotated data is scarce and, if available, exhibits bias due to the limited representation of positive classes. This hampers the development of accurate models, impacting the ability to generalize effectively to different domains [112]

- **Neglect of Scientific Texts’ Unique Challenges:** Limited attention has been given to the distinctive challenges associated with understanding scientific texts. These challenges include idiosyncratic writing styles, specialized article organizations, and domain-specific vocabularies uncommon in other text genres [47].
- **Ambiguity and Variability in Scientific Language:** The automated extraction of claims from scientific papers faces difficulty due to inherent ambiguity and variability in natural language. Even seemingly straightforward tasks, such as isolating reported values for physical quantities, encounter complications arising from domain-specific conventions and referencing of named entities.
- **Esoteric Encoding of Literature:** Scientific literature often employs esoteric encoding, adding another layer of complexity to cited text span retrieval and NER. Deciphering and interpreting such encoding require specialized attention, making it essential for the effective application of these techniques in scientific contexts.

In summary, the challenges in Cited Text Span Retrieval and Named Entity Recognition extend beyond conventional constraints, necessitating domain-specific adaptations to effectively address the intricacies of scientific literature.

1.6 Key Contributions and Thesis Outline

1. We comprehensively explore the research landscape pertaining to Word and Word Sequence Representation, Cited Text Span Retrieval, Sequence Labelling, and Named Entity Recognition. Our survey encompasses various aspects, investigating how these tasks pose challenges in complex settings marked by structural ambiguity, low resources, or the need for domain expertise. We discuss different tasks and challenges organized to address these issues, highlighting the frequent utilization of various pre-trained models and other techniques to tackle these complexities. We elaborate on the related work in Chapter 2.
2. We introduce CitRet, a novel hybrid Cited Text Span Retrieval (CTSR) model. This model, designed with simplicity and efficacy in mind, demonstrates a remarkable capability to operate with reduced data requirements for fine-tuning. CitRet utilizes the distinctive semantic and syntactic structural characteristics inherent in scientific documents. This unique approach allows us to achieve superior performance with significantly less data for fine-tuning—specifically, a mere 1040 documents. CitRet combines mildly-trained SBERT-based contextual embeddings with pre-trained non-contextual Word2Vec embeddings, harnessing semantic textual similarity to enhance accuracy. Moreover, its computational efficiency stands as a key advantage in resource-conscious environments.

The proposed CTSR model surpasses the current state-of-the-art (SOTA) by a substantial margin, achieving a notable advancement of over 15% in the identification of cited text spans. We present the details of CitRet in Chapter 3.

3. We address Named Entity Recognition (NER), with a specific focus on Complex NER in English and Low Resource NER in languages such as Spanish and Chinese. Through experimentation with transformer-based models like BERT-Linear, BERT-CRF, and BERT-BiLSTM-CRF, we demonstrate that simpler models outperform larger ensembles and even surpass certain systems trained on additional gazetteer-based data. In the Low Resource NER setting, our investigation extends to Spanish and Chinese, employing various BERT-based architectures. Notably, the Whole Word Masking (WWM) strategy proves effective, particularly for languages like Chinese. Our models achieve competitive rankings in the MultiCoNER shared task, underscoring their effectiveness. For a more in-depth exploration of our approach to this task, refer to Chapters 4 and 5.
4. We perform extensive experimentation, result compilation, and error analysis, specifically focusing on Cited Text Span Retrieval (CTSR) and Named Entity Recognition (NER) in diverse settings. The conducted experiments offer nuanced insights into the challenges and intricacies of CTSR and NER, forming a foundational pillar for our research contributions. This thorough exploration of experimental results enriches the scholarly discourse on these tasks in different contexts.
5. In Chapter 6, we summarize our findings and explore potential directions for future research in these domains.

Chapter 2

Related Work

In this section, we embark on a comprehensive exploration of the related work that lays the foundation for our research. We endeavor to contextualize our study within the broader landscape of natural language processing and information retrieval. We delve into various research domains that not only serve as influential methodologies but also encapsulate tasks intrinsic to our work. Among these, Cited Text Span Retrieval (CTSR) and Named Entity Recognition (NER) are a pivotal focus, given their paramount role in our research objectives.

We first briefly touch upon text and sentence representation, which is foundational to all natural language processing (NLP) challenges.

For Cited Text Span Retrieval, we dissect the methodologies employed for retrieving specific spans of text within documents, particularly in the context of academic literature.

The examination of NER encompasses a meticulous review of sequence labeling approaches, shedding light on the diverse strategies employed for entity recognition. We also scrutinize the influence of pre-trained language models on NER, exploring how these models have revolutionized the accuracy and efficiency of named entity identification. Our investigation extends to the occurrence of Named Entities in low-resource and complex settings, acknowledging the importance of making NER accessible and effective across diverse linguistic landscapes.

2.1 Word and Word Sequence Representation

In the realm of text representation, the foundational task involves capturing the meaning of words and word sequences. The distributional hypothesis, asserting that words with similar meanings occur in similar contexts, has driven this field. Early count-based models, exemplified by Sahlgren [107], relied on term co-occurrence matrices and matrix factorization. Mikolov et al. introduced Word2Vec in 2013, with variants such as Continuous Bag of Words (CBOW) and Skip-Gram optimizing context word prediction [90, 91]. These embeddings revolutionized the field by efficiently representing word meanings. Moving beyond individual words, understanding complex language units necessitates encoding word sequences. Traditional methods

like one-hot vectors, Bag of Words (BoW)[12, 86], and TF-IDF [113, 106] suffered from high dimensionality and sparsity. With the rise of deep learning, word embeddings gained prominence for sequence representation. Compositional models aimed to represent sequences as dense vectors, applying composition functions to word embeddings. Vector averaging emerged as a common composition function [65, 40, 91, 92, 49, 42, 93, 7, 35, 111]. Notably, it involves calculating the component-wise mean of word embeddings in a sequence. Various works, including [91] and [42], demonstrated the effectiveness of such representations. In recent advancements, pre-trained transformer-based language models, exemplified by BERT and its variants (e.g., ELECTRA, ALBERT, and RoBERTa), have dominated NLP tasks with state-of-the-art results. These models leverage attention mechanisms and multi-task learning [32, 26, 64, 73]. BERT, for instance, uses self-attention to generate embeddings for tokens, including a special [CLS] token representing the entire sequence. While BERT embeddings are commonly obtained by averaging token embeddings or using the [CLS] token, research by Reimers and Gurevych [103] suggests that, in an unsupervised regime, BERT embeddings for sentence embeddings might perform inferiorly compared to averaging GloVe embeddings. The authors' rationale for the inferior performance of BERT embeddings in an unsupervised regime is attributed to the incompatibility of BERT-generated sentence embeddings with cosine similarity or Euclidean distance metrics [103].

2.2 Cited Text Span Retrieval

2.2.1 History of Cited Text Span Retrieval Tasks

The history of Cited Text Span Retrieval tasks traces back to the Text Analysis Conference (TAC) 2014, specifically BioMedSumm Task3. This pilot task marked a significant milestone as the first initiative to provide annotated resources with citing and cited sentences, laying the foundation for biomedical article summarization. The task introduced sub-tasks addressing various steps crucial for an efficient scientific summarization system, focusing on cited text spans.

Building on this initiative, the CL-SciSumm Shared Tasks emerged, further promoting the identification of cited text spans and their utilization in generating scientific summaries. In 2016, the second CL-SciSumm [52] Shared Task took place as part of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) workshop at the Joint Conference on Digital Libraries (JCDL 2016). Subsequently, from 2017 to 2019, CL-SciSumm was colocated with BIRNDL at the ACM Conference on Research and Development in Information Retrieval (ACM SIGIR 20172019) [51, 53, 21]. CL-SciSumm 2020 was organised as a part of Scholarly Document Processing at EMNLP 2020 [20].

2.2.2 Prior Approaches of CTSR

The primary objective of the CL-SciSumm Shared Task is to unite the summarization community in addressing challenges related to scientific communication summarization.

The task of CTSR requires modeling the relationship (similarity) between a citing and a candidate cited sentence. Early systems proposed using features based on TFIDF [137, 17, 97] and n-grams or sentence graph overlap [5, 58] in order to calculate similarity scores between the citing sentence and candidate sentences. Similarity measures such as Jaccard similarity and cosine similarity were commonly used to solve this task. [15, 30, 57, 96]. The problem has also been posed as a binary classification problem in [29, 137, 139]. In addition to traditional features such as TF-IDF and n-grams, prior methods have also proposed using learned distributed vector space representation (word embeddings) based features since they contain the semantic similarity information at the word level. Models using both non-contextual embeddings such as Word2Vec and contextual embedding methods like BERT have been utilized to find these word embeddings. These extracted features are further used as an input to machine learning algorithms like SVM [79], random forests [122], Word Mover’s Distance [68], CNN [69, 4] or XGBoost [114, 96]. Furthermore, many approaches even adopted voting mechanisms and ensemble techniques on top of their models to improve their metrics [19, 122, 79, 80, 98]. The current best-performing models exploit transformers fine-tuned on very large datasets [19, 138]. [19] also experimented with adding document-level features to the model using special tokens. Other noteworthy approaches, like [10] formulated the task as a search problem and used a two-step approach for retrieving relevant sentences for a given citation. They first find candidate sentences using Apache Solr and BM25 and then re-rank the retrieved sentences using a computationally expensive BERT-based re-ranker.

CTSR as Semantic Textual Similarity: We model the problem as a semantic textual similarity (STS) task. To this end, learning sentence embeddings instead of word embeddings has shown promise and improvement in performance [103]. Using pooling strategies such as mean or max pooling of word embeddings has proven to be an efficient way of obtaining sentence embeddings. SBERT [103] by default uses mean pooling. [23] further explored generalized pooling strategies to enhance sentence embeddings. CNN-based models have also been used to encode sentences into fixed length vectors [56]. To improve performance on sentence matching tasks, [72] proposed syntax- and semantics-aware BERT(SS-BERT), which implicitly integrates syntactic and semantic information of sentences. [120] showed that sentence embeddings could be further improved by employing principal component removal based denoising as a post-processing step.

2.3 Named Entity Recognition

2.3.1 Evolution of Named Entity Recognition

The inception of Named Entity Recognition (NER) tasks gained prominence through benchmark datasets like CoNLL 2002 [117] and CoNLL-2003 [108] shared tasks, focusing on language-independent named entity recognition. These tasks concentrated on identifying persons, locations, organizations, and miscellaneous entities. As NER tasks expanded into real-world settings, addressing long-tailed entities became imperative, leading to the exploration of deep learning models for supervised training. However, the annotation of large amounts of token-level data for NER tasks remained a substantial challenge.

In response to these challenges, and to address the need for real-world applicability in complex settings, MultiCoNER 2022 [84] was organized, featuring 13 tracks focusing on detecting semantically ambiguous and complex entities in various languages. The dataset encompassed 11 languages and included entities such as media titles, products, and groups. By dividing the task into 13 tracks, MultiCoNER 2022 encouraged participants to build monolingual NER models for individual languages, as well as multilingual models capable of working across all languages.

Building upon the success of MultiCoNER 2022, MultiCoNER 2 (SemEval-2023 Task 2) [38] continued to push the boundaries of NER. The dataset, known as MULTICONER V2, comprised 2.2 million instances across 12 languages. The data extraction strategy involved sentences from localized versions of Wikipedia, with entities interlinked and resolved using Wikidata as a reference. To enhance the challenge for models, the text was preprocessed by lowercase conversion and punctuation removal, resulting in more representative and challenging sentences reflective of real-world data.

Furthermore, MultiCoNER 2 introduced a fine-grained NER taxonomy building upon the WNUT 2017 classification [31]. It featured 33 fine-grained classes grouped across six coarse types, allowing for a more nuanced understanding of complex entities. This fine-grained taxonomy enabled the identification of specific types of entities, including those with complex structures (e.g., Creative Works) or entities that are ambiguous without context (e.g., distinguishing between SCIENTIST and ATHLETE within the PER coarse-grained type).

In the landscape of Named Entity Recognition (NER), the challenges posed by identifying named entities in scientific documents are particularly intricate, given their long-tailed nature, rarity, and domain specificity. The inherent complexity of entities in scientific literature underscores the necessity for specialized approaches tailored to this unique domain.

One notable initiative on this scientific front is the DEAL (Detecting Entities in the Astrophysics Literature) shared task, documented in the Workshop on Information Extraction from Scientific Publications (WIESP) during ACL-IJCNLP 2021 [44]. DEAL specifically addressed NER within astrophysics publications, necessitating the recognition of entities ranging from sim-

ple elements like URLs to highly unstructured components such as formulas. The labels for this task were meticulously curated by domain experts, encapsulating entities deemed relevant to the astrophysics community. This ranged from entities useful to researchers, like "Telescope," to those valuable to archivists and administrators, such as "Grant." DEAL effectively navigated the complexities inherent in entity recognition within scientific literature, highlighting the diverse nature of entities within this domain.

In addition to DEAL, recent works have delved into various aspects of NER within scientific documents. Some focus on detecting biomedical entities [59] or scientific entities like tasks, methods and datasets [74, 55, 89], while others concentrate on specific entity types like dataset names or polymer names in materials science publications [47]. The diverse array of scientific NER tasks emphasize the critical need for specialized approaches capable of addressing the intricacies inherent in different scientific domains.

2.3.2 Prior Approaches of NER

- **Sequence Labelling in NER:** In the realm of Named Entity Recognition (NER), sequence labelling is a fundamental task involving the classification of each token within a sequence by assigning a specific label. Traditional approaches to NER treated it as a sequence labelling problem, employing models such as CRF (Conditional Random Fields) [61] and HMM (Hidden Markov Model) [24]. Recent advancements have introduced deep learning models, exemplified by Bidirectional LSTM with a CRF layer [66], to address the inherent challenges. The high cost of labelling, especially for rare, long-tailed entity types, prompted the exploration of methods such as Active Learning [43], Distant Supervision [124], and Reinforcement Learning-based Distant Supervision [95, 135]. Liu et al. [71] focused on detecting dataset mentions in scientific text, utilizing data augmentation to mitigate label scarcity successfully.
- **Pre-trained Language Models for NER:** The introduction of pre-trained language models, particularly transformer-based models like BERT [33], has revolutionized NER methodologies. These models leverage transfer learning, with approaches such as Bidirectional LSTM with a CRF on top [48] and BERT with a CRF layer [50]. Utilizing a BERT-based model with a CRF layer, competitive performance has been achieved in low-resource NER tasks across multiple languages, surpassing baseline performance significantly. CAIR-NL [109] employs a multi-objective joint learning system (MOJLS) that intricately enhances the representation of low-context and fine-grained entities. The training procedure involves minimizing Representation Gaps (Addressing gaps between fine-grained entity types within a coarse-grained type.), Information Augmentation (Reducing representation gaps between an input sentence and the input augmented with external information for a given entity.), Negative Log-Likelihood Loss (Employing this

loss function to refine model predictions.), Biaffine Layer Label Prediction Loss (Incorporating a biaffine layer to predict label losses.) Furthermore, CAIR-NLP leverages external context retrieval via search engines for input text

- **Large Language Models on IE** Recent strides in NLP have scaled the parametric number of language models to hundreds of billions, yielding exceptional performance from models like GPT-3 [16], OPT-175B [140], Flan-PaLM [25], LLaMA [119], and ChatGPT2. In the domain of Information Extraction (IE), ChatIE [130] harnesses ChatGPT for extraction, demonstrating potential for further improvement. Ongoing experiments explore instruction fine-tuning [129] and simplified training objectives [123] to adapt large language models to extraction tasks.
- **Gazetteer-based Models for NER:** Top-performing teams in the MultCoNER challenge extensively utilized external knowledge bases, such as Wikipedia and gazetteers, to enhance context [128, 22]. Examples include DAMO-NLP’s unified retrieval-augmented system (U-RaNER) [115], which incorporated knowledge from Wikipedia paragraphs and the Wikidata knowledge graph. Other top systems, like PAI [78] and USTC-NELSLIP [77], explored the use of gazetteers to provide additional contextual knowledge for NER. Gazetteer-based models played a crucial role in accurately identifying complex entities, as demonstrated by their success in the MultiCoNER challenge.

NER in Low Resource Settings: Addressing NER in low-resource settings has been a focus of recent research. Cross-lingual knowledge transfer [36], bilingual dictionaries [132], and Bayesian graphical models [101] have been proposed to leverage cross-lingual contextual information. Other approaches include the creation of soft-gazetteers for low-resource languages [104] and unsupervised methods for circumventing label scarcity [13]. Additionally, multilingual transfer learning [100] and distant supervision [46] have been employed to enhance NER performance in low-resource languages. Recent approaches also involve innovative techniques, such as the concatenation of embeddings [125] and co-regularization frameworks [142], contributing to the state-of-the-art in NER tasks.

These diverse approaches collectively highlight the evolving landscape of NER, emphasizing both traditional and innovative strategies to address challenges in different settings and scenarios.

Chapter 3

A Hybrid Model for Cited Text Span Retrieval

3.1 Overview

This chapter aims to identify cited text spans in the reference paper related to the given citance in the citing paper. We refer to it as cited text span retrieval (CTSR). Most current methods attempt this task by relying on pre-trained, off-the-shelf deep learning models like SciBERT. Though these models are pre-trained on large datasets, they underperform in out-of-domain settings. We introduce CitRet, a novel hybrid model for CTSR that leverages unique semantic and syntactic structural characteristics of scientific documents. This enables us to use significantly less data for finetuning. We use only 1040 documents for finetuning. Our model augments mildly-trained SBERT-based contextual embeddings with pre-trained non-contextual Word2Vec embeddings to calculate semantic textual similarity. We demonstrate the performance of our model on the CLSciSumm shared tasks. It improves the state-of-the-art results by over 15% on the F1 score evaluation.

3.2 Introduction

Citations are an integral part of scientific literature as they help better understand the relationships between scientific documents. Authors cite other papers to acknowledge their contributions, compare to their work, criticize, and improve upon their work. Citances often focus on the most important components of a scientific document. Moreover, citance-based summarization is also a widely studied field because it covers some insights that might not be present in abstract-based summarization [34].

However, a citance depends on the intention and opinion of the citing author and can be affected by epistemic value drift¹ [27]. Also, a citance in itself lacks sufficient details to capture the exact content of the referenced paper. Hence, identifying the correct context of the cited text

¹An example of epistemic value drift is citing a claim as a fact.

can enable us to verify the biases [139], overcome epistemic value drift, build dense knowledge graphs, and generate better summaries [54, 20]. Furthermore, it also helps in qualitative analysis of the citations [116]. Motivated by these, research tasks and tracks such as BiomedSumm² and CLSciSumm lay significant emphasis on this fundamental and challenging problem of finding the exact cited text span. We refer to this task as cited text span retrieval (CTSR).

Most of the current methods targeting this problem are centered around fine-tuning deep neural networks. In this regard, transformer [121] based encoders such as BERT [32] and SciBERT [14] have proven to be very effective and have outperformed standard baselines like LDA and TF-IDF. However, a major drawback of these methods is that they require large domain-specific datasets, often exceeding 1 million documents, to fine-tune.

This chapter proposes CitRet, a hybrid CTSR model that performs well even in low-resourced domain-specific settings. We model the problem as a semantic textual similarity (STS) task. We exploit the distinctive semantic and syntactic structural characteristics of scientific literature, i.e., when a paper is cited, the cited text of the reference paper is often paraphrased in such a way that it still expresses the same central idea while also preserving certain keywords. Hence, we use these keywords, which are common to both the citance and the cited sentence, to find weighted contextual embeddings for the sentences. To find these weighted contextual embeddings, we use Sentence-BERT (SBERT) [103] fine-tuned to minimize cosine similarity loss on training data. However, when the training data is scarce, these contextual embeddings fail to capture out-of-domain knowledge. To overcome this, we further leverage pre-trained non-contextual embeddings like Word2Vec [90] to capture the general domain knowledge. We use Word Mover’s Distance (WMD) [60] to find (dis)similarity scores based on these non-contextual embeddings. This hybrid approach of utilizing contextual and non-contextual embeddings enables CitRet to generalize well over unseen datasets. Definitions of the terms used throughout the chapter are:

- **Reference paper (RP):** A scientific document of which one or more sentences have been cited by another paper(s).
- **Citing paper (CP):** A document that contains one or multiple citations to an RP.
- **Citance:** A sentence in CP that contains the reference to the RP.
- **Cited sentence:** The exact piece of the text belonging to the RP that a citance refers to.
- **Cited text span:** Span of the cited sentence(s) belonging to the RP corresponding to a citance.

The major contributions of this work are: 1) Proposing a simple yet effective CTSR model that requires less data for fine-tuning and is computationally inexpensive. We train only

²<http://www.nist.gov/tac/2014/BiomedSumm/>

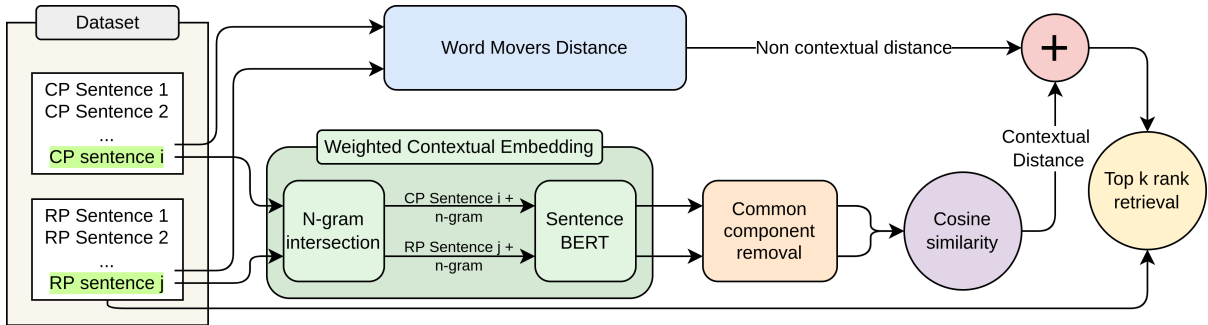


Figure 3.1: Illustration of the CitRet model. WMD and weighted contextual embeddings (WCEs) are calculated for an input pair. The WCEs are then denoised using the common component removal technique. These denoised WCEs are used to find cosine similarity between the sentences of the input pair. Finally, WMD and cosine scores are added, and top k similar sentences in an RP for a citance are retrieved.

on the CL-SciSumm training dataset that consists of 40 manually annotated articles and 1000 automatically annotated articles.

- 2) Advancing the state-of-the-art (SOTA) to identify cited text span by over 15%.
- 3) Empirically validating the advantage of using the semantic and syntactic structure for CTSR.

3.3 Method

We formulate this task of CTSR as finding semantic textual similarity between a citance and all the sentences of an RP, i.e., to find the cited text span for a given citance, we pick the top k similar sentences in the RP. We refer to a <citance, a sentence in the RP> pair as an *input pair*. As shown in Figure 3.1, an input pair is first pre-processed by lowercasing the tokens, removing the stop words, and removing the special characters. Then to find the final similarity scores, CitRet employs a mix of cosine scores using weighted contextual embeddings (contextual distance) and Word Movers Distance scores (non-contextual distance) using pre-trained non-contextual embeddings. Now, we explain each component of the pipeline in detail.

3.3.1 Background

3.3.1.1 SBERT

Sentence-BERT (SBERT) [103], is a modification of the pre-trained BERT (Bidirectional Encoder Representations from Transformers) [32] model. BERT is a popular attention mechanism-based model that takes a sentence (an arbitrary sequence of tokens) as an input and learns con-

textual embeddings for each token in the sentence. Though BERT has achieved state-of-the-art performance in a wide variety of NLP tasks, its design renders it inappropriate for semantic similarity search and unsupervised tasks because BERT doesn't compute independent sentence embeddings and instead learns embeddings for each token of the sentence.

To overcome this problem, SBERT builds over the BERT's innovation of using a bidirectional encoder. SBERT leverages BERT-based siamese network architecture to embed sentences into a fixed-length vector by adding a pooling layer on top of the BERT layer. The SBERT siamese network architecture can be fine-tuned using different losses such as triplet loss, contrastive loss, and cosine similarity loss. Moreover, SBERT is computationally inexpensive compared to BERT [103].

3.3.1.2 Word Mover's Distance

Given pre-trained embeddings for the words, Word Mover's Distance (WMD) [60] measures the distance between a pair of sentences (sequence of words). It exploits the underlying geometry of the word embeddings to represent a sentence as a weighted point cloud in the word embedding space. It formulates the problem of finding the distance between two sentences as a transportation problem based on Earth Movers Distance. It defines the dissimilarity between two sentences as the minimum amount of work (distance traveled) required to transport words from one sentence to the words of another sentence in the word embedding space. This minimum cumulative travel cost between words of two sentences is calculated by solving the following linear optimization problem.

$$\begin{aligned} & \min_{T \geq 0} \sum_{i,j=1}^n T_{ij} c(i,j) \\ \text{subject to: } & \sum_{j=1}^n T_{ij} = s_i \quad \forall i \in \{1, \dots, n\} \\ & \sum_{i=1}^n T_{ij} = s'_j \quad \forall j \in \{1, \dots, n\} \end{aligned}$$

Here, s and s' are the normalized bag-of-words representation of two sentences. T is a flow matrix, where the $T_{ij} \geq 0$ entry indicates how much of word i in sentence s travels to word j in sentence s' . The total outgoing flow from a word i in sentence s to all the words j in sentence s' equals to the normalized frequency of word i , i.e. ($\sum_j T_{ij} = s_i$). The distance between two words in the embedding space is given by $c(i,j)$ and calculated using Euclidean distance between the word embeddings. The final distance between two sentences is $\sum_{i,j} T_{ij} c(i,j)$.

3.3.2 Contextual Distance

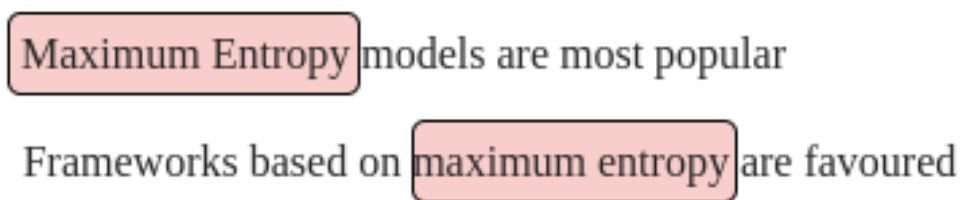
Contextual distance between the sentences is calculated using contextual sentence embeddings. The proposed model uses finetuned SBERT to learn these contextual embeddings for an input pair. SBERT returns a fixed-length dense vector for an input sentence (sentence embedding), irrespective of the length of the input sentence. To yield the final sentence embeddings, CitRet follows three steps: 1) Finetuning the SBERT, 2) Finding the weighted contextual embeddings for each sentence pair, and 3) Denoising the embeddings. We explain each step in detail below.

3.3.2.1 Finetuning the SBERT

To finetune SBERT siamese networks, we use cosine similarity loss. As training examples, we pass sentence pairs annotated with cosine similarity scores on a scale of 0 to 1. For each citance, we pass 5 sentence pairs of 3 different types, i.e., one pair with the actual cited text having a similarity score of 1, two pairs with randomly selected sentences from other RP having a similarity score of 0, and two pairs with randomly selected sentences belonging to the same RP having similarity score of 0.3. This helps us model relations between the sentences of the same documents and sentences of different documents.

3.3.2.2 Weighted contextual embeddings

When an RP is cited, the information that can be extracted from a citance about the RP depends upon the intention, and the opinion of the citing author(s) [139]. However, when the cited sentences are referred to, some key ideas and keywords are preserved, as depicted in Figure 2. CitRet exploits this characteristic of the scientific documents to find weighted contextual embeddings for the input pair.



Maximum Entropy models are most popular
Frameworks based on maximum entropy are favoured

Figure 3.2: n-gram intersection of two sentences

SBERT takes the mean of all the word embeddings to calculate the sentence embedding. After fine-tuning SBERT on domain-specific data, it is able to learn contextual embedding for a sentence. To leverage this contextual learning capability of SBERT and to find weighted contextual embeddings (WCEs) for the sentences, we use a very simple and intuitive strategy

of concatenating the common keywords to the input pair before passing it to the SBERT (we concatenate the keyword to both the sentences of the pair). These keywords are extracted by finding common n-gram intersections between the sentences of the input pair. In the example shown in Figure 3.2, *maximum entropy* is the common keyword (bigram). Concatenating these n-grams results in the common keywords having more weight in the sentence embeddings due to the mean pooling operation. Therefore, the sentence embedding vectors of the pair come closer in the dense vector space if they share some keywords. Here, number n can be optimized empirically, and in our tests, we get the best results for bigrams.

3.3.2.3 Denoising

We further modify the WCEs that we get from the previous step by using a denoising technique adapted from *piecewise common component removal* method proposed in [35]. Here, the common components refer to the common topics (discourse themes) that exist throughout the document (RP and CP) and can be considered as noise. Thus, removing these common components can be understood as downgrading the unimportant components (common discourse) and focusing on the components that have more discriminatory power. This helps in denoising the embeddings [8]. Since cosine-similarity treats all dimensions equally denoising becomes critical in making it more focused. Consequently, the cosine similarity scores calculated using denoised embeddings become more relevant

$$\tilde{v} = v - \sum_i^m \lambda_i \text{proj}_{pc_i} v, \text{ where } \lambda_i = \frac{\sigma_i^2}{\sum_{j=1}^m \sigma_j^2}$$

These common discourse vectors are estimated as the principal components for a set of WCEs. These principal components are calculated by singular value decomposition of $A_{l \times d}$ matrix, where l is the number of sentences in the document (RP and CP), and d is the dimension of the WCEs. To get the final denoised sentence vector \tilde{v} , we subtract from the original sentence vector v , the weighted sum of the projections of the vector v on the first $m(= 3)$ principal components $pc_{i..m}$. The projections $\text{proj}_{pc_i} v$ are weighted by λ_i , where λ_i is the proportion of variance σ_i (singular value) captured by the principal component pc_i .

3.3.3 Non-contextual Distance

CitRet uses both supervised and unsupervised techniques to calculate the final similarity scores to generalize well over unseen datasets. It augments contextual distance calculated using mildly-trained SBERT with non-contextual distance calculated using unsupervised WMD technique Figure 3.3 demonstrates WMD’s ability to capture relations in the general domain setting. The arrows represent the flow between two words of an input pair. It may be observed how *models* flows to *frameworks* and *popular* to *favoured*. It can be noted that the words

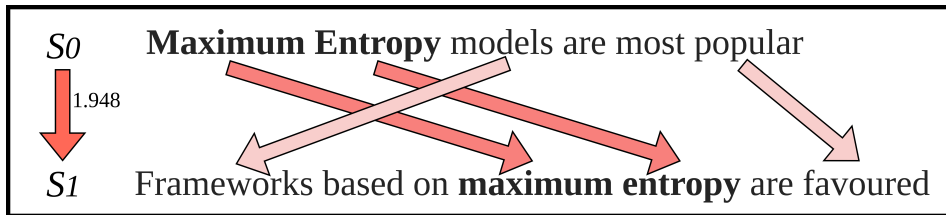


Figure 3.3: Flow between 2 sentences S_0 and S_1 using WMD

popular and *favoured* are general domain words (non-scientific terms) and might not appear very frequently in a scarce domain-specific dataset. Hence, the semantic relationships between these general domain words are better captured by WMD.

3.4 Detailed Experimental Setup and Analysis

We demonstrate the performance of the proposed method on CL-SciSumm shared task [54, 21, 20] task 1(a), where for each citance, we need to identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. These cited text spans are of the granularity of a sentence fragment, a full sentence, or several consecutive sentences (no more than 5). For this, we pick top k (we picked $k = 3$) semantically similar candidate cited sentences for a given citance by sorting their similarity scores. We evaluate the predictions against gold label annotations using F1 score.

We compare the performance of the proposed model CitRet, with the best 3 systems (of each category) submitted by NaCTeM-UoM [138] and the best 2 systems submitted by team NLP-PINGAN-TECH [19], over CL-SciSumm test set.

The systems submitted by NaCTeM-UoM are based on BERT. Along with *BERT 2018/19 OV + 2018 FT* (a BERT model fine-tuned on the CL-SciSumm 2018-2019 dataset), they submitted models *ACL 2018* and *SciBERT 2018*. Both these models are first trained on significantly large domain-specific corpora and then fine-tuned on CL-SciSumm dataset. *ACL 2018* is trained ACL-ARC [99] whereas *SciBERT 2018* is based on SciBERT model [14], which is pre-trained on a collection of 1.14M documents from Semantic Scholar [6].

NLP-PINGAN-TECH team also centered their approach around fine-tuning BERT-based models using larger domain-specific datasets. Their best-performing system *SciBERT-SemBERT* is an ensemble of SciBERT, SemBERT [141] based on SciBERT, *Sci-BERT-fake-token* (tokens for position and section details like $[method][sid=xx][ssid=xx]$ are added as prefixes to the sentences) and *SciBERT-special-token* (tokens for position and section details like $[method],[sid=1]$, etc. are added to the SciBERT dictionary to avoid split during tokenization). The other method *SciBer-ACLBERT*, submitted by the NLP-PINGAN-TECH team that achieved a high score, also leverages SciBERT and ACL corpora.

In comparison, the proposed model is trained only on the CL-SciSumm training dataset that consists of 40 manually annotated articles, which were used in the 2018 CL-SciSumm challenge as well, and 1000 document sets that were automatically annotated using neural networks. These 1000 document sets were introduced in 2019 and are of lower quality compared to the manually annotated dataset. Also, we do not use any external corpora to fine-tune our model.

3.5 Results

We demonstrate the performance of the proposed method on CL-SciSumm shared task [54, 21, 20] task 1(a), where for each citance, we need to identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. These cited text spans range from the granularity of a sentence fragment to several consecutive sentences. We pick top $k = 3$ similar candidate cited sentences for a given citance. CitRet is trained only on the CL-SciSumm training dataset that consists of 40 manually annotated articles and 1000 low-quality document sets that were automatically annotated using neural networks. We do not use any external corpora to fine-tune our model. We evaluate our model’s performance against gold label annotations for the CL-SciSumm test set of 20 documents.

We consider the SOTA models of 2019 and 2020 CL-SciSumm tasks as baselines. Table 3.1 shows that CitRet performs the best in quantitative metrics (F1 and Precision) and outperforms 2019 SOTA (*ACL 2018*) by over 57% and 2020 SOTA (*SciBERT-SemBERT*) by over 15% on F1 score evaluation. It can be noted that using just the *SBERT + WCE* component outperformed all the baseline SOTA models that use much larger datasets (exceeding 1 million) for finetuning³. This empirically validates that using the semantic and syntactic structure for CTSR can significantly improve the results.

Moreover, as evident from the ablation study, individual components of our pipeline also help in increased performance. The most significant improvement, of 30% over fine-tuned SBERT, was achieved by weighted contextual embeddings (*SBERT + WCE*). It can be noted from Table 3.1 that using just *SBERT + WCE* component of our pipeline outperformed all the SOTA models. This empirically validates that utilizing the unique structural characteristics of scientific documents can significantly improve the results. Further denoising the weighted contextual embeddings (*SBERT + WCE + D*) for $m = 3$ improved the performance by around 5%. Moreover, augmenting the contextual embeddings-based similarity scores with WMD achieved a new SOTA by advancing the results of *SBERT + WCE + D* by over 9%.

We also performed experiments to check how the performance of the model varies with the train and test sets’ size. The proposed method showed improvement when we used 1000 document sets that were automatically annotated using neural networks along with the 40

³Please refer to Appendix 3.4 for details of the experimental setup of the baseline models and ablation study analysis.

Method	Recall	Precision	F1
ACL 2018	-	-	0.126
BERT 2018/19 OV+2018FT	-	-	0.120
SciBERT 2018	-	-	0.078
SciBERT-SemBERT	0.2459	0.1318	0.1716
SciBer-ACLBERT	0.2265	0.1244	0.1606
SBERT [†]	0.1879	0.1023	0.1325
SBERT + WCE [†]	0.1815	0.1647	0.1727
Denoising (SBERT+WCE+D) [†]	0.1901	0.1724	0.1808
CitRet (SBERT+WCE+D+WMD) [†]	0.2080	0.1888	0.1979

Table 3.1: Performance comparison of our model with the baseline models. The last four rows show the ablation study of our model marked with †. D denotes the denoising step.

manually annotated documents. We obtained 0.17790 F1 (0.1869 Recall and 0.1697 Precision) when we trained with just the manually annotated dataset that contained only 40 documents. We also experimented with the 1000 noisy training samples by randomly splitting them into the train (80%) and test (20%) sets and obtained 0.2779 F1 (0.4836 Recall and 0.195 Precision).

3.6 Discussion

As can be observed from Table 3.1, the proposed method significantly improves the F1 score (+15%) and Precision(+43%) with some loss in Recall(15%). Our approach focuses on Precision (a measure of the quality of retrieval) over Recall (a measure of quantity) because, for the given task, the probability of getting false positives is very high. Hence, a higher precision results in a more concise and accurate summarization.

The proposed approach is in line with the recommendation made by the task organisers to exploit the structural and semantic characteristics that are unique to scientific documents to enrich the embeddings. The chapter proposes a simple and computationally inexpensive alternative to the current state-of-art model in the form of CitRet. It leverages both contextual and non-contextual embeddings. CitRet also combines a supervised model and an unsupervised model. This hybrid architecture provides performance and robustness against noisy training

samples. The components of the model are lightweight (do not require extensive fine-tuning), faster, explainable, and intuitive. This highlights how other statistical machine learning techniques can be leveraged along with modern deep neural network architectures to compensate for the lack of quality training data and outperform computationally expensive architectures.

It may also be noted that while our method beats the baselines by large margins and achieves a new SOTA, the absolute values are still rather low because of the non-triviality of the task. The task becomes particularly challenging because of the low-quality training data and subjectivity of the annotators. Hence, we believe that there is a scope for further improvement, and the problem demands greater exploration.

3.7 Implementation details

We use the PyTorch framework to implement our NER model. We use the pre-trained SciBERT tokenizer and embeddings as input to a dropout layer with a dropout probability of 0.5 to prevent overfitting. We use a learning rate of 1e-5 and train all models for 10 epochs. We pass the output from the dropout layer through a linear layer with an input dimension the same as the hidden dimension of SciBERT embeddings (768), and an output dimension the same as the number of labels (4). For Sentence-BERT, we use pre-trained models available in Pytorch.

Chapter 4

Transformer Based Architectures for Complex NER in English

4.1 Overview

In this chapter, we investigate the task of complex NER for the English language. The task is non-trivial due to the semantic ambiguity of the textual structure and the rarity of occurrence of such entities in the prevalent literature. Using pre-trained language models, we obtain a competitive performance on this task. We qualitatively analyze the performance of multiple architectures for this task. All our models are able to outperform the baseline by a significant margin. Our best-performing model advances the baseline F1-score by over 9%.

4.2 Introduction

Named Entity Recognition is an Information Extraction task that aims to detect entities from unstructured text and classify them into predefined categories. Although the task of NER has been investigated adequately by previous research work [87, 94, 62, 39, 105], the detection of named entities in a multilingual setting is non-trivial. Furthermore, the introduction of additional layers of complexity - in the form of semantic ambiguity and a lower amount of contextual availability poses further challenges. NER in low-resource languages further enhances the difficulty of such tasks due to the scarcity of available data. Recently, deep learning models have gained popularity for NER [133, 67, 45]. However, these approaches are data-intensive and become ineffective when there is a lack of labeled data. Hence, the NER task for low-resource languages becomes further challenging.

To foster research in this area, the SemEval MultiCoNER challenge [85] was introduced that deals with multiple low-resource language NER with semantically ambiguous entities. In this chapter, we describe our approach to tackle this task using state-of-the-art deep learning models and introduce a simple neural network architecture that builds on top of pre-trained language models. Our approach beats the baseline by a significant margin. We compare multiple architectures on the test and validation set of the shared task. All our models beat

the baseline by a significant margin. We provide the formal task description in Section 4.3, the dataset details in Section 4.5, the method and the model architecture in Section 4.5. We provide details about the experimental implementation in Section 4.6. We discuss the results obtained and error analysis in Sections 4.7 and 4.8, respectively.

4.3 Task Description

The objective of this shared task is to build complex Named Entity Recognition systems for multiple languages such as English, Spanish, Chinese, Hindi, Bangla, etc. The task presents a unique challenge in the form of detecting the entities in semantically ambiguous and low-context settings. Moreover, the shared task also tests the generalization capability and domain adaptability of the proposed systems by testing the system over additional (low-context) data sets containing questions and short search queries, such as Google Search queries.

For this task, the systems had to identify the B-I-O format [102] (short for beginning, inside, outside) tags for six NER-tags classes, namely Person, Product, Location, Group, Corporation, AND Creative Work.

Earlier works have also tried to address the problem of NER, but usually, the datasets consisted of well-formed texts of easy entities [9], and little has been done to tackle the problem of identifying semantically and syntactically ambiguous entities like Creative Works. For example: *Eternal Sunshine of the Spotless Mind* and *Among Us* are complex entities that may be considered as Named Entities in some very selective contexts for e.g. *Among Us* is not a NE in "There is not much disagreement among us," but a CW in " *Among Us* is a super fun game to play." This task also aims at tackling such problems.

4.4 Dataset

The MultiCoNER dataset [82] introduced consists of labelled complex Named Entities. For the monolingual track, the participants have to train a model that works for one language only. For training and validation purposes, train and dev set is provided with labelled entities. The monolingual model trained needs to be used for the prediction of named entities in the test set. The labels from the test set are not provided directly. In this system description for the monolingual track, we have considered the English NER dataset for our task. The dataset follows a BIO tagging scheme and there are 6 entity types in the label space. The statistics for the English dataset in the monolingual track for the train and dev set are provided in Table 4.1 and the description of the label space in Table 4.2.

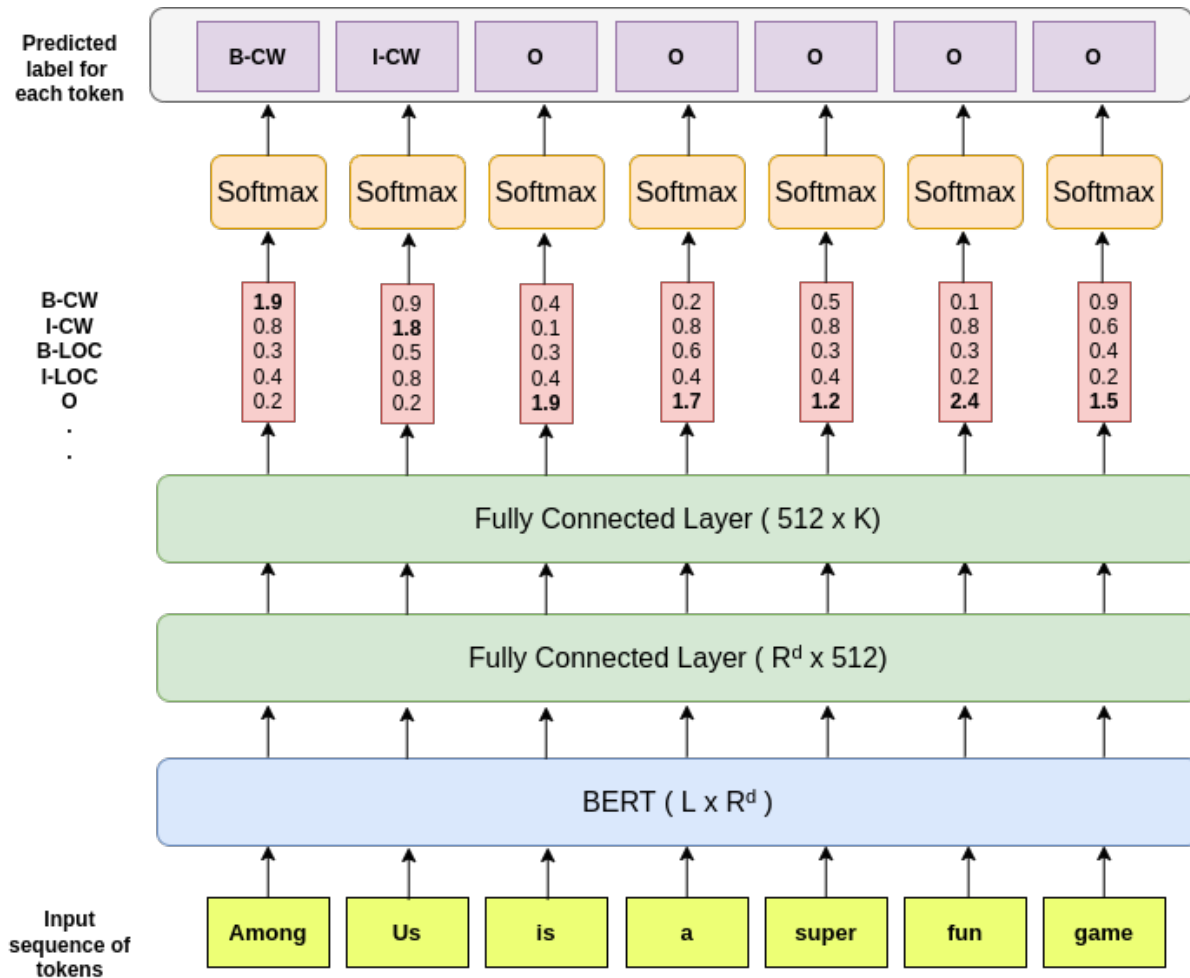


Figure 4.1: BERT-based architecture

4.5 Models

This section describes our approach to designing a system to solve the problem of classifying the tokens of a given sentence into one of the six NE categories. We also briefly describe the BERT model architecture employed in our system.

As is the case with most of the NLP tasks, the performance of the model boils down to learning the best-distributed representation for the tokens. With the advent of transformer-based models, the whole domain of NLP has been revolutionized because they provide us with some of the most feature-rich embeddings. Contextual embeddings learned using transformer-based models give better performance than embeddings learned using traditional methods such as TF-IDF, word2vec, etc., for downstream tasks such as NER, since such tasks require greater contextual awareness.

We also adopted the simple strategy of finetuning various architectures based on pre-trained language models such as Bert on our task-specific data.

BERT+CRF : We use a pre-trained BERT model to obtain the token embeddings. These embeddings are passed to a token-level classifier followed by a Linear-Chain CRF. The CRF produces a probability distribution over the entire label space for each token among the sequence of tokens. More formally: 1) For a sequence of tokens $x = (x_1, x_2, x_3, \dots, x_m)$, where x_i is the i th token among the sequence of tokens, we obtain a low-dimensional dense embedding, $x_i \in R^d$ where d is the embedding dimension. 2) This embedding is mapped to a lower dimensional space $x_i \in R^k$ where k is the total number of labels. 3) The output scores from the linear layer are obtained as $P \in R^{m \times k}$, where m is the number of tokens. These scores are passed to the CRF layer, whose parameters are $A \in R^{k+2 \times k+2}$. Each element A_{ij} signifies the transition score from the i th label to the j th label. The 2 additional states in A are the start and the end state of a sequence. For a series of tokens $x = (x_1, x_2, x_3, \dots, x_m)$ we obtain a series of predictions $y = (y_1, y_2, y_3, \dots, y_m)$. As described in [63], the score of the entire sequence is defined as :

$$s(x, y) = \sum_{i=0}^m A_{y_i, y_{i+1}} + \sum_{i=1}^m P_{i, y_i}$$

The model is trained to maximize the log probability of the correct label sequence:

$$\log(p(y|x)) = s(x, y) - \log \left(\sum_{\tilde{y} \in Y_X} e^{s(x, \tilde{y})} \right)$$

where Y_X are all possible label sequences.

BERT+BiLSTM+CRF : We use a pre-trained BERT model to obtain the contextual embeddings from the sentences. These embeddings are passed to the BiLSTM layer. The BiLSTM layer captures these into a hidden state representation. This representation is passed to a CRF layer that obtains the probability distributions across the labels. Specifically, the pre-trained language model is used to map the tokens in each sentence to a distributed representation. This is used as the word embedding layer of the BiLSTM-CRF model. The BiLSTM-CRF layer is used to sequence label the sentence, and the predicted labels are obtained. The supervised learning algorithm iterates to improve its predicted label accuracy over every iteration. More formally, the process can be described as follows : 1) The target sentence comprising of m tokens, is represented as $x = (x_1, x_2, x_3, \dots, x_m)$, where x_i represents the i th token of the entire target sentence. 2) x_i is mapped to a low dimensional dense vector, $x_i \in R^d$ using the pre-trained BERT embeddings, where d is the dimension of dense embedding. 3) The sequence of tokens x is taken as an input to the BiLSTM in each time step, and the forward hidden states $\vec{h}_f = (\vec{h}_1, \vec{h}_2, \vec{h}_3, \dots, \vec{h}_m)$ and the backward hidden states $\overleftarrow{h}_b = (\overleftarrow{h}_1, \overleftarrow{h}_2, \overleftarrow{h}_3, \dots, \overleftarrow{h}_m)$ are concatenated to form the combined hidden state representation $h = [\vec{h}_f, \overleftarrow{h}_b]$. 4) The combined hidden state representation $h \in R^{m \times n}$ is reduced to a k dimensions using a linear layer, where k is

	Train	Dev
# sentences	15300	800

Table 4.1: Total sentences in English monolingual track

Label	Description
PER	Person
LOC	Location
GRP	Group
CORP	Corporation
PROD	Product
CW	Creative Work

Table 4.2: Entity types in the label space

Class Label	BERT+Linear			BERT+CRF			BERT+BiLSTM-CRF		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
LOC	0.9304	0.9145	0.9224	0.903	0.9145	0.9087	0.9025	0.9103	0.9064
PER	0.9659	0.9759	0.9708	0.936	0.9586	0.9472	0.8882	0.9586	0.9221
PROD	0.7365	0.8367	0.7834	0.7785	0.7891	0.7838	0.7372	0.7823	0.7591
GRP	0.8923	0.9158	0.9039	0.8341	0.9000	0.8658	0.8466	0.8421	0.8443
CW	0.7955	0.7955	0.7955	0.7963	0.733	0.7633	0.7353	0.7102	0.7225
CORP	0.893	0.8653	0.8789	0.8877	0.8601	0.8737	0.8837	0.7876	0.8329
Average	0.8689	0.8839	0.8758	0.8559	0.8592	0.8571	0.8322	0.8318	0.8312

Table 4.3: Results of our models on validation dataset

the number of labels to distribute the probabilities across. 4) Finally, the CRF layer is used to obtain the probability distribution across all the labels to obtain the final prediction.

BERT+Linear: The token sequence is mapped to a lower dimensional space using pre-trained BERT embeddings. These embeddings are then passed to a linear layer that maps these embeddings to a lower dimension of label space. The output scores are then softmaxed to provide a probability distribution across all labels.

4.6 Implementation Details

We implement all our transformer-based models using Pytorch and Huggingface libraries. We implemented 3 models, BERT+Linear, BERT+CRF and BERT+BILSTM+CRF. We also tried adding POS embeddings as extra features to the models and compared the results. We use a dropout from 0.2 to 0.5 in all models, and found that 0.3 gave the best results throughout. We used 2 linear layers in the BERT+Linear model. We added a softmax layer to obtain the probability distribution across all the labels. For the BERT+Linear model, we run our experiments across 1-20 epochs. We find that the model starts to overfit after 10 epochs, and the best results are obtained after 5 epochs of training. For BERT+CRF, we experiment across 1-100 epochs. We find the model gives the most optimal result at the 20th epoch, after which it starts to overfit. We use a learning rate of $1e^{-4}$ for all the models. WE validate the results of all models using our dev set.

4.7 Results

We compare the performance of our models in the validation set against the baseline. We use the best-performing model for the final submission in the evaluation phase. We provide details of the performance of the best performing over the blind test dataset provided in the evaluation phase. We provide a detailed comparison of the performance of our models across all the class labels in the validation dataset in Table 4.3. We observe that the simple BERT+Linear model performs the best as compared to other larger models. We attribute this to the limited number of samples in the training dataset. The lack of a sufficient number of training samples limits the ability of larger models to generalize properly over the entire training set. We notice that the performance of the BERT+Linear model is consistent across all class labels except for PRODUCT.

	Precision	Recall	F1-Score
Baseline System	0.773	0.780	0.776
BERT + CRF	0.855	0.859	0.857
BERT+BILSTM-CRF	0.832	0.831	0.831
BERT + Linear	0.868	0.883	0.875

Table 4.4: Comparison of model performances with the baseline on the validation dataset

Class Label	BERT+Linear		
	Precision	Recall	F1-Score
LOC	0.7292	0.7614	0.7449
PER	0.8776	0.8922	0.8848
PROD	0.7079	0.6460	0.6755
GRP	0.7699	0.6600	0.7107
CW	0.5527	0.6299	0.5888
CORP	0.7253	0.6759	0.6998
Average	0.7271	0.7109	0.7174

Table 4.5: Performance of model on test dataset

4.8 Error Analysis

We perform error analysis for all 3 different model performances on the validation dataset. We find that for all 3 models, each model has the highest difficulty in accurately predicting the *CW* (*Creating Work*) label. This can be attributed to the higher degree of ambiguity when it comes to *CW* named entities, as these often share similar type of textual structure as normal non-named entity text tokens. It can be inferred that all 3 models are memorizing entity names from the training data to some extent. It is most prevalent in **BERT+BiLSTM+CRF** model, as we can see that it has the least amount of prediction accuracy among other models. This is consistent with our reasoning that heavier models tend to overfit the dataset faster. Hence, we deduce that named entity memorization can be attributed to a type of overfitting behavior by the model in question. The **BERT+Linear** model, which is the lightest model with the least amount of trainable parameters among all 3, is found to be significantly less prone to memorize entity names.

We perform error analysis for all 3 different model performances on the validation dataset. We find that for all 3 models, each model has the highest difficulty in accurately predicting the *CW* (*Creating Work*) label. This can be attributed to the higher degree of ambiguity when it comes to *CW* named entities, as these often share a similar type of textual structure as normal non-named entity text tokens. It can be inferred that all 3 models are memorizing entity names from the training data to some extent. It is most prevalent in **BERT+BiLSTM+CRF** model, as we can see that it has the least amount of prediction accuracy among other models. This is consistent with our reasoning that heavier models tend to overfit the dataset faster. Hence, we deduce that named entity memorization can be attributed to a type of overfitting behavior by the model in question. The **BERT+Linear** model, which is the lightest model with the least

amount of trainable parameters among all 3, is found to be significantly less prone to memorize entity names.

Furthermore, upon qualitative analysis, we found that our models often have difficulty in recognizing longer-named entities (entities comprising 5 or more tokens). This can be attributed to the lack of such entities in the training dataset. The models are majorly exposed to a shorter set of entity spans and texts that occur out of the BIO tag and are non-named entities. Due to the lack of exposure of the models to adequate training instances of longer spans, the models are often unable to predict such longer entity spans.

Chapter 5

Complex NER in Semantically Ambiguous Settings for Low Resource Languages

5.1 Overview

In this chapter, we leverage pre-trained language models to solve the task of complex NER on 2 low-resource languages- Chinese and Spanish. We use the technique of Whole Word Masking (WWM) to boost the performance of Masked Language Modeling objective on large, unsupervised corpora. We experiment with multiple neural network architectures, incorporating CRF, BiLSTMs and Linear Classifiers on top of pre-trained BERT embeddings. All our models outperform the baseline by a significant margin and our best performing model obtains a competitive position on the evaluation leaderboard for the blind test set. We hope this work facilitates further research in the challenging domain of ambiguous, low-resource, complex NER.

5.2 Introduction

We investigate the task of complex, semantically ambiguous, and low-resource NER [85]. The most popular NER task in the English language is CoNLL [11], which is widely used as a benchmark for most NER models. Multiple models have been able to obtain sufficiently high performances in this task setting [126, 142, 75, 110, 136, 134, 127]. The CoNLL training set consists of 14,987 train sentences comprising 203,621 tokens for English data. The entity space consists of 4 different types of entity type labels (locations, persons, organisations and miscellaneous) to classify each named token. The English data was taken from the Reuters Corpus, which comprises of Reuters News Stories for 1 year. The training data source, and by extension, the labelled named entities comprises of majorly popular entities found in the general English textual content prevalent in the media. Hence, these entities were easier to classify into the correct tokens due to the large prevalence of training data. With the use of pre-trained transformer-based language models, which are already trained on a large unlabelled corpus of

English text, this task became even less challenging, as the nature of textual structure in these corpora largely overlaps with that of CoNLL.

However, there is a multitude of varieties of named entities possible, ones that comprise of complex, ambiguous textual structural content. Such named entities are harder, in general, to predict for language models due to the semantically ambiguous nature of the textual structure of the named entities and the lower amount of occurrence of such entity types in general English text. The shared task of MultiCoNER (which stands for multilingual, complex NER) adds additional challenges by introducing rarer label types (like creative work, products, etc.).

Another way to increase the difficulty of NER tasks is to perform them in low-resource languages. There is a significant dearth of both labeled and unlabelled data for such languages. The complexity is further enhanced by using rarer entity types in such languages. Combined with a lack of unlabelled data, the lack of occurrence of rarer entity token types becomes even harder for the fine-tuned language models to overcome. The shared task of MultiCoNER introduces datasets in multiple low-resource languages.

We leverage large pre-trained language models trained in low-resource language corpora to obtain competitive performances in the low-resource, complex NER setting. We show that simpler architectures successfully outperform other heavier counterparts. We use standard BERT-CRF based models to obtain high performances in the evaluation set. We experiment on two low-resource dataset: Spanish and Chinese.

Our approach beats the baseline by a significant margin. We compare multiple architectures on the test and validation set of the shared task. All our models beat the baseline by a significant margin. We provide the formal task description in Section 5.3, the dataset details in Section 5.5, the method and the model architecture in Section 5.5. We provide details about the experimental implementation in Section 5.6. We discuss the results obtained and error analysis in Sections 5.7 and 5.8, respectively.

5.3 Task Description

The objective of this shared task is to build complex Named Entity Recognition systems for multiple languages such as English, Spanish, Chinese, Hindi, Bangla, etc. The task presents a unique challenge in the form of detecting the entities in semantically ambiguous and low-context settings. Moreover, the shared task also tests the generalization capability and domain adaptability of the proposed systems by testing the system over additional (low-context) data sets containing questions and short search queries, such as Google Search queries.

For this task, the systems had to identify the B-I-O format [102] (short for beginning, inside, outside) tags for six NER-tags classes, namely Person, Product, Location, Group, Corporation, and Creative Work.

Earlier works have also tried to address the problem of NER, but usually, the datasets consisted of well-formed texts of easy entities [9], and little has been done to tackle the problem of identifying semantically and syntactically ambiguous entities like Creative Works. For example: *Eternal Sunshine of the Spotless Mind* and *Among Us* are complex entities that may be considered as Named Entities in some very selective contexts for eg. *Among Us* is not a NE in "There is not much disagreement among us", but a CW in "*Among Us* is a super fun game to play". This task also aims at tackling such problems.

5.4 Dataset

The MultiCoNER dataset [82] consists of multiple low-resource languages. We consider Chinese and Spanish languages in this work. For the monolingual track, the participants must train a model that only works for one language. We train the language model on the train set to obtain predictions for dev and test sets. The labels from the blind test set are not provided directly. The dataset follows a BIO tagging scheme with 6 entity types in the label space. The statistics for the Chinese and Spanish datasets in the monolingual track for the train and dev set are provided in Table 5.1 and the description of the label space in Table 5.2.

5.5 System Overview

At first, we pre-train the BERT language model on unlabelled corpora for the target low-resource language. For Chinese, we use the strategy outlined by [28]. BERT uses the WordPiece tokenizer [131] to split tokens into smaller fragments. It is easier for the Masked Language Model to predict these masked fragments. However, for the Chinese textual texture, the Chinese characters are not formed by alphabet-like symbols, so the WordPiece tokenizer is unable to split the words into small fragments. Hence, we use the Chinese Word Segmentation (CWS) tool to split the text into separate words and then use the Whole Word Masking strategy for the Masked Language Model objective. This removes the drawback of masking small fragments, making it harder for the model to predict whole masked words.

For the Spanish variant, we adopt the strategy outlined by [18]. Similar to [28], they use the strategy of whole word masking for pre-training the BERT language model on unlabelled Spanish corpus.

We adopt the strategy of finetuning these pre-trained BERT models on the downstream NER task for each language, respectively.

BERT+CRF: We obtain token-level dense representations using BERT-based pretrained embeddings. We pass these embeddings to the CRF layer to obtain the probability distribution across the label space. For a sequence of tokens $x = (x_1, x_2, x_3, \dots, x_n)$, we obtain the i th token representation x_i of dimension d , which is the dimension of the dense vector representations

of the BERT-based embeddings obtained from the pre-trained language model. The token embedding x_i is passed to a dense linear layer to transform the representation from d to k dimensional space, where k is the number of labels. The output scores, obtained from the linear layer as $P \in R^{m \times k}$, are passed to the CRF layer whose parameters are $A \in R^{k+2 \times k+2}$. Element A_{ij} denotes the transition score from the i th to the j th label. 2 additional states are added to the start and end of the sequence. For a series of tokens $x = (x_1, x_2, x_3, \dots, x_n)$ we obtain a series of predictions $y = (y_1, y_2, y_3, \dots, y_n)$.

As described in [63], the score of the entire sequence is defined as :

$$s(x, y) = \sum_{i=0}^m A_{y_i, y_{i+1}} + \sum_{i=1}^m P_{i, y_i}$$

The model is trained to maximize the log probability of the correct label sequence:

$$\log(p(y|x)) = s(x, y) - \log \left(\sum_{\tilde{y} \in Y_X} e^{s(x, \tilde{y})} \right)$$

where Y_X are all possible label sequences.

BERT+BiLSTM+CRF : We obtain token-level contextual dense representations using BERT-based pre-trained embeddings. These embeddings are passed to a BiLSTM layer, which obtains the hidden-state representation of these tokens. We pass these hidden states to the CRF layer to obtain the probability distribution across the label space. We use the pre-trained language model to map the tokens in each sentence to a dense embedding representation. The BERT-based dense embeddings are passed to the BiLSTM-CRF layer, which is used to obtain the predicted labels for each token in the entire sequence. More formally, For a sequence of tokens $x = (x_1, x_2, x_3, \dots, x_n)$, we obtain the i th token representation x_i of dimension d , which is the dimension of the dense vector representations of the BERT-based embeddings obtained from the pre-trained language model. The token embedding x_i is passed to a dense linear layer to transform the representation from d to k dimensional space, where k is the number of labels. The sequence of tokens x is taken as an input to the BiLSTM in each time step, and the forward hidden states $\vec{h}_f = (\vec{h}_1, \vec{h}_2, \vec{h}_3, \dots, \vec{h}_n)$ and the backward hidden states $\overleftarrow{h}_b = (\overleftarrow{h}_1, \overleftarrow{h}_2, \overleftarrow{h}_3, \dots, \overleftarrow{h}_n)$ are concatenated to form the combined hidden state representation $h = [\vec{h}_f, \overleftarrow{h}_b]$. The combined hidden state representation $h \in R^{m \times n}$ is transformed to a k dimensional space using a linear layer, where k is the number of labels to distribute the probabilities across. Finally, the CRF layer outputs the probability distribution for each token across the label space.

BERT+Linear: The token sequence is mapped to a lower dimensional space using pre-trained BERT embeddings. These embeddings are then passed to a linear layer that maps these embeddings to a lower dimension of label space. The output scores are then softmaxed to provide a probability distribution across all labels.

	Train	Dev
# sentences	15300	800

Table 5.1: Total sentences in Chinese and Spanish monolingual track

Label	Description
PER	Person
LOC	Location
GRP	Group
CORP	Corporation
PROD	Product
CW	Creative Work

Table 5.2: Entity Types in the label space

Class Label	BERT+CRF			BERT+Linear			BERT+BiLSTM-CRF		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
LOC	0.8368	0.8796	0.8577	0.8194	0.8613	0.8399	0.8219	0.8759	0.8481
PER	0.9065	0.9028	0.9047	0.8933	0.9150	0.9040	0.9177	0.9028	0.9102
PROD	0.6970	0.7468	0.7210	0.6864	0.7532	0.7183	0.7278	0.7468	0.7372
GRP	0.7952	0.7857	0.7904	0.8061	0.7917	0.7988	0.7751	0.7798	0.7774
CW	0.7965	0.7135	0.7527	0.8107	0.7135	0.7590	0.7654	0.7135	0.7385
CORP	0.8657	0.8227	0.8436	0.8529	0.8227	0.8375	0.8397	0.7801	0.8088
Average	0.8163	0.8085	0.8117	0.8115	0.8096	0.8096	0.8079	0.7998	0.8034

Table 5.3: Results of our models on validation dataset for the Spanish language.

Class Label	BERT+CRF			BERT+Linear			BERT+BiLSTM-CRF		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
LOC	0.9239	0.9312	0.9275	0.9186	0.9259	0.9223	0.9465	0.9365	0.9415
PER	0.8971	0.9457	0.9208	0.8955	0.9302	0.9125	0.8497	0.9225	0.9084
PROD	0.8662	0.8504	0.8582	0.8593	0.8248	0.8417	0.8867	0.8285	0.8566
GRP	0.7727	0.6538	0.7083	0.6923	0.6923	0.6923	0.7500	0.6923	0.7200
CW	0.8556	0.8191	0.8370	0.8370	0.8191	0.8280	0.8265	0.8617	0.8437
CORP	0.8808	0.8854	0.8831	0.8883	0.8698	0.8789	0.8615	0.8750	0.8682
Average	0.8660	0.8476	0.8558	0.8485	0.8437	0.846	0.8610	0.8527	0.8564

Table 5.4: Results of our models on validation dataset for the Chinese language.

5.6 Implementation Details

We implement all our transformer-based models using Pytorch and Huggingface libraries. The Chinese language model with the Whole Word Masking (WWM) objective is trained on the Chinese Wikipedia unlabelled text corpus. We use the same training corpus as [18] to pre-train the BERT language model on Spanish data. We implement 3 models: BERT-BiLSTM-CRF, BERT+Linear, and BERT+CRF for our low-resource NER task setting. We run our experiments between 1-100 epochs. We find that the best results are obtained after 10 epochs of training for each model, after which the model starts to overfit. We use a cyclic learning rate between $1e^{-4}$ to $1e^{-6}$. We use a dropout from 0.2 to 0.5 for all models. We validate the results of all models using our validation dataset.

5.7 Results

We compare the performances of all models in the low-resource language setting for both languages. We observe that the **BERT+CRF** model performs the best across both languages. We choose the best-performing model to evaluate our results on the blind test set. Our approach beats the baseline by a significant margin and outperforms multiple models in the competition. We provide detailed comparisons of all 3 models in Tables 5.3 and 5.4 for Spanish and Chinese languages, respectively. We also compare the results between the baseline and our models for the validation dataset in Tables 5.5 and 5.6.

We observe the **BERT+CRF** model beats **BERT+Linear** by a slender margin. This can be attributed to the addition of the CRF layer, which has been popularly used for sequence labeling tasks by various neural architectures. The **BERT+BiLSTM+CRF** model is much

Class Label	BERT+CRF		
	Prec	Rec	F1
LOC	0.5768	0.6571	0.6144
PER	0.7641	0.7739	0.7690
PROD	0.6292	0.5141	0.5659
GRP	0.5727	0.5560	0.5642
CW	0.5331	0.5257	0.5294
CORP	0.6605	0.6005	0.6291
Average	0.6227	0.6046	0.6120

Table 5.5: Performance of the Spanish model on the test dataset.

heavier with a larger number of parameters and overfits the training dataset due to the smaller number of training instances.

5.8 Error Analysis

We perform error analysis on all 3 different models. We qualitatively analyze the predictions on the validation dataset for both languages. As the final evaluation test set in blind, we are unable to perform analysis on the same.

We find that the labels GRP (Group), PROD (Product), and CW (Creative Work) are the most inaccurately predicted labels for the Spanish models. This conforms to our hypothesis that the long-tailed nature of these entities (which means the frequency of occurrence of such entity types in the general literature of the target language is rare). Hence, the model has the most difficulty in recognising these entities from the contextual sentences. The other label types are more common and were present in the CoNLL dataset as well. We also notice that the **BERT+Linear** does marginally better than **BERT+CRF** on predicting such labels, despite it not being the best performing model overall. This can be attributed to it being a lighter model, imparting it the capability of generalising better while training on a relatively lower amount of training instances. The other 2 models have a larger number of parameters, leading them to overfit due to label scarcity. For the Chinese language as well, we notice a similar phenomenon for the GRP label.

We perform error analysis for all 3 different model performances on the validation dataset. We find that for all 3 models, each model has the highest difficulty in accurately predicting the *CW (Creating Work)* label. This can be attributed to the higher degree of ambiguity when it comes to *CW* named entities, as these often share a similar type of textual structure as normal

Class Label	BERT+CRF		
	Prec	Rec	F1
LOC	0.6930	0.7955	0.7407
PER	0.7952	0.6377	0.7078
PROD	0.6853	0.7232	0.7038
GRP	0.7254	0.4608	0.5636
CW	0.5520	0.6798	0.6093
CORP	0.6526	0.7361	0.6918
Average	0.6839	0.6722	0.6695

Table 5.6: Performance of the Chinese model on the test dataset.

non-named entity text tokens. It can be inferred that all 3 models are memorizing entity names from the training data to some extent. It is most prevalent in **BERT+BiLSTM+CRF** model, as we can see that it has the least amount of prediction accuracy among other models. This is consistent with our reasoning that heavier models tend to overfit the dataset faster. Hence, we deduce that named entity memorization can be attributed to a type of overfitting behavior by the model in question. The **BERT+Linear** model, which is the lightest model with the least amount of trainable parameters among all 3, is found to be significantly less prone to memorizing entity names.

Furthermore, upon qualitative analysis, we found that our models often have difficulty in recognizing longer-named entities (entities comprising 5 or more tokens). This can be attributed to the lack of such entities in the training dataset. The models are majorly exposed to a shorter set of entity spans and texts that occur out of the BIO tag and are non-named entities. Due to the lack of exposure of the models to adequate training instances of longer spans, the models are often unable to predict such longer entity spans.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

- In Chapter 3, we have introduced CitRet, a novel model for cited text span retrieval. CitRet outperforms the current SOTA models by significant margins (15% F1). The proposed model is quite simple, computationally inexpensive, improves generalization, and does not require any large external datasets to fine-tune. However, considering the non-triviality of the task, we propose a new approach for further exploration of the task.
- In Chapter 4, we experimented with 3 model architectures for a novel dataset introduced for the shared task of detecting complex NER. Our best-performing model comprises of a simple linear classifier on top of BERT based pretrained language model. We find that this simple approach performs competitively as compared to its heavier counterparts. It also beats numerous teams in the performance in the final evaluation dataset. Upon analysis, we attribute this observation to the scarcity of labeled training data. We find this simpler approach to give a higher performance as it can utilize the contextual information from a sequence of tokens to accurately predict the named entity tokens. It can optimally avoid overfitting to a more significant extent and hence performs better than other heavier models.
- In Chapter 5, we have introduced vital improvements over the baseline for the shared task of complex NER for low-resource languages. We leverage the Whole Word Masking objective to perform better in this low-resource setting. We perform extensive experiments and find that simple BERT-CRF-based models perform strongly against other heavier models even in such low resource semantically ambiguous settings as evidenced by the final evaluation rankings. We also conduct qualitative error analysis and describe our findings.

6.2 Future Work

In addition to advancing the current research directions, our future work plans involve exploring the integration of Cited Text Span Retrieval (CTSR) and Named Entity Recognition (NER) into broader applications, such as question-answering, summarization, and retrieval-augmented generation. These extensions aim to leverage the foundational capabilities developed in CTSR and NER to enhance diverse aspects of information retrieval and comprehension.

- **Addressing Label Scarcity in Low-Resource Languages:** To overcome label scarcity challenges in low-resource languages, we plan to leverage data augmentation and distant supervision techniques. Our future work will focus on utilizing additional data augmentation methods and distant supervision to create clean silver labels, thereby increasing training instances. We believe that this approach will enable us to leverage larger models for training, enhancing the performance of our systems.
- **Question Answering (QA) Systems:** Integrating CTSR and NER into QA systems holds the potential to enhance the accuracy and depth of responses. By efficiently identifying relevant cited text spans and recognizing named entities within these spans, QA models can provide more precise and contextually rich answers to user queries.
- **Summarization Techniques:** Our future work plans include exploring how CTSR and NER can contribute to the improvement of summarization techniques. Extracting key information from cited text spans and accurately recognizing named entities can aid in generating concise and informative summaries of scientific documents.
- **Retrieval-Augmented Generation:** Extending our models into retrieval-augmented generation frameworks can enhance the generation of informative content. By incorporating the context retrieved through CTSR and enriching it with identified named entities, the generated content can be more contextually relevant and coherent.
- **Knowledge Graph Integration:** In our future work, we aim to explore the construction and maintenance of knowledge graphs from extracted entities and their relationships. This structured representation of scientific knowledge could facilitate more advanced analysis and visualization.
- **Evaluation in Real-World Scenarios:** Our future work plans involve deploying and evaluating CTSR and NER in real-world scenarios. Collaborating with researchers and institutions will help us understand the practical impact of our models on their literature review processes.
- **Interdisciplinary Applications:** Exploring the application of CTSR and NER beyond traditional academic literature to interdisciplinary domains is an exciting avenue for future

research. Adapting the models to diverse fields and knowledge contexts will broaden the impact of our work.

- **Ethical Implications and User Feedback:** Continual attention to ethical considerations and gathering user feedback will be integral to our future work plans. This involves addressing potential biases, ensuring model transparency, and refining the models based on insights from users in the academic community.

By integrating CTSR and NER into these broader applications, we plan to provide researchers with powerful tools that not only facilitate literature review but also enhance their ability to engage with and extract knowledge from scholarly documents efficiently. For future work, we aim to leverage data augmentation and distant supervision techniques to circumvent the label scarcity problem in low-resource languages. Additionally, we plan to explore other data augmentation techniques and distant supervision to create clean silver labels, believing that this would enable us to leverage larger models for training purposes.

Related Publications

Publications related to thesis

1. *Amit Pandey*¹, Avani Gupta¹, and Vikram Pudi. 2022. **CitRet: A Hybrid Model for Cited Text Span Retrieval**. In Proceedings of the 29th International Conference on Computational Linguistics, pages 45284536, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
2. *Amit Pandey*², Swayatta Daw², Narendra Unnam, Vikram Pudi **Complex NER in Semantically Ambiguous Settings for Low Resource Languages** In Proceedings of 16th SemEval at the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics
3. *Amit Pandey*², Swayatta Daw², Vikram Pudi **Transformer Based Architecture for Complex NER** In Proceedings of 16th SemEval at the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics

Other publications

1. Narendra Babu Unnam, P.Krishna Reddy, *Amit Pandey*, Naresh Manwani; **Journey to the center of the words: Word weighting scheme based on the geometry of word embeddings**; 34th International Conference on Scientific and Statistical Database Management (SSDBM 2022); July 6-8, 2022 Copenhagen, Denmark.
2. Narendra BabuUnnam, Krishna Reddy Polepalli, *Amit Pandey* and Naresh Manwani; **Text Representation Models based on the Spatial Distributional Properties of Word Embeddings**; Proceedings of the 7th International Conference on Data Science and Management of Data, 11th CoDS and 29th COMAD 2024.

¹Equal Contribution

²Equal Contribution

Bibliography

- [1] Knowledge synthesis: Systematic & scoping reviews. url: <https://guides.lib.uwo.ca/knowledgesynthesis>.
- [2] Sciassist. url: <https://github.com/WING-NUS/SciAssist>.
- [3] The systematic review process. url: <https://guides.hsict.library.utoronto.ca/c.php?g=430254&p=5018365>.
- [4] A. G. T. AbuRa'ed, À. Bravo Serrano, L. Chiruzzo, and H. Saggion. Lastus/taln+ inco@ cl-scisumm 2018-using regression and convolutions for cross-document semantic linking and summarization of scholarly literature. In *Mayr P, Chandrasekaran MK, Jaidka K, editors. BIRNDL 2018. 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries; 2018 Jul 21; Ann Arbor, MI.[place unknown]: CEUR; 2018. p. 150-63. CEUR Workshop Proceedings, 2018.*
- [5] P. Aggarwal and R. Sharma. Lexical and syntactic cues to identify reference scope of citance. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)*, pages 103–112, 2016.
- [6] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, R. Kinney, S. Kohlmeier, K. Lo, T. Murray, H.-H. Ooi, M. Peters, J. Power, S. Skjonsberg, L. L. Wang, C. Wilhelm, Z. Yuan, M. van Zuylen, and O. Etzioni. Construction of the literature graph in semantic scholar, 2018.
- [7] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of international conference on learning representations (ICLR)*, 2017.
- [8] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*, 2017.
- [9] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*, 2017.
- [10] D. Aumiller, S. Almasian, P. Hausner, and M. Gertz. UniHD@CL-SciSumm 2020: Citation extraction as search. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 261–269, Online, Nov. 2020. Association for Computational Linguistics.

- [11] A. Baeovski, S. Edunov, Y. Liu, L. Zettlemoyer, and M. Auli. Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [12] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.
- [13] M. S. Bari, S. R. Joty, and P. Jwalapuram. Zero-resource cross-lingual named entity recognition. *CoRR*, abs/1911.09812, 2019.
- [14] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.
- [15] À. Bravo, L. Chiruzzo, H. Saggion, et al. Lastus/taln+ inco@ cl-scisumm 2018-using regression and convolutions for cross-document semantic linking and summarization of scholarly literature. In *BIRNDL@ SIGIR*, 2018.
- [16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [17] Z. Cao, W. Li, and D. Wu. Polyu at cl-scisumm 2016. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)*, pages 132–138, 2016.
- [18] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.
- [19] L. Chai, G. Fu, and Y. Ni. Nlp-pingan-tech@ cl-scisumm 2020. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 235–241, 2020.
- [20] M. K. Chandrasekaran, G. Feigenblat, E. Hovy, A. Ravichander, M. Shmueli-Scheuer, and A. de Waard. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online, Nov. 2020. Association for Computational Linguistics.
- [21] M. K. Chandrasekaran, M. Yasunaga, D. Radev, D. Freitag, and M.-Y. Kan. Overview and results: Cl-scisumm shared task 2019. *arXiv preprint arXiv:1907.09854*, 2019.
- [22] B. Chen, J.-Y. Ma, J. Qi, W. Guo, Z.-H. Ling, and Q. Liu. Ustc-nelslip at semeval-2022 task 11: gazetteer-adapted integration network for multilingual complex named entity recognition. *arXiv preprint arXiv:2203.03216*, 2022.
- [23] Q. Chen, Z.-H. Ling, and X. Zhu. Enhancing sentence embedding with generalized pooling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1815–1826, 2018.

- [24] H. L. Chieu and H. Ng. Named entity recognition with a maximum entropy approach. In *CoNLL*, 2003.
- [25] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models, 2022.
- [26] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv:2003.10555*, 2020.
- [27] A. Cohan, L. Soldaini, and N. Goharian. Matching citation text and cited spans in biomedical literature: a search-oriented approach. In *proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 1042–1048, 2015.
- [28] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, Nov. 2020. Association for Computational Linguistics.
- [29] E. Davoodi, K. Madan, and J. Gu. Clscisumm shared task: On the contribution of similarity measure and natural language processing features for citing problem. In *BIRNDL@ SIGIR*, 2018.
- [30] D. Debnath, A. Achom, and P. Pakray. Nlp-nitmz@ clscisumm-18. In *BIRNDL@ SIGIR*, pages 164–171, 2018.
- [31] L. Derczynski, E. Nichols, M. Van Erp, and N. Limsopatham. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, 2017.
- [32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [34] A. Elkiss, S. Shen, A. Fader, G. Erkan, D. States, and D. Radev. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1):51–62, 2008.
- [35] K. Ethayarajh. Unsupervised random walk sentence embeddings: A strong but simple baseline. In I. Augenstein, K. Cao, H. He, F. Hill, S. Gella, J. Kiros, H. Mei, and D. Misra, editors, *Proceedings*

of the Third Workshop on Representation Learning for NLP, pages 91–100, Melbourne, Australia, July 2018. Association for Computational Linguistics.

- [36] X. Feng, X. Feng, B. Qin, Z. Feng, and T. Liu. Improving low resource named entity recognition using cross-lingual knowledge transfer. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18*, page 40714077. AAAI Press, 2018.
- [37] B. Fetahu, A. Fang, O. Rokhlenko, and S. Malmasi. Gazetteer enhanced named entity recognition for code-mixed web queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681, 2021.
- [38] B. Fetahu, S. Kar, Z. Chen, O. Rokhlenko, and S. Malmasi. Semeval-2023 task 2: Fine-grained multilingual named entity recognition (multiconer 2). *arXiv preprint arXiv:2305.06586*, 2023.
- [39] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 168–171, 2003.
- [40] P. W. Foltz, W. Kintsch, and T. K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, Volume 25(2-3):pages 285–307, 1998.
- [41] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.
- [42] J. W. M. B. K. Gimpel and K. Livescu. Towards universal paraphrastic sentence embeddings. *arXiv:1511.08198*, 2015.
- [43] S. Goldberg, D. Z. Wang, and C. Grant. A probabilistically integrated system for crowd-assisted text labeling and extraction. *J. Data and Information Quality*, 8(2), Feb. 2017.
- [44] F. Grezes, S. Blanco-Cuaresma, T. Allen, and T. Ghosal. Overview of the first shared task on detecting entities in the astrophysics literature (deal). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 1–7, 2022.
- [45] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.
- [46] M. A. Hedderich, L. Lange, and D. Klakow. ANEA: distant supervision for low-resource named entity recognition. *CoRR*, abs/2102.13129, 2021.
- [47] Z. Hong, R. Tchoua, K. Chard, and I. Foster. Sciner: Extracting named entities from scientific literature. In V. V. Krzhizhanovskaya, G. Závodszy, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira, editors, *Computational Science – ICCS 2020*, pages 308–321, Cham, 2020. Springer International Publishing.
- [48] Z. Huang, W. Xu, and K. Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015.
- [49] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd annual meeting of*

the association for computational linguistics and the 7th international joint conference on natural language processing, pages 1681–1691. Association for computational linguistics, 2015.

- [50] S. A. Jadhav. Detecting potential topics in news using bert, CRF and wikipedia. *CoRR*, abs/2002.11402, 2020.
- [51] K. Jaidka, M. K. Chandrasekaran, D. Jain, and M. Kan. The cl-scisumm shared task 2017: Results and key insights. In K. Jaidka, M. K. Chandrasekaran, and M. Kan, editors, *Proceedings of the Computational Linguistics Scientific Summarization Shared Task (CL-SciSumm 2017) organized as a part of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017) and co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), Tokyo, Japan, August 11, 2017*, volume 2002 of *CEUR Workshop Proceedings*, pages 1–15. CEUR-WS.org, 2017.
- [52] K. Jaidka, M. K. Chandrasekaran, S. Rustagi, and M.-Y. Kan. Overview of the cl-scisumm 2016 shared task. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)*, pages 93–102, 2016.
- [53] K. Jaidka, M. Yasunaga, M. K. Chandrasekaran, D. Radev, and M.-Y. Kan. The cl-scisumm shared task 2018: Results and key insights. *arXiv preprint arXiv:1909.00764*, 2019.
- [54] K. Jaidka, M. Yasunaga, M. K. Chandrasekaran, D. Radev, and M.-Y. Kan. The cl-scisumm shared task 2018: Results and key insights, 2019.
- [55] S. Jain, M. van Zuylen, H. Hajishirzi, and I. Beltagy. Scirex: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, jul 2020.
- [56] X. Jiao, F. Wang, and D. Feng. Convolutional neural network for universal sentence embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2470–2481, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.
- [57] H. Kim and S. Ou. Nju@cl-scisumm-19. In *BIRNDL@SIGIR*, volume 2414 of *CEUR Workshop Proceedings*, pages 247–255. CEUR-WS.org, 2019.
- [58] S. Klampfl, A. Rexha, and R. Kern. Identifying referenced text in scientific publications by summarisation and classification techniques. In *Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL)*, pages 122–131, 2016.
- [59] V. Kocaman and D. Talby. Biomedical named entity recognition at scale. *CoRR*, abs/2011.06315, 2020.
- [60] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

- [61] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.
- [62] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.
- [63] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.
- [64] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv:1909.11942*, 2019.
- [65] T. K. Landauer and S. T. Dumais. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, Volume 104(2):pages 211, 1997.
- [66] J. Li, A. Sun, J. Han, and C. Li. A survey on deep learning for named entity recognition. *ArXiv*, abs/1812.09449, 2018.
- [67] J. Li, A. Sun, J. Han, and C. Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2020.
- [68] L. Li, J. Chi, M. Chen, Z. Huang, Y. Zhu, and X. Fu. Cist@ clscisumm-18: Methods for computational linguistics scientific citation linkage, facet classification and summarization. In *BIRNDL@SIGIR*, 2018.
- [69] L. Li, Y. Zhu, Y. Xie, Z. Huang, W. Liu, X. Li, and Y. Liu. Cist@clscisumm-19: Automatic scientific paper summarization with citances and facets. In *BIRNDL@SIGIR*, volume 2414 of *CEUR Workshop Proceedings*, pages 196–207. CEUR-WS.org, 2019.
- [70] X. Li, Y.-H. Lee, and J. Ouyang. Cited text spans for citation text generation. *arXiv preprint arXiv:2309.06365*, 2023.
- [71] Q. Liu, P. cheng Li, W. Lu, and Q. Cheng. Long-tail dataset entity recognition based on data augmentation. In *EEKE@JCDL*, 2020.
- [72] T. Liu, X. Wang, C. Lv, R. Zhen, and G. Fu. Sentence matching with syntax-and semantics-aware bert. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3302–3312, 2020.
- [73] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019.
- [74] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

- [75] J. Luoma and S. Pyysalo. Exploring cross-sentence contexts for named entity recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 904–914, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics.
- [76] J. Lala, O. O’Donoghue, A. Shtedritski, S. Cox, S. G. Rodrigues, and A. D. White. Paperqa: Retrieval-augmented generative agent for scientific research, 2023.
- [77] J.-Y. Ma, J.-C. Gu, J. Qi, Z.-H. Ling, Q. Liu, and X. Zhao. Ustc-nelslip at semeval-2023 task 2: Statistical construction and dual adaptation of gazetteer for multilingual complex ner. *arXiv preprint arXiv:2305.02517*, 2023.
- [78] L. Ma, Z. Sun, J. Jiang, and X. Li. Pai at semeval-2023 task 4: A general multi-label classification system with class-balanced loss function and ensemble module. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 256–261, 2023.
- [79] S. Ma, H. Zhang, J. Xu, and C. Zhang. Njust@ clscisumm-18. In *BIRNDL@ SIGIR*, 2018.
- [80] S. Ma, H. Zhang, T. Xu, J. Xu, S. Hu, and C. Zhang. Ir&tm-njust @ clscisumm-19. In *BIRNDL@SIGIR*, volume 2414 of *CEUR Workshop Proceedings*, pages 181–195. CEUR-WS.org, 2019.
- [81] S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics.
- [82] S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition. 2022.
- [83] S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States, July 2022. Association for Computational Linguistics.
- [84] S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022)*, pages 1412–1437, 2022.
- [85] S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, 2022.
- [86] C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

- [87] A. Mansouri, L. S. Affendey, and A. Mamat. Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2):339–344, 2008.
- [88] T. Meng, A. Fang, O. Rokhlenko, and S. Malmasi. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512, Online, June 2021. Association for Computational Linguistics.
- [89] S. Mesbah, C. Lofi, M. V. Torre, A. Bozzon, and G.-J. Houben. Tse-ner: An iterative approach for long-tail entity extraction in scientific publications. In *International Semantic Web Conference*, pages 127–143. Springer, 2018.
- [90] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [91] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of advances in neural information processing systems*, pages 3111–3119, 2013.
- [92] J. f. Mitchell and M. Lapata. Vector-based models of semantic composition. In *Proceedings of association for computational linguistics*, pages 236–244, 2008.
- [93] J. Mu and P. Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *Proceedings of international conference on learning representations*, 2018.
- [94] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.
- [95] F. Nooralahzadeh, J. T. Lønning, and L. Øvrelid. Reinforcement-based denoising of distantly supervised NER with partial annotation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 225–233, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [96] Y. Pitarch, K. Pinel-Sauvagnat, G. Hubert, G. Cabanac, and O. Fraïsier-Vannier. IRIT-IRIS at cl-scisumm 2019: Matching citances with their intended reference text spans from the scientific literature. In *BIRNDL@SIGIR*, volume 2414 of *CEUR Workshop Proceedings*, pages 208–213. CEUR-WS.org, 2019.
- [97] A. Prasad. Wing-nus at cl-scisumm 2017: Learning from syntactic and semantic similarity for citation contextualization. In *BIRNDL@ SIGIR (2)*, 2017.
- [98] M. L. Quatra, L. Cagliero, and E. Baralis. Poli2sum@cl-scisumm-19: Identify, classify, and summarize cited text spans by means of ensembles of supervised models. In *BIRNDL@SIGIR*, volume 2414 of *CEUR Workshop Proceedings*, pages 233–246. CEUR-WS.org, 2019.
- [99] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944, 2013.

- [100] A. Rahimi, Y. Li, and T. Cohn. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy, July 2019. Association for Computational Linguistics.
- [101] A. Rahimi, Y. Li, and T. Cohn. Multilingual ner transfer for low-resource languages. *ArXiv*, abs/1902.00193, 2019.
- [102] L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer, 1999.
- [103] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [104] S. Rijhwani, S. Zhou, G. Neubig, and J. Carbonell. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8118–8123, Online, July 2020. Association for Computational Linguistics.
- [105] A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1524–1534, 2011.
- [106] S. E. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *J. Documentation*, Volume 60:pages 503–520, 2004.
- [107] M. Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, 2006.
- [108] E. F. Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.
- [109] N. Sangeeth, B. Paul, and C. Chaudhary. Cair-nlp at semeval-2023 task 2: A multi-objective joint learning system for named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1926–1935, 2023.
- [110] S. Schweter and A. Akbik. FLERT: document-level features for named entity recognition. *CoRR*, abs/2011.06993, 2020.
- [111] P. Singh and A. Mukerjee. Words are not equal: Graded weighting model for building composite document vectors. In *Proceedings of the 12th international conference on natural language processing*, pages 11–19. NLP Association of India, 2015.
- [112] S. Singha Roy and R. E. Mercer. Building a synthetic biomedical research article citation linkage corpus. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5665–5672, Marseille, France, June 2022. European Language Resources Association.
- [113] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, Volume 28(1):pages 11–21, 1972.

- [114] B. Syed, V. Indurthi, B. V. Srinivasan, and V. Varma. Helium @ cl-scisumm-19 : Transfer learning for effective scientific research comprehension. In *BIRNDL@SIGIR*, volume 2414 of *CEUR Workshop Proceedings*, pages 214–223. CEUR-WS.org, 2019.
- [115] Z. Tan, S. Huang, Z. Jia, J. Cai, Y. Li, W. Lu, Y. Zhuang, K. Tu, P. Xie, F. Huang, et al. Damo-nlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. *arXiv preprint arXiv:2305.03688*, 2023.
- [116] S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia, July 2006. Association for Computational Linguistics.
- [117] E. F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.
- [118] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.
- [119] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.
- [120] N. B. Unnam, K. Reddy, A. Pandey, and N. Manwani. Journey to the center of the words: Word weighting scheme based on the geometry of word embeddings. In *34th International Conference on Scientific and Statistical Database Management, SSDBM 2022, New York, NY, USA, 2022*. Association for Computing Machinery.
- [121] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.
- [122] P. Wang, S. Li, T. Wang, H. Zhou, and J. Tang. Nudt@ clscisumm-18. In *BIRNDL@ SIGIR*, 2018.
- [123] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, and G. Wang. Gpt-ner: Named entity recognition via large language models, 2023.
- [124] X. Wang, Y. Guan, Y. Zhang, Q. Li, and J. Han. Pattern-enhanced named entity recognition with distant supervision. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 818–827, 2020.
- [125] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu. Automated concatenation of embeddings for structured prediction. *CoRR*, abs/2010.05006, 2020.
- [126] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu. Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

- Language Processing (Volume 1: Long Papers)*, pages 2643–2660, Online, Aug. 2021. Association for Computational Linguistics.
- [127] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online, Aug. 2021. Association for Computational Linguistics.
- [128] X. Wang, Y. Shen, J. Cai, T. Wang, X. Wang, P. Xie, F. Huang, W. Lu, Y. Zhuang, K. Tu, et al. Damo-nlp at semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv:2203.00545*, 2022.
- [129] X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui, J. Kang, J. Yang, S. Li, and C. Du. Instructuie: Multi-task instruction tuning for unified information extraction, 2023.
- [130] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, Y. Jiang, and W. Han. Zero-shot information extraction via chatting with chatgpt, 2023.
- [131] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.
- [132] J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.
- [133] V. Yadav and S. Bethard. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*, 2019.
- [134] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, Nov. 2020. Association for Computational Linguistics.
- [135] Y. Yang, W. Chen, Z. Li, Z. He, and M. Zhang. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.

- [136] D. Ye, Y. Lin, and M. Sun. Pack together: Entity and relation extraction with levitated marker. *CoRR*, abs/2109.06067, 2021.
- [137] J.-Y. Yeh, T.-Y. Hsu, C.-J. Tsai, and P.-C. Cheng. Reference scope identification for citances by classification with text similarity measures. In *proceedings of the 6th international conference on software and computer applications*, pages 87–91, 2017.
- [138] C. Zerva, M. Nghiem, N. T. H. Nguyen, and S. Ananiadou. Nactem-uom @ cl-scisumm 2019. In *BIRNDL@SIGIR*, volume 2414 of *CEUR Workshop Proceedings*, pages 167–180. CEUR-WS.org, 2019.
- [139] C. Zerva, M.-Q. Nghiem, N. T. H. Nguyen, and S. Ananiadou. Cited text span identification for scientific summarisation using pre-trained encoders. *Scientometrics*, 125:3109–3137, 2020.
- [140] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.
- [141] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou. Semantics-aware bert for language understanding, 2020.
- [142] W. Zhou and M. Chen. Learning from noisy labels for entity-centric information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5381–5392, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.