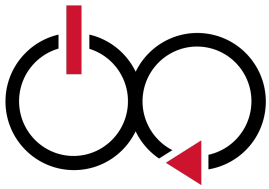
SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA



Department of **Informatics, Systems and Communication** and **P.M.I. Reboot srl**

Ph. D. program in **Computer Science**, **XXXVIII** cycle

Resource-Efficient and Knowledge-Enriched Information Extraction: From Named Entity Recognition to Relation Extraction

Gianmaria Balducci
807141

Tutor: **Prof. Alberto Leporati** (Università degli Studi di Milano-Bicocca)

Supervisor: **Prof. Elisabetta Fersini** (Università degli Studi di Milano-Bicocca), **Alberto Nosari** (P.M.I. Reboot srl)

Coordinator: **Prof. Leonardo Mariani** (Università degli Studi di Milano-Bicocca)

Academic Year **2024/2025**

ABSTRACT

With the exponential growth of unstructured textual data, extracting structured information has become essential. Named Entity Recognition (NER) and Relation Extraction (RE) are fundamental tasks in Information Extraction (IE), enabling the identification of entities and the relationships between them. These components are critical for numerous Natural Language Processing (NLP) applications, including sentiment analysis, chatbots, Retrieval Augmented Generation, Question Answering, and knowledge base construction. This thesis explores advanced methods for NER and RE, examining both traditional and deep learning approaches with particular emphasis on supervised learning paradigms. Conducted within an industrial PhD framework, this research addresses the dual challenge of advancing the state of the art while ensuring practical applicability and deployability in production environments. The work investigates three key challenges in information extraction: class imbalance in NER, cross-domain transfer learning for entity recognition, and knowledge-enhanced Relation Extraction. The first contribution investigates imbalance learning techniques for NER, including resampling and data augmentation methods to improve recognition of underrepresented entity types. The second contribution develops a Cross-Domain NER adaptation technique, which combines multiple shallow classifiers trained on different features extracted from pre-trained transformers to enable effective transfer learning across different domains. The third contribution focuses on Relation Extraction in both English and Italian Language, introducing methods that leverage Large Language Models (LLMs) and demonstrates how knowledge extracted from larger models can be effectively distilled into smaller, more efficient architectures through Knowledge Distillation techniques. Through systematic investigation and experimentation across multiple benchmark datasets, this work advances information extraction by addressing both theoretical challenges and practical constraints, including data scarcity, domain adaptation, computational efficiency, and resource limitations characteristic of real-world deployment scenarios.

ABSTRACT

Con la crescita esponenziale dei dati testuali non strutturati, l'estrazione di informazioni strutturate è diventata essenziale. Il riconoscimento delle entità denominate (NER) e l'estrazione delle relazioni (RE) sono compiti fondamentali nell'estrazione delle informazioni (IE), che consentono l'identificazione delle entità e delle relazioni tra di esse. Questi componenti sono fondamentali per numerose applicazioni di elaborazione del linguaggio naturale (NLP), tra cui l'analisi del sentiment, i chatbot, la RAG, il question-answering e la costruzione di basi di conoscenza. Questa tesi esplora metodi avanzati per NER e RE, esaminando sia gli approcci tradizionali che quelli di deep learning, con particolare enfasi sui paradigmi di apprendimento supervisionato. Condotta nell'ambito di un dottorato di ricerca industriale, questa ricerca affronta la duplice sfida di far progredire lo stato dell'arte garantendo al contempo l'applicabilità pratica e l'implementabilità in ambienti di produzione. Il lavoro indaga tre sfide chiave nell'estrazione di informazioni: Imbalance learning per la NER, Cross-Domain NER, e l'estrazione di relazioni. Il primo contributo indaga le tecniche di apprendimento in casi di dataset sbilanciati per il task di NER, compresi i metodi di ricampionamento e di aumento dei dati per migliorare il riconoscimento dei tipi di entità sottorappresentate. Il secondo contributo sviluppa una tecnica di adattamento NER cross-domain, che combina più classificatori addestrati su diverse features estratte da trasformers pre-addestrati per consentire un efficace apprendimento trasferibile tra diversi domini. Il terzo contributo si concentra sull'estrazione di relazioni sia in lingua inglese che italiana, introducendo metodi che sfruttano i modelli linguistici di grandi dimensioni (LLM) e dimostra come le conoscenze estratte da modelli più grandi possano essere efficacemente distillate in architetture più piccole ed efficienti attraverso tecniche di distillazione della conoscenza. Attraverso un'indagine sistematica e la sperimentazione su più dataset di riferimento, questo lavoro fa progredire l'estrazione di informazioni affrontando sia le sfide teoriche che i vincoli pratici, tra cui la scarsità di dati, l'adattamento del dominio, l'efficienza computazionale e le limitazioni delle risorse caratteristiche degli scenari di implementazione nel mondo reale.

LIST OF PUBLICATIONS IN THE SCOPE OF THE PHD THESIS

- *Cross-domain NER: a resource-efficient transfer-learning approach*
  Gianmaria Balducci, Elisabetta Fersini, Enza Messina at 30th Natural Language Information Systems 2025 conference.
- *Beyond Raw Text: Knowledge-Augmented Italian Relation Extraction with Large Language Models*
  Gianmaria Balducci, Elisabetta Fersini, Enza Messina at 11th Italian Conference on Computational Linguistics 2025
- *Knowledge-Augmented Distillation for Italian Relation Extraction*
  Gianmaria Balducci, Elisabetta Fersini, Enza Messina to be submitted at 12th Italian Conference on Computational Linguistics 2026
- *From Enrichment to Distillation: A Knowledge-Augmented Framework for Relation Extraction*
  Gianmaria Balducci, Elisabetta Fersini, Enza Messina submitted at the Information Processing & Management Journal (2026).

OTHER PUBLICATIONS

- *Bias Mitigation in Misogynous Meme Recognition: A Preliminary Study*
  Gianmaria Balducci, Giulia Rizzi, Elisabetta Fersini at Ninth Italian Conference on Computational Linguistics CLiC-it 2023.
- *Misogynous Memes Recognition: Training vs Inference Bias Mitigation strategies*
  Gianmaria Balducci, Giulia Rizzi, Elisabetta Fersini at Italian Journal of Computational Linguistics, 2025 (IJCOL).

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to all those who have supported me throughout this doctoral journey. I am profoundly grateful to my supervisor, Professor Elisabetta Fersini, for her invaluable guidance, expertise, and unwavering support throughout this research. Her insights and encouragement have been instrumental in shaping this work and my development as a researcher and as human being. I extend my sincere thanks to PMI Reboot S.R.L. for providing the computational infrastructure and server resources that made this research possible. Their technical support has been essential to the completion of this thesis. Finally, I owe my deepest appreciation to my family, whose love, patience, and constant encouragement have sustained me through the challenges of this doctoral program. To all who have contributed to this journey, named and unnamed, I am truly grateful.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

BACKGROUND

1 INTRODUCTION

From social media posts and news articles to scientific publications and enterprise documents, the volume of unstructured textual content continues to grow at an exponential rate. As this wealth of information accumulates, efficiently extracting knowledge becomes both a pressing challenge and a promising opportunity. Knowledge, in this context, refers to the factual content embedded within text that describes real-world phenomena, states, and connections. This knowledge manifests as assertions about the world: claims about what exists (objects, agents, concepts), descriptions of their properties and attributes, statements about how they connect or interact, and accounts of events that unfold over time. Raw text presents information in narrative, conversational, or rhetorical forms optimized for human comprehension. However, extracted knowledge must be explicit, unambiguous, and computationally accessible.

Information extraction (IE)—the process of automatically retrieving structured information from unstructured text—lies at the core of this endeavor. By transforming implicit factual content scattered throughout free-form text into explicit, formal representations, IE systems enable powerful downstream applications. These include systematic analysis across large corpora, pattern identification, cross-source claim verification, and evidence-based decision-making across diverse domains.

A foundational task within the broader IE pipeline is **Named Entity Recognition (NER)** [1]. NER is responsible for identifying and classifying spans of text that refer to real-world entities, such as persons, organizations, locations, dates, monetary values, and more. This task has become indispensable in numerous Natural Language Processing (NLP) applications, ranging from information retrieval and sentiment analysis to automatic summarization and question answering systems. In addition to functioning as a standalone task, NER serves as a critical precursor to more complex semantic analysis tasks. Accurate entity identification provides the bedrock upon which further linguistic and semantic reasoning can be constructed. However, identifying named entities in text is only the first step toward deeper understanding.

Occurrences of entities in a sentence are often linked through well-defined relations. **Relation extraction (RE)** [2] identifies these relation-

ships between entities in written content, forming the foundation for many natural language processing (NLP) and information retrieval applications, including knowledge graph completion and question answering systems. Relationships between entities are usually represented as relational triplets [3]: ⟨Entity 1, Relation, Entity 2⟩, where Entity 1 and Entity 2 are the connected entities, and Relation describes their connection. For instance, from the sentence "Bill Gates is the founder of Microsoft," we extract the relational triple ⟨Bill Gates, founder of, Microsoft⟩. Relation extraction provides practical value across numerous domains. In automatic question answering systems, it links related questions with their answers to improve response accuracy. In search and retrieval systems, it enables semantic retrieval that understands query meaning rather than merely matching keywords. In ontology learning, it discovers new entity relationships to enrich knowledge structures [4], [5], [6]. In semantic web labeling, it automatically associates knowledge units to create interconnected content. Through these applications, Relation Extraction transforms unstructured text into structured, machine-understandable knowledge.

Extracting such relationships is essential for building knowledge graphs and semantic databases [7], which serve as structured repositories of knowledge and enable powerful reasoning capabilities across various applications [8] [9], [10]. Despite significant progress in both NER and RE, a number of challenges persist. A key challenge that emerges concerns the adaptability of NER systems to diverse domains, where shifting semantics, contextual variability, and limited annotated data complicate the creation of universally robust models. Across all possible semantic definitions and interpretations of an entity, it's quite complex to find a resource that meets personalized needs.

Owing to the data scarcity issue in practical scenarios, obtaining adequate domain-related data is usually labour-intensive. The naive idea of training models with rich-resource domain data (source) and transferring knowledge to a new specific domain (target) may struggle handing the semantic gap and limited data problem [11]. Hence, cross-domain NER, capable of learning information from the source domain to specific target domains with limited data, has been proposed to alleviate this issue. Despite the empirical success of previous works, several issues remain which have not been be appropriately solved. Firstly, previous methods often rely on task-specific architectures for various domains with different entity categories. Some approaches design different model architectures to adapt a pre-trained Language Model (PLM) to other target domains, which restricts the model's us-

ability in more applications. Secondly, most current methods are computationally inefficient and require tuning all parameters of the PLMs. The evolving nature of language presents unique obstacles. New concepts, entities, and linguistic patterns continuously emerge, making it difficult for static models to keep pace. This linguistic dynamism necessitates models that are robust to domain shifts, temporal changes, and the introduction of previously unseen entity types or relations. Additionally, while supervised learning remains the dominant approach for training NER and RE models, it often requires large volumes of annotated data, which are expensive and time-consuming to produce. This limitation has spurred interest in alternative learning paradigms, such as distant supervision, semi-supervised learning, and unsupervised methods, which seek to reduce the dependence on labeled data by leveraging existing knowledge bases or unannotated corpora. These approaches, while promising, introduce their own set of challenges, such as noisy labeling, weak supervision, and difficulty in generalization.

For Relation Extraction a wide range of studies [12], [13], [14] has explored how integrating external knowledge can improve understanding of the task, especially when training data are limited. Some approaches embed prior knowledge from lexical databases and semantic resources directly into neural encoders, while others enrich sentence representations by combining structured information from knowledge graphs with semantic cues drawn from large text corpora. Despite these advantages, external knowledge sources often introduce issues such as noise, missing data, or contextually irrelevant information that can hinder performance in specialized domains [15]. This limitation has motivated research into leveraging Large Language Models (LLMs) as more reliable and contextually aware knowledge sources.

This thesis lies at the intersection of Named Entity Recognition (NER) and Relation Extraction (RE), addressing some of the most persistent challenges in information extraction under resource-constrained conditions. In real-world scenarios, the scarcity of annotated data, domain-specific pre-trained models, and reusable linguistic resources significantly limits the robustness and scalability of NER and RE systems. To mitigate these limitations, the thesis investigates Cross-Domain NER, examining how models can adapt to new domains with limited annotated data by exploiting latent representations and transfer-learning strategies derived from source-domain features. This approach seeks to maintain high performance while reducing the computational and annotation costs typically associated with domain-specific fine-tuning.

Beyond entity recognition, Relation Extraction in low-resource settings introduces further complexity, as it often depends on external knowledge bases or lexical resources to capture semantic relations. These approaches are frequently static, incomplete, or poorly aligned with specific contexts. To address this problem, the thesis leverages Large Language Models (LLMs) as dynamic and context-aware sources of background knowledge, capable of reasoning beyond fixed resources. Finally, this work introduces a knowledge distillation framework that transfers the reasoning and contextualization capabilities of LLMs into lightweight models, enabling efficient and scalable deployment without additional knowledge. Overall, the proposed methodologies aim to tackle challenges arising both from the intrinsic complexity of natural language (ambiguity, polysemy, linguistic dynamism) and from practical constraints related to efficiency, data scarcity, and domain adaptability.

1.1 CONTRIBUTION OF THE THESIS

This thesis introduces a resource-efficient framework for **Named Entity Recognition (NER)** and **Relation Extraction (RE)**, designed to confront the central challenges of information extraction: limited computational resources, data scarcity, class imbalance, and cross-domain generalization. The proposed framework explores different techniques such as transfer learning, ensemble modeling, and knowledge distillation to enable high-performance information extraction even under constrained conditions, offering a scalable alternative to computationally demanding large-scale systems. The research progresses systematically through the information extraction pipeline, beginning with fundamental challenges in entity recognition, advancing through cross-domain adaptation strategies, and culminating in methods that leverage and distill the reasoning capabilities of Large Language Models for Relation Extraction. By bridging the gap between resource-intensive state-of-the-art models and the practical constraints of real-world deployment, these contributions establish a foundation for building accurate, interpretable, and computationally sustainable information extraction systems deployable across diverse domains and languages. The main contributions of the PhD thesis are summarized as follows:

- The first contribution introduces a study on the **imbalance** problem on NER. Class imbalance is a long-standing problem in machine learning tasks, posing challenges to researchers and practitioners in many domains [16]. Addressing class imbalance is

crucial in NER tasks to ensure that the models generalize well across all named entity types. The study experiments with models trained on OntoNotes 5.0 [17] and CoNLL03 [18], well-known benchmarks for English NER. Investigating different imbalance learning techniques across both span-based and transformer-based architectures, results demonstrate that the optimal strategy depends on model architecture: sentence-level resampling methods like smoothed count prove most effective for SpanCategorizer, while simple random oversampling suffices for BERT models leveraging pre-trained representations. Notably, random undersampling fails catastrophically across all configurations, underscoring the fundamental reliance of modern NER systems on non-entity tokens for accurate boundary detection.

- The second contribution involves in **Cross-Domain NER**, we explore several combinations of transfer learning between different sources and target domains. This contribution investigate the possibility, under a limited resources scenario, if the target domain labels and their distributions can be learned by exploiting features obtained from the source domain and reaching comparable performance with respect to other transfer-learning techniques, but at lower computational costs. Through comprehensive evaluation across CoNLL03 [18], OntoNotes 5.0, and MultiNERD [19] benchmarks, results demonstrate that leveraging transformer-based model characteristics, including hidden states, attention vectors, and probability distributions, can achieve performance comparable to fine-tuning while significantly reducing computational costs.

- The third contribution addresses Relation Extraction by leveraging the **reasoning capabilities of Large Language Models** to generate reliable background knowledge. The proposed method systematically enriches training data by extracting related knowledge from an LLM. The approach then formalizes a **Knowledge Distillation** process through explicit minimization of Kullback–Leibler (KL) divergence between output distributions of models trained on enriched versus raw sentence inputs. Experimental validation on the CoNLL04 and SemEval 2010 Task 8 datasets demonstrates the effectiveness of this knowledge augmentation and distillation strategy for English Relation Extraction. Subsequently, the approach is extended to the **Italian language**, adapting the methodology for knowledge generation and

validating the cross-lingual applicability on translated Italian CoNLL04 dataset, confirming that the proposed framework maintains its effectiveness across different languages.

1.2 ORGANIZATION OF THE THESIS

In the following, we detail each Chapter of this thesis, highlighting the research questions that have been addressed.
Chapter 1: Introduction. In this Chapter, we have introduced the content of this thesis, outlining the main concepts and the most relevant content related to Named Entity Recognition and Relation Extraction in resource-constrained scenarios. The rest of the Chapter also outlines the main research questions that this work aims to answer.
Chapter 2: Preliminaries This chapter establishes the core theoretical basis for this research, providing readers with the necessary knowledge to understand the work that follows. This section offers a thorough introduction to the research conducted in this thesis, presenting concepts, structures, computational methods, and key ideas that support our study of Named Entity Recognition and Relation Extraction across different domains.
Chapter 3: Related Work. This Chapter summarizes the state of the art in Named Entity Recognition and Relation Extraction. We explore traditional and modern approaches to NER, particularly focusing on cross-domain adaptation. We also review Relation Extraction techniques, including knowledge-augmented approaches and the use of Large Language Models.
Chapter 4: Imbalance Learning in Named Entity Recognition: This chapter examines the significant issue of unequal class distribution in Named Entity Recognition (NER) applications, a core problem that substantially impacts how models perform when identifying various entity categories. Within the larger investigation of extracting detailed named entities and their relationships from text content, this chapter analyzes different approaches to tackle the natural disproportion found in conventional NER datasets. The research question answered in this Chapter is:

RQ 4.1 - How can we effectively address class imbalance in Named Entity Recognition tasks?

Chapter 5: Cross-Domain Named Entity Recognition. This Chapter addresses the challenges of domain adaptation in NER when dealing

with limited annotated resources. Public NER resources, such as annotated datasets and annotation services exist across various domains. No single resource typically supports all entity types required for specific downstream applications. Additionally, the availability of training data to effectively develop NER systems for different domain classification schemas is often limited due to constraints on time, quality, and annotation costs. We explore a transfer-learning approach that exploits different features obtained from the source domain to train several models and compose a heterogeneous ensemble that learns target domain labels and their distributions. The research questions answered in this Chapter are:

RQ 5.1 - How can features obtained from a source domain be leveraged to learn target domain labels and distributions efficiently, achieving performance close to fine-tuned transformers but with lower computational cost?

RQ 5.2 - Which representations matter most to deal with the problem of transfer learning for NER?

Chapter 6: Knowledge-Augmented Relation Extraction with Large Language Models. In this Chapter, we explore how Large Language Models can be leveraged to enhance Relation Extraction performance through systematic knowledge augmentation. We investigate the generation of background knowledge and its impact on model performance in English and Italian texts. Furthermore, this Chapter investigates the distillation of knowledge from models trained on enriched data to lightweight models operating on raw text. We explore a distillation technique that preserves reasoning capabilities while eliminating dependencies on auxiliary information during inference. The research questions answered in this Chapter are:

RQ 6.1 - Which complementary knowledge component can be obtained by Large Language Models to augment training data for Relation Extraction?

RQ 6.2 - What is the effect of knowledge distillation when considering only model that takes as input only the raw sentence?

Chapter 7: Conclusions and Future Works. This thesis concludes with a summary of the main findings and an overview of the most

significant contributions to research in Named Entity Recognition and Relation Extraction under resource-constrained scenarios. It also suggests directions for future research that could build upon the findings presented in this thesis.

2

PRELIMINARIES

This chapter establishes the foundational concepts and theoretical frameworks essential for understanding the research presented in this thesis. We begin by describing the critical need for Named Entity Recognition (NER) and Relation Extraction (RE) as fundamental capabilities for understanding document content, enabling precise information retrieval, and supporting downstream analytical tasks in domain-specific applications. The chapter then provides comprehensive coverage of Named Entity Recognition as a structured prediction problem, examining fundamental concepts including tagging schemes and entity type taxonomies. Given the data constraints typical in real-world scenarios, we dedicate substantial attention to transfer learning paradigms-including homogeneous and heterogeneous transfer, with particular emphasis on fine-tuning strategies that enable effective adaptation of pre-trained models to domain-specific tasks. We also examine knowledge distillation as a critical technique for model compression, enabling the deployment of performant yet efficient models while preserving the capabilities learned from larger teacher models. Relation Extraction is introduced as the natural extension of NER, enabling the identification of semantic connections between recognized entities across both general and domain-specific contexts. We conclude with a discussion of evaluation metrics that provide a rigorous assessment of model performance for these structured prediction tasks.

2.1 NAMED ENTITY RECOGNITION: FOUNDATIONS AND FORMULATIONS

Named Entity Recognition (NER) constitutes a fundamental information extraction task that involves identifying and classifying mentions of named entities in unstructured text into predefined categories [20]. As a core component in NLP pipelines, NER serves as a prerequisite for numerous downstream applications, including Relation Extraction, question answering, and knowledge base construction. Named Entity Recognition can be understood as a structured prediction problem where we need to make decisions about each token in a sentence while considering the context and dependencies between these decisions.

The fundamental challenge lies in determining not just what type of entity each token belongs to, but also understanding the boundaries of these entities within the text. When approaching NER as a machine learning problem, we typically frame it as finding the most probable sequence of labels for a given input text. Given a sequence of tokens, our goal is to assign each token a label that indicates whether it's part of an entity and, if so, what type of entity it belongs to. This optimal label assignment maximizes the probability of the entire sequence being correct, which can be expressed as:

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in \mathcal{Y}^n} P(\mathbf{y}|\mathbf{x})$$

This formulation captures the essence of NER: we seek the label sequence $\mathbf{y}^*$ that is most likely given our observed token sequence $\mathbf{x}$.
The way we assign labels to tokens significantly impacts the performance and interpretability of NER systems. Two primary encoding schemes have emerged as standards in the field, each with distinct advantages for different applications [21], [22].
The BIO (Beginning-Inside-Outside) encoding scheme provides a straightforward approach to entity labeling. In this scheme, each token receives one of three types of labels: B-tags mark the beginning of an entity of a specific type, I-tags indicate continuation of an entity, and O-tags denote tokens that are not part of any entity. This scheme effectively handles most entity recognition scenarios and has become widely adopted due to its simplicity and effectiveness.
The BILOU (Beginning-Inside-Last-Outside-Unit) encoding extends the BIO scheme by adding more explicit boundary information. This scheme introduces L-tags to mark the last token of multi-token entities and U-tags for single-token entities. The additional boundary information provided by BILOU has shown to improve entity boundary detection performance, particularly for complex entity structures [23].

Table 1: Common NER tagging schemes comparison

BIO Encoding	**BILOU Encoding**
B – T: Beginning of entity type T	B – T: Beginning of multi-token entity
I – T: Inside entity type T	I – T: Inside of multi-token entity
O: Outside any entity	L – T: Last token of multi-token entity
	U – T: Unit-length entity
	O: Outside any entity

The effectiveness of any NER system depends heavily on the entity types it recognizes and how these types are defined. Entity type selection represents a fundamental design decision that influences both the system's capabilities and its practical applications [24].
Most general-purpose NER systems focus on four primary entity categories that appear frequently across diverse text domains.
Person entities encompass individual names in various forms, from simple first names to complete formal names with titles and suffixes.
Location entities capture geographical references ranging from specific addresses to broad regional designations like countries and continents.
Organizational entities identify institutional references, including corporations, government agencies, educational institutions, and non-profit organizations.
The *miscellaneous* category serves as a catch-all for named entities that don't fit the other categories, typically including events, products, awards, and cultural references.
Specialized applications often require recognition of additional entity types beyond the core categories. Temporal entities capture time-related expressions including specific dates, time ranges, and duration references. Numerical entities encompass monetary amounts, percentages, measurements, and other quantitative expressions. Domain-specific entity types emerge in specialized fields: biomedical NER systems recognize genes, proteins, diseases, and drug names [25] [26], [27], [28], while legal text processing systems identify laws, court cases, and legal precedents [29], [30], [31].
Real-world text presents several challenging scenarios that complicate straightforward entity recognition. Nested entities occur when one entity contains another, such as "University of California, Berkeley" containing both an organization and a location reference. Overlapping entities present situations where the same text span can be interpreted as multiple entity types, like "Apple Inc." functioning as both an organization name and potentially a product reference. Discontinuous entities appear when entity mentions are split across non-contiguous text spans, such as "New York" and "Yankees" being separated by intervening tokens while together forming a single organization reference.

2.1.1 *Transfer Learning*

Transfer learning represents one of the most practical paradigms in modern machine learning, addressing the fundamental challenge of leveraging knowledge from data-rich domains to improve performance

in data-scarce scenarios. The field can be systematically organized along several key dimensions that reflect both the nature of the learning problem and the mechanisms through which knowledge is transferred [32].

The relationship between source and target environments fundamentally shapes the transfer learning approach. In homogeneous transfer learning, both domains share identical feature spaces and similar data distributions, making knowledge transfer relatively straightforward [33]. This scenario often occurs when applying models across different but related datasets, such as transferring sentiment analysis models between product review domains.

Conversely, heterogeneous transfer learning [34] tackles the more challenging scenario where domains differ in their feature representations or underlying distributions. This might involve transferring knowledge from text to images, or adapting models trained on one language to another with different linguistic structures.

The task-oriented perspective introduces additional distinctions. Inductive transfer learning addresses scenarios where source and target tasks differ, necessitating some labeled data in the target domain to guide the adaptation process. This is common in specialized applications where general models need task-specific refinement. Transductive transfer learning [35] maintains task consistency while adapting to different domains, focusing purely on distribution shift challenges. Finally, represents the most challenging scenario, operating without labeled data in either domain and relying entirely on structural similarities in the data.

Parameter-based transfer learning has become the dominant paradigm in deep learning, largely due to its simplicity and effectiveness. The fundamental insight is that parameters learned on large-scale source tasks often encode generally useful representations that can be adapted to target tasks with minimal modification [36], [37].

Fine-tuning represents the most straightforward approach to parameter transfer. The process begins with a model pre-trained on a large source dataset, followed by continued training on the target task. This seemingly simple procedure involves numerous strategic decisions that significantly impact performance [36], [37]. Full fine-tuning updates all model parameters during target task training, providing maximum flexibility but risking overfitting when target data is limited. Partial fine-tuning freezes lower layers while updating higher ones, based on the intuition that early layers capture general features while later layers encode task-specific patterns. Layer-wise fine-tuning adopts a

gradual unfreezing strategy, starting with the highest layers and progressively incorporating lower layers as training progresses [38]. The choice among these strategies depends on target dataset size, domain similarity, and computational constraints. Large target datasets with sufficient diversity generally benefit from full fine-tuning, while small datasets often perform better with partial approaches that limit the risk of overfitting [39].

Recent advances have introduced parameter-efficient methods that achieve competitive performance while modifying only a small fraction of pre-trained parameters. These approaches address both computational efficiency and the practical challenges of maintaining multiple model variants. Adapter layers insert small bottleneck modules between existing model layers, keeping the original parameters frozen while learning task-specific adaptations through these compact modules [40]. This approach enables easy switching between tasks and reduces storage requirements for multi-task scenarios. Low-Rank Adaptation (LoRA) operates on the insight that parameter updates during fine-tuning often have low intrinsic dimensionality. By approximating updates through low-rank matrix factorizations, LoRA achieves comparable performance to full fine-tuning while updating orders of magnitude fewer parameters [41]. Prefix tuning takes a fundamentally different approach by optimizing continuous task-specific vectors that are prepended to model inputs, leaving all original parameters unchanged [42]. This method is particularly effective for language models where the prefix can guide generation towards task-appropriate outputs.

2.1.2 *Knowledge Distillation*

Knowledge distillation has emerged as one of the most elegant solutions to a fundamental challenge in modern machine learning: how to capture the sophisticated understanding of large, powerful models in smaller, more practical systems. This paradigm becomes increasingly critical as models grow ever larger while deployment environments become more resource-constrained. The core insight driving knowledge distillation is that large models contain far more information than their final predictions alone—they possess rich internal representations and nuanced understanding that can be transferred to more efficient architectures [43].

The teacher-student metaphor that defines knowledge distillation reflects a profound shift in how we think about model training. Rather

than learning directly from raw data labels, student models can benefit from the accumulated wisdom of their larger counterparts. This approach often yields better performance than training smaller models independently, suggesting that the learning process itself can be optimized through careful knowledge transfer.
Consider a practical example: compressing a large BERT model for mobile deployment. The teacher BERT-Large with 340M parameters produces rich probability distributions over vocabulary tokens, revealing not just the most likely answer but also plausible alternatives and their relative probabilities. A student model like DistilBERT with only 66M parameters learns to mimic these soft predictions, capturing much of the teacher's linguistic understanding while being 6× smaller and 2× faster. Other common knowledge distillation techniques often used in state-of-the-art works are base on Kullback-Leibler (KL) divergence [44] and Jensen-Shannon (JS) divergence [45].

2.2 RELATION EXTRACTION

Relation Extraction (RE) constitutes a fundamental information extraction task that identifies and classifies semantic relationships between entities mentioned in text [46]. Building upon named entity recognition, Relation Extraction enables the construction of structured knowledge from unstructured text by discovering how entities interact and relate to one another. As a critical component in knowledge base construction, question answering, and semantic search systems, RE transforms unstructured narratives into machine-readable structured representations that capture the relational information. Relation extraction can be formally defined as a classification problem over entity pairs. Given a text segment containing a set of identified entities $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$, the task is to determine whether any semantic relation exists between entity pairs and, if so, to classify that relation into predefined types from a relation taxonomy $\mathcal{R}$.

For a given entity pair (e_i, e_j) within a textual context $\mathbf{x}$, the Relation Extraction model must predict:

$$r^* = \arg\max_{r \in \mathcal{R}} P(r|e_i, e_j, \mathbf{x})$$

where r^* represents the most likely relation type, and the special label "no_relation" indicates that no semantic relationship exists between the entity pair. This formulation captures the directional nature of many relations—the relationship from entity e_i to e_j may differ from

that of e_j to e_i (e.g., "founded" versus "was founded by").
The complexity of Relation Extraction varies significantly based on the assumptions and constraints imposed on the task. In sentence-level RE [47], entities and their relations are constrained to occur within single sentences, simplifying context modeling but potentially missing inter-sentential relationships. Document-level RE [48] relaxes this constraint, requiring models to identify relations between entities that may be mentioned multiple sentences apart, demanding more sophisticated context aggregation and coreference handling. Relation extraction approaches can be categorized along several dimensions that reflect different methodological perspectives and practical considerations. The pipeline approach treats NER and RE as sequential, independent tasks. Entities are first identified through a dedicated NER model, and the recognized entities are subsequently passed to a separate RE model that classifies relationships between entity pairs. While conceptually simple and modular, this approach suffers from error propagation
mistakes in entity recognition cascade into the Relation Extraction stage, and the RE model cannot provide feedback to improve entity boundary detection [49].
Joint extraction models address these limitations by simultaneously identifying entities and their relations in a unified framework [50]. By modeling the interdependencies between entity recognition and relation classification, joint models can leverage relational information to improve entity detection and vice versa. For instance, recognizing that "founded" likely connects a person to an organization can help disambiguate entity types. However, joint models introduce additional complexity in model architecture and training procedures [51].
Traditional closed RE operates with a predefined relation taxonomy, classifying entity pairs into one of several known relation types. This approach enables focused extraction of specific relationship types relevant to a domain but cannot generalize to unseen relations. The model is fundamentally limited by the relation types encountered during training. Open Information Extraction (Open IE) takes a more flexible approach, extracting relation phrases directly from text without constraining them to a fixed schema. For the sentence “Einstein developed the theory of relativity,” an Open IE system would extract the triple (*Einstein, developed, theory of relativity*) without requiring “developed” to match any predefined relation type. This paradigm offers greater flexibility and coverage but produces less standardized outputs that may be more difficult to integrate into structured knowledge bases [52, 53].

Supervised Relation Extraction relies on manually annotated training data where human annotators have identified and labeled relationships between entities. While this produces high-quality training signals, manual annotation is expensive, time-consuming, and difficult to scale to large corpora or specialized domains.

Distant supervision offers an alternative by automatically generating training examples from existing knowledge bases. The fundamental assumption is that if a knowledge base contains a relation between two entities, then any sentence mentioning both entities is likely to express that relation. For example, if a knowledge base records that "Steve Jobs founded Apple," sentences containing both "Steve Jobs" and "Apple" can be automatically labeled as expressing a "founded" relation [54, 55]. While enabling training on much larger datasets, distant supervision introduces label noise since not all co-occurrences express the assumed relation, requiring models to handle noisy labels robustly [55, 56].

The definition and organization of relation types fundamentally shapes the capabilities and applications of Relation Extraction systems. Relation taxonomies must balance comprehensiveness with practicality, providing sufficient coverage while maintaining clear, distinguishable categories that can be reliably identified by both human annotators and machine learning models [57].

Most general-purpose Relation Extraction systems focus on common relationship types that appear frequently across diverse text domains. Employment relationships (*works_for*, *employed_by*) capture professional affiliations between people and organizations. Residential relationships (*lives_in*, *based_in*) identify geographic associations of individuals or institutions. Organizational founding relations (*founded*, *established*, *created*) document the origins of institutions and companies. Personal relationships, including familial connections (*parent_of*, *spouse_of*, *sibling_of*) and social associations (*friend_of*, *colleague_of*), represent important categories for biographical and social information extraction. Part-whole relations (*part_of*, *contains*, *member_of*) capture compositional and membership structures.

Specialized applications often require recognition of domain-specific relationship types that capture the unique semantic patterns of particular fields. In biomedical Relation Extraction, critical relations include drug-disease interactions (*treats*, *causes*, *prevents*), protein-protein interactions (*binds_to*, *activates*, *inhibits*), and gene-disease associations (*associated_with*, *causes*, *increases_risk_of*). Legal text processing requires recognition of citation relationships (*cites*, *overrules*, *distinguishes*), reg-

ulatory connections (*regulates*, *enforces*, *violates*), and jurisdictional relations. Financial documents contain investment relationships (*invests_in*, *acquires*, *divests*), ownership structures (*owns*, *controls*, *subsidiary_of*), and market influence patterns (*competes_with*, *supplies_to*, *distributes_for*).
Real-world text presents several challenging scenarios that significantly complicate Relation Extraction and require sophisticated modeling approaches.
When text contains multiple entities, their relationships may overlap in complex ways. Consider "The University of California, Berkeley professor received funding from the National Science Foundation." This sentence contains multiple entities (university, location, person, organization) with various relationships (affiliation, location, funding) that a comprehensive extraction system must disentangle.
Not all relations are explicitly stated through surface lexical patterns. The sentence "John entered the restaurant and ordered coffee" implicitly contains a customer-business relationship that requires commonsense reasoning to infer. Models must go beyond pattern matching to understand implied relationships based on world knowledge and contextual understanding.
Many relations are inherently directional—"founded" and "was founded by" represent inverse relations with different head and tail entities. However, some relations are symmetric (*spouse_of*, *sibling_of*), while others may be transitive (*larger_than*, *part_of*) or follow other logical properties. Relation extraction systems must appropriately handle these logical properties, potentially enforcing consistency constraints during prediction.

2.2.1 *Evaluation Metrics*

Relation extraction performance is typically evaluated using precision, recall, and F1-score, computed over predicted relation instances. A predicted relation triple (e_i, r, e_j) is considered correct if it matches a gold-standard annotation with the same head entity, relation type, and tail entity.
Different evaluation settings impose varying requirements on entity boundary matching. In strict evaluation, entity boundaries must match exactly for a relation to be counted as correct. In relaxed evaluation, entities are considered matching if they overlap, accommodating minor boundary variations. For scenarios prioritizing relation discovery over precise entity spans, relation-level evaluation focuses solely on whether the correct relation type is identified between roughly the cor-

rect entities, regardless of exact boundaries.
Beyond instance-level metrics, macro-averaging across relation types provides insights into performance consistency across the relation taxonomy, highlighting whether models perform uniformly well or exhibit significant disparities between common and rare relation types.

3

RELATED WORK

This chapter provides a comprehensive review of the existing literature on information extraction, with a particular focus on **Named Entity Recognition (NER)** and **Relation Extraction (RE)**. The discussion is structured to reflect the historical evolution, methodological innovations, and recent advances that have shaped modern research in these areas.

We begin with an overview of the *evolution of NER systems*, tracing their development from early rule-based and statistical models to contemporary deep learning and large language model-based architectures. This section highlights key benchmark datasets, competitions, and milestones that established the foundations of the field. The following section delves into *neural approaches to NER*, examining how deep architectures—ranging from BiLSTM-CRF models to transformer-based encoders—have redefined the task by enabling richer contextual understanding and more robust generalization. Within this context, we discuss the role of attention mechanisms, domain adaptation strategies, and the recent integration of Large Language Models as reasoning-aware NER systems capable of leveraging implicit knowledge during prediction.

The final part of the chapter is dedicated to *Relation Extraction (RE)*, which extends the information extraction pipeline by identifying semantic relations between recognized entities. We first review benchmark datasets and evaluation frameworks that have supported empirical research in this domain. Then, we discuss the deep learning revolution that transformed RE, emphasizing both supervised and distant supervision paradigms. Lastly, we analyze how Large Language Models have reshaped the landscape of Relation Extraction—offering new opportunities for zero-shot generalization, contextual reasoning, and explainable knowledge discovery. Overall, this chapter situates NER and RE within a unified perspective, emphasizing their interdependence and the technological continuum that connects early sequence labeling techniques with modern reasoning-capable architectures.

3.1 EVOLUTION OF NAMED ENTITY RECOGNITION

Named Entity Recognition emerged as a fundamental task in natural language processing during the Message Understanding Conferences (MUC) in the 1990s b[58]. MUC-6/MUC-7 marked the starting point for NER as a shared task, establishing the foundation for identifying entities such as persons, locations, and organizations in text. In the Figure 1 is represented a generic example of NER output that classifies entities like Person and Product.

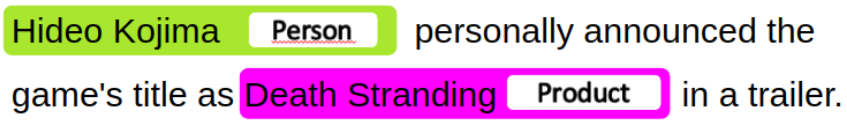


Figure 1: Example of Named Entity Recognition extracting Person and Product categories

3.1.1 *Early Benchmark Datasets and Competitions*

The development of NER systems has been closely tied to the availability of standardized datasets and evaluation frameworks. The CoNLL-2003 [59] shared task introduced a significant milestone by focusing on language-independent NER, providing training and test data for English and German with four entity types: persons, locations, organizations, and miscellaneous entities. Sixteen systems participated in the CoNLL-2003 shared task, utilizing a wide variety of machine learning techniques and different feature sets. The CoNLL 2002 and CoNLL 2003 datasets were among the early efforts to create NER datasets using multiple languages and introducing new entity types, responding to trends in technology at that time. These competitions established important precedents for evaluation metrics and methodologies that continue to influence NER research today. Following these foundational works, the ACE (Automatic Content Extraction) [60] project, conducted since 2002 as a successor to the MUC named entity recognition task, shifted the emphasis from initial entity recognition to entity resolution. ACE introduced not only more entity types and subtyping but also considered the annotation of nested entities.

3.1.2 *Sequence Labeling with Conditional Random Fields*

Traditional statistical approaches to NER often employ Conditional Random Fields (CRFs), which provide a principled framework for sequence labeling tasks. CRFs were first introduced by Lafferty et al. [61] as a framework for building probabilistic models to segment and label sequence data. Their application to NER was pioneered by McCallum et al. [62], who demonstrated their effectiveness in capturing dependencies between adjacent labels while incorporating rich feature representations of the input text.

The fundamental insight behind CRFs for NER lies in their ability to model the conditional probability of an entire label sequence given the input text. Rather than making independent decisions about each token's label, CRFs consider the global context and constraints across the entire sequence. This global perspective allows the model to avoid inconsistent label sequences, such as an I-tag following an O-tag without an intervening B-tag, a common problem in traditional classification approaches.

Early applications of CRFs to NER, [63], [64], [65] demonstrated significant improvements over previous approaches. In this work [66] is presented a framework for biomedical named entity recognition using CRFs with a variety of features, showing that CRFs with simple orthographic features could achieve competitive performance on biomedical texts. This work established CRFs as a viable approach for domain-specific NER tasks.

McCallum et al. [62] conducted comprehensive experiments comparing CRFs to Hidden Markov Models and Maximum Entropy Markov Models for information extraction tasks, including NER. Their results demonstrated that CRFs consistently outperformed alternative approaches, particularly in scenarios with rich feature representations and overlapping features.

Sha et al. [67] provided theoretical analysis of CRFs for sequence labeling, establishing convergence properties and optimization techniques that became standard in the field. Their work laid the mathematical foundation for efficient training of CRF models on large-scale NER datasets.

The power of CRFs lies in their flexible feature representation capabilities, which was extensively explored in early NER research. Finkel et al. [68] developed a comprehensive feature set for English NER that included orthographic features (capitalization patterns, digit patterns,

punctuation), linguistic features (part-of-speech tags, phrase chunks), and contextual features (word windows, character n-grams).

3.2 NEURAL APPROACHES TO NAMED ENTITY RECOGNITION

The introduction of neural network architectures revolutionized NER systems, moving from feature-heavy approaches to end-to-end learning models. The emergence of neural approaches has dramatically transformed the research landscape [69], enabling more sophisticated modeling of semantic relationships and contextual dependencies that were challenging to capture with traditional methods. This section provides a comprehensive overview of the evolution and current state-of-the-art in neural NER systems.

3.2.1 *Foundational Neural Architectures*

The transition from traditional machine learning approaches to neural methods began with the introduction of word embeddings and basic neural network architectures. Early neural approaches to NER leveraged the power of distributed representations, where words are encoded as dense vectors in high-dimensional spaces, capturing semantic similarities and relationships that were previously difficult to model explicitly [70]. The foundational work by Collobert et al. [71] demonstrated that deep learning could be successfully applied to various NLP tasks, including NER. This seminal work introduced a unified neural architecture capable of solving multiple sequence labeling problems without task-specific feature engineering. The model employed convolutional neural networks over word embeddings, showing that neural approaches could match or exceed the performance of traditional methods while requiring minimal domain knowledge. Building on this foundation, Guimarães et al[72] proposed a character-level convolutional neural network for NER that could capture morphological patterns and handle out-of-vocabulary words more effectively. This approach was particularly valuable for languages with rich morphology and demonstrated the importance of character-level information in entity recognition tasks.

3.2.2 *BiLSTM-CRF Architecture*

The BiLSTM-CRF architecture, proposed by Huang et al. [73], became a cornerstone in neural NER systems. This architecture combines the sequence modeling capabilities of bidirectional Long Short-Term Memory [74] networks with the structured prediction advantages of Conditional Random Fields. The CRF layer takes the output of the BiLSTM layer and manages the sequence tagging process, utilizing contextualized tag information to produce better tagging accuracy through the incorporation of label dependencies and constraints. The BiLSTM-CRF model exhibits promising results across various domains, including biomedical applications where two works [75, 76], demonstrated its effectiveness. In biomedical research, the attention-based BiLSTM-CRF approach has been successfully applied to document-level chemical named entity recognition, addressing the tagging inconsistency problem common in sentence-level NER methods [77]. The architecture's ability to capture both forward and backward contextual information while maintaining structured output constraints has made it particularly suitable for complex entity recognition tasks in specialized domains. Building upon the success of BiLSTM-CRF models, researchers developed more sophisticated recurrent architectures to address specific challenges in NER. This paper [78] proposed a multi-task learning framework that jointly trains NER with other sequence labeling tasks, improving generalization and reducing overfitting. Their model incorporates additional linguistic features and demonstrates significant improvements over single-task approaches. Another approach [79] introduced a transfer learning mechanism for neural NER, showing how pre-trained language models can be fine-tuned for entity recognition tasks. This work laid the groundwork for the transformer-based approaches that would later dominate the field. The authors demonstrated that cross-lingual transfer learning could be particularly effective for low-resource languages.

3.2.3 *Attention Mechanisms and Transformer Architecture*

The transformer architecture, Figure 2, represents a paradigm shift in sequence modeling, replacing recurrent connections with attention mechanisms that enable parallel processing and more effective modeling of long-range dependencies [80]. Rather than processing sequences sequentially like RNNs and LSTMs, transformers allow all positions in a sequence to be processed simultaneously, dramatically improving

computational efficiency while maintaining or improving model performance.

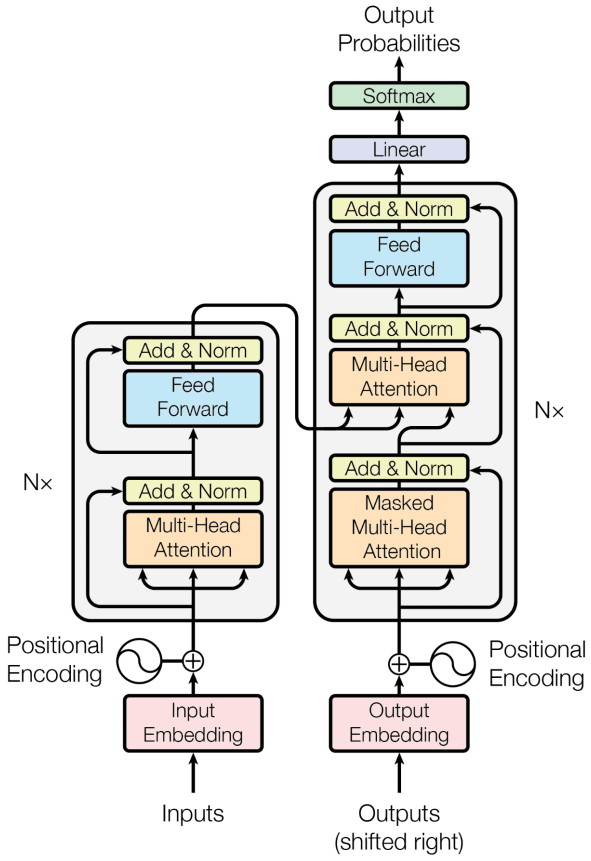


Figure 2: Schematic representation of the Transformer architecture. Source: [80]

The fundamental innovation of transformers lies in their use of attention mechanisms as the primary building block for sequence modeling. Where traditional approaches rely on recurrent connections to maintain information about sequence history, transformers use attention to directly model relationships between any pair of positions in a sequence, regardless of their distance. This design choice enables better modeling of long-range dependencies while eliminating the sequential bottleneck that limits the parallelization of recurrent models [80].

Attention mechanisms represent a powerful paradigm for allowing models to dynamically focus on different parts of input data when making predictions. The core insight behind attention is that not all

parts of the input are equally relevant for producing a particular output, and the model should be able to learn which parts deserve more focus. The modern understanding of attention mechanisms centers around the query–key–value framework, which provides an elegant abstraction for thinking about how attention works. In this framework, we have three types of vectors: queries represent what we're looking for, keys represent what's available to look at, and values represent the actual information we want to retrieve. The attention mechanism works by computing compatibility scores between each query and all available keys. These scores determine how much attention to pay to each key–value pair when constructing the output. The process begins by computing a compatibility function that measures how well each key matches the current query. Common approaches to measuring this compatibility include additive methods that use learned neural networks, multiplicative methods that compute dot products, and scaled versions that normalize the results.

Scaled dot-product attention represents the core attention mechanism used in transformer architectures. This approach computes attention over sets of queries, keys, and values simultaneously, enabling efficient parallel processing while maintaining the flexibility to model complex relationships between sequence positions. The scaled dot-product attention mechanism operates on three matrices: a query matrix Q containing all queries, a key matrix K containing all keys, and a value matrix V containing all values. Rather than computing attention for individual query–key–value triplets, this approach processes entire matrices simultaneously, enabling efficient implementation on parallel hardware.

Formally:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

This formulation computes all query–key compatibilities through a single matrix multiplication $QK^\top$, applies the scaling factor $\frac{1}{\sqrt{d_k}}$, normalizes through softmax, and computes the final weighted combination of values through multiplication with V.

Multi-head attention extends the basic attention mechanism by computing attention in multiple representation subspaces simultaneously. This extension allows the model to attend to information from different positions and capture different types of relationships within the same layer [80]. The fundamental insight behind multi-head attention is that different types of relationships may require different represen-

tation spaces to be effectively captured. For example, syntactic relationships between words might be best captured in one representation subspace, while semantic relationships might be better modeled in a different subspace. By computing attention in multiple heads, the model can specialize each head to capture different aspects of the input relationships. The attention mechanism can be applied in different configurations depending on the source of queries, keys, and values. Two primary configurations have emerged as fundamental building blocks in transformer architectures: self-attention and cross-attention.
Self-attention occurs when queries, keys, and values all come from the same sequence. This setup allows each position in the sequence to attend to all positions in the same sequence, enabling the model to capture complex relationships and dependencies within the input. Each position can directly access information from any other position in the sequence, creating a fully-connected graph where edge weights are determined by attention scores.
Cross-attention occurs when queries come from one sequence while keys and values come from a different sequence. This configuration is particularly important in encoder–decoder architectures, where the decoder needs to attend to the encoder's output while generating the target sequence. The queries come from the decoder's current state, while the keys and values come from the encoder's output. This arrangement enables the decoder to selectively focus on relevant parts of the input sequence as it generates different parts of the output [80]. Cross-attention is essential for tasks that involve mapping between different sequences, such as machine translation, summarization, and question answering. It provides a flexible mechanism for the model to determine which parts of the input are most relevant for generating each part of the output. Moreover, subsequent works have recognized the computational and memory limitations of the standard (dense) attention mechanism, and have proposed modifications or sparse / hybrid attention techniques to address these. For example, the Long-Short Transformer introduces a hybrid of long-range and local attention to reduce complexity for long sequences [81], and Cluster-Former uses clustering to yield sparse attention across chunked sequences to better encode long-range dependencies in QA settings [82].

3.2.4 *Transformer-Based Approaches*

The advent of transformer architectures, particularly BERT (Bidirectional Encoder Representations from Transformers) [83], marked an-

other paradigm shift in NER research. BERT's bidirectional training approach allows it to capture context from both directions simultaneously, making it particularly well-suited for sequence labeling tasks like NER.

Several variants of BERT have been developed to improve general language understanding and downstream task performance across diverse domains. **RoBERTa** [84]: Optimizes BERT's training procedure by removing the Next Sentence Prediction task and using longer training with larger batches. For NER tasks, RoBERTa often outperforms BERT due to its more robust pre-training.

DeBERTa [85]: Introduces disentangled attention mechanisms that separately encode content and position information. This architectural improvement leads to better performance on NER tasks, particularly for entities where positional information is crucial.

ELECTRA [86]: Uses a replaced token detection objective instead of masked language modeling, leading to more efficient pre-training and competitive NER performance with smaller computational requirements.

The success of general-domain transformers led to the development of domain-specific variants that achieve superior performance in specialized fields:

Moreover, several variants of BERT have been developed specifically for improved NER performance: **BioBERT** [87] and **SciBERT** [88] have demonstrated impressive capabilities in biomedical and scientific domains by incorporating domain-relevant pre-training data. These models show significant improvements over general BERT on biomedical NER tasks, with BioBERT achieving state-of-the-art performance on several biomedical entity recognition benchmarks.

ClinicalBERT [89] focuses specifically on clinical text, pre-trained on electronic health records and clinical notes. This specialization leads to substantial improvements in recognizing medical entities and understanding clinical language patterns. **FinBERT** [90] models have been developed for financial domain NER, showing improved performance in recognizing financial entities, organizations, and market-specific terminology. **BERT-NER** [91] works in general domain recognizing Person, Location, Organization and Miscellaneous entity types. Finally, BERT-BiLSTM-CRF combinations have been extensively studied, with models achieving significant performance improvements across multiple domains [92], [93],[94], and languages [95] [96].

3.2.5 *Large Language Model-Based Methods for NER*

Large Language Models (LLMs) represent the most recent paradigm shift in neural NER, bringing unprecedented capabilities in few-shot and zero-shot learning scenarios [97], [98]. These models, characterized by their substantial number of parameters extending into tens or hundreds of billions, have fundamentally changed how we approach entity recognition tasks. Models such as GPT [99], BloomZ [100], and LLaMA [101] exemplify this category. While originally designed for text generation, their application to sequence labeling tasks like NER has required innovative adaptations. The fundamental challenge in applying LLMs to NER lies in the mismatch between their generative nature and the structured prediction requirements of entity recognition. To address this gap, several innovative approaches have emerged: This work [102] introduced GPT-NER, which reformulates sequence labeling as a text generation task. Instead of traditional BIO tagging, entities are marked using special tokens. For example, a location entity "Paris" is transformed into "Paris## is a city," where "" and "##" denote entity boundaries. This approach achieves performance comparable to fully supervised baselines while excelling in low-resource and few-shot scenarios.
Ashok at al. [103] developed PromptNER, which leverages entity type definitions within prompts to enable few-shot learning. The model generates entity lists with explanations, achieving state-of-the-art performance on few-shot NER benchmarks across various datasets. This approach demonstrates the power of natural language instructions in guiding LLM behavior for structured tasks. A research by De Toni et al. [104] explored the zero-shot abilities of the To multitask model [105] for NER —a prompt-based LLM developed as part of the BigScience project for open research — , particularly in historical texts. While showing potential for languages lacking labeled datasets, the study revealed challenges in handling domain-specific terminology and inconsistent spelling patterns common in historical documents. Several sophisticated approaches have been developed to enhance LLM performance on NER tasks: Zhu et al. [106] introduced GL-NER, a generation-aware LLM specifically designed for few-shot NER. The model employs novel prompt templates with label-injected instructions and uses masking-based loss optimization. This approach can generate entity names or signal "does not exist" when no entity is present, significantly improving few-shot learning performance over traditional prompt-based methods. Other works [107] explored combining different LLMs for en-

tity disambiguation in dialogue systems. Their approach uses GPT-2 as a generator during training and BERT for evaluation during inference, effectively addressing ambiguity and entity comparison in real-time dialogue scenarios.
This paper [108] developed UniversalNER through targeted distillation from LLMs, creating a model with broad coverage of entity types suitable for clinical applications. Similarly, another approach [109] introduced a self-improving zero-shot framework that leverages unlabeled corpora for continuous improvement through self-annotated pseudo-demonstrations. For the use of an LLM to enhance the context, this work [110] developed CALM, which generates additional context for entities offline using LLMs, particularly useful in low-resource settings where annotated data is limited. Through prompt engineering and self-improvement techniques, LLMs have demonstrated promising results across a wide range of applications. However, their high computational and memory requirements make them costly and often inaccessible for smaller-scale implementations. In addition, the heavy reliance on prompt engineering introduces a degree of unpredictability and limits scalability. LLMs also tend to struggle with intricate cases that demand deep, domain-specific knowledge unless extensively fine-tuned on specialized data.
The integration of LLMs into NER marks a significant advancement in the field, especially for scenarios with scarce training data. Nevertheless, their substantial resource demands and sensitivity to prompt design underscore the need for continued research into more efficient and robust methodologies. Future work will likely focus on hybrid models that combine the rich contextual understanding of LLMs with the efficiency and reliability of traditional neural architectures.

3.2.6 *Domain Adaptation in Named Entity Recognition*

Domain adaptation represents a critical challenge in NER systems when shifting from source to target domains with different entity distributions and vocabularies.
Early work in domain adaptation for NER primarily focused on rule-based mappings and simple statistical approaches [111], [112], [113], [24]. However, the emergence of neural approaches has dramatically transformed the research landscape. Methods employing noise-contrastive estimation align semantically similar concepts in the embedding space while using regularization techniques to preserve original semantic knowledge [114]. Word embedding techniques have proven particu-

larly effective for domain adaptation. Fersini et al. [115] explored domain adaptation specifically for NER systems, showing that carefully constructed word embeddings can effectively bridge the gap between source and target domains by capturing underlying semantic relationships. A common practice consists in transferring pre-trained word embeddings to downstream tasks, for example, by training medical-specific embeddings and applying them to NER, with several techniques from concatenation to fine-tuning being explored.
Recent advances have leveraged self-alignment and contrastive learning techniques to improve cross-domain adaptation, with approaches like self-alignment pretraining specifically designed for biomedical entities [116]. Semi-supervised learning [117], [118] has emerged as another powerful paradigm for domain adaptation, particularly beneficial when limited labeled data is available in target domains.
Although the above-mentioned investigations represent a fundamental step towards the adaptation of NER systems, one main issue remains open: most of the available approaches are often computationally expensive, requiring resources that lead to high energy consumption, making them costly and less sustainable for real-world applications.

3.3 RELATION EXTRACTION

This section provides a comprehensive overview of the evolution and current state of relation extraction (RE) techniques, with particular emphasis on approaches leveraging language models. We organize our discussion around the foundational concepts, methodological advances, and contemporary challenges that define the current landscape of relation extraction research. Figure 3 represents an example of Relation Extraction task that link through a semantic relationship the entities found based on the text. Relation extraction, as a fundamen-

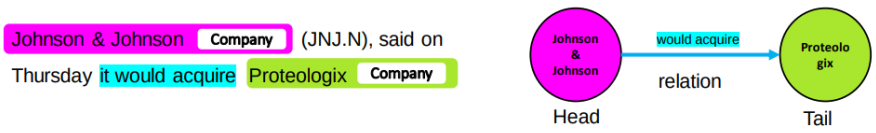


Figure 3: Example of Relation Extraction task.

tal task in natural language processing, aims to identify and classify semantic relationships between entities mentioned in text [119]. The field has undergone significant evolution from early rule-based systems to sophisticated neural architectures that leverage the power of

pre-trained language models. Early approaches to relation extraction predominantly relied on pipeline architectures that decomposed the task into sequential stages: Named Entity Recognition (NER), Relation Identification, and Relation Classification. These pipeline systems, while conceptually straightforward, suffered from error propagation issues where mistakes in entity recognition adversely affected downstream relation classification accuracy [120, 121].
The transition from traditional pipeline approaches to joint modeling represents a significant paradigm shift in relation extraction research. Joint models attempt to simultaneously perform entity recognition and relation classification, thereby capturing the inherent interdependencies between these tasks [122]. This architectural evolution has proven particularly beneficial in mitigating error propagation while improving overall system performance, as demonstrated by recent comparative studies that show joint approaches consistently outperforming pipeline methods across various metrics and datasets. In this context, recent work has explored hybrid joint architectures that combine Large Language Models with structured neural components; for example, [123, 124] propose joint entity and relation extraction frameworks that integrate LLM-based representations with Graph Convolutional Networks or other joint learning mechanisms to better model structural dependencies between entities and relations.

3.3.1 *Benchmark Datasets and Evaluation Frameworks*

The development of standardized benchmark datasets has been crucial for advancing relation extraction research. The CoNLL04 dataset [125] remains widely used for evaluating relation extraction systems, containing annotations for entities and relations including *Live_In*, *Located_In*, *OrgBased_In*, *Kill*, and *Work_For*. Recent surveys [122] identify TACRED [126], DocRED [127], and FewRel [128] as the most prevalent benchmark datasets, corresponding to sentence-level, document-level, and few-shot relation extraction paradigms respectively. TACRED serves as the primary benchmark for traditional sentence-level approaches, providing a robust testbed for comparing different methodologies in controlled settings, while DocRED has become instrumental for evaluating document-level methodologies that require understanding of complex inter-sentential relationships and multi-hop reasoning capabilities.
FewRel has emerged as the standard for few-shot evaluation, enabling assessment of model generalization capabilities across diverse relation

types and providing crucial insights into the adaptability of different architectural approaches when faced with limited training data. The diversity of evaluation datasets reflects the broad applicability of relation extraction techniques across various domains, including biomedical literature, social media content, financial documents, and legal texts [122]. This diversity underscores the importance of developing robust and generalizable relation extraction methodologies that can maintain performance across different textual genres, linguistic styles, and domain-specific vocabularies while adapting to the unique challenges presented by each application area.

In addition to these widely used corpora, several further benchmark datasets enrich the evaluation landscape. For example, SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals provides a widely-adopted sentence-level dataset for semantic relation extraction across nine directional relations and an "Other" relation [129]. Similarly, the ADE (Adverse Drug Events) corpus offers a domain-specific biomedical benchmark focusing on drug–adverse-effect relations, thereby highlighting the challenge of fine-grained relation extraction in high-specialty sub-domains [130, 131]. Together, this benchmark corpus portfolio—from general-domain sentence-level tasks to document-level and domain-specialized challenges—provides the foundation for rigorous evaluation of relation extraction systems. It emphasizes the dual need for high performance in controlled settings and robust generalizability to domain-specific, few-shot, and cross-sentence extraction scenarios.

3.3.2 *Deep Learning Revolution in Relation Extraction*

The advent of deep neural networks marked a transformative period in relation extraction research. Pre-trained language models, particularly BERT [132] and its variants, have emerged as the cornerstone of modern relation extraction systems. LUKE [133] introduced entity-aware self-attention mechanisms that explicitly model entity representations alongside contextual token representations, achieving substantial improvements on relation extraction benchmarks. The transformer architecture has fundamentally reshaped relation extraction methodologies, with BERT's bidirectional training objective, which predicts masked tokens based on both left and right context, aligning naturally with the relational reasoning required in extraction tasks [122].

Research has shown that BERT-based models achieve state-of-the-art performance across multiple relation extraction benchmarks, with RoBERTa

demonstrating particular effectiveness in document-level relation extraction scenarios [122]. Most past approaches focus on pre-trained transformer models such as BERT[134] to downstream RE tasks [135]. Recent advances in RE have been driven by deep neural networks, with large pre-trained language models achieving state-of-the-art performance. However, despite these advances, several fundamental challenges persist in real-world deployment scenarios. The primary limitation stems from the long-tail distribution of relations in natural datasets. While frequent relations benefit from abundant training examples, the majority of relations suffer from severe data scarcity. This creates a significant bottleneck since deep learning approaches require substantial labeled corpora resources that are often unavailable in low-resource settings [136]. This limitation becomes particularly problematic for semantically complex tasks like RE, which require deep domain-specific knowledge and robust generalization capabilities. Recent investigations have explored the application of generative approaches to relation extraction, where REBEL [137] employs an autoregressive model based on BART [138] that generates relation triplets using structured linearization formats. This generative paradigm has been successfully extended to multilingual contexts with mREBEL [139], demonstrating the versatility of sequence-to-sequence architectures for relation extraction tasks and establishing a new direction that treats relation extraction as a text generation problem rather than traditional classification.

A substantial body of research has investigated the incorporation of external knowledge to enhance relation extraction performance, particularly in low-resource scenarios. Li et al. [140] proposed knowledge-attention encoders that integrate prior knowledge from external lexical resources such as FrameNet and Thesaurus.com into deep neural networks, while Gao et al. [141] introduced enriched sentence-level representations by incorporating both structured knowledge from external knowledge graphs and semantic knowledge derived from textual corpora. More recently, ontology-guided validation approaches have been proposed to improve the reliability of extracted relations; [142, 143] introduce methods that combine ontological constraints with LLM-based validation to verify and correct predicted relations, resulting in improved precision, semantic consistency, and domain-specific adaptation. However, external knowledge sources present inherent challenges, including noise, incompleteness, and potential misleading information that may not align with specific contextual and domain requirements.

3.3.3 *Large Language Models*

Large Language Models (LLMs) represent perhaps the most transformative development in artificial intelligence in recent decades. Rather than requiring task-specific architectures and training procedures, LLMs demonstrate that a single model trained on diverse text can exhibit remarkable competence across an extraordinary range of natural language tasks [144]. The revolution brought by LLMs extends beyond mere performance improvements. These models have introduced entirely new paradigms for human-AI interaction, where complex tasks can be accomplished through natural language prompts rather than traditional programming interfaces. This shift has democratized access to AI capabilities, enabling users without technical expertise to harness sophisticated language understanding and generation for their specific needs.

Large models can learn to perform new tasks simply by observing examples within their input context, without any parameter updates. This capability effectively turns the model's forward pass into a form of rapid adaptation, enabling few-shot learning across diverse domains. Chain-of-thought reasoning allows models to solve complex problems by breaking them down into intermediate steps, much like human problem-solving strategies. Rather than attempting to jump directly to answers, models can articulate their reasoning process, leading to dramatic improvements on mathematical, logical, and commonsense reasoning tasks. The instruction following enables models to understand and execute complex natural language instructions, even for tasks they haven't explicitly seen during training. This capability bridges the gap between human intent and machine execution, making LLMs far more versatile and user-friendly. Modern LLM training typically progresses through multiple phases, each serving specific purposes in developing the model's capabilities. The foundation of LLM training rests on self-supervised learning objectives that enable models to learn from vast amounts of unlabeled text. Causal language modeling trains models to predict the next token in a sequence given all previous tokens, following the objective:

$$\mathcal{L}_{CLM} = -\sum_{t=1}^{T} \log P(x_t | x_{<t}; \theta)$$

This autoregressive approach naturally enables text generation capabilities while fostering deep understanding of language structure and content. The simplicity of this objective belies its power—by learn-

ing to predict the next word across billions of text snippets, models develop remarkably sophisticated understanding of syntax, semantics, and world knowledge.

3.3.4 *Large Language Models for Relation Extraction*

The emergence of Large Language Models has introduced new paradigms for relation extraction research. Pan et al. [145] provide a comprehensive analysis of the integration between LLMs and knowledge graphs, highlighting the potential of LLMs to function as dynamic knowledge bases with superior contextual understanding capabilities. Unlike static external knowledge resources, LLMs demonstrate remarkable ability to capture implicit relationships, resolve ambiguities, and interpret complex linguistic phenomena including metaphors, negations, and conditional statements [122]. Recent studies [146], [122] have shown that LLMs, particularly models like T5, Flan-T5 and GPT variants, exhibit promising performance in few-shot relation extraction scenarios, demonstrating their capacity to identify previously unseen relations with minimal training examples. Beyond direct inference, recent research has investigated distillation-based strategies to transfer LLM-generated knowledge into less resource consumption extraction models. Additionally, automatic prompt design strategies, such as PSOPL [147], have been shown to optimize LLM prompts for medical information extraction tasks, improving performance and reducing human effort. Zero-shot relational triplet extraction has also benefited from LLM generalization through multi-task data fusion [148], where generation and understanding tasks are integrated to improve performance on unseen relations. Despite their promise, LLMs contribute only approximately 25% to overall state-of-the-art results in relation extraction [122]. This disparity suggests that while LLMs excel in generative and few-shot scenarios, traditional transformer models like BERT and RoBERTa remain more effective for standard relation extraction benchmarks. The phenomenon can be attributed to the inherent suitability of BERT-like models for downstream classification tasks, whereas LLMs demonstrate their strength in tasks requiring extensive contextual generation and reasoning with limited examples. Despite these advances, the potential of LLMs to support and improve downstream RE remains largely underexplored. Given their demonstrated utility, further investigation into their integration with RE workflows is both timely and necessary.

Part II

NAMED ENTITY RECOGNITION

4

IMBALANCE LEARNING IN NAMED ENTITY RECOGNITION

This chapter investigates the pervasive issue of class imbalance in Named Entity Recognition systems, a fundamental problem that significantly affects model performance across different entity types. Given such scenario, the investigation evaluates multiple rebalancing strategies applied to widely-used English NER benchmarks, specifically OntoNotes 5.0 and CoNLL-2003, to discuss which approaches best mitigate performance degradation on minority entity classes.

RESEARCH QUESTIONS

RQ 4.1 - How can we effectively address class imbalance in Named Entity Recognition tasks?

4.1 METHODOLOGY

Class imbalance represents a persistent challenge in sequence labeling tasks, where certain entity types are significantly underrepresented compared to others. In NER datasets, this manifests as a predominance of non-entity tokens (labeled as "O" in BIO tagging schemes shown in Table 1) and common entity types such as PERSON, LOCATION, and ORGANIZATION, while fine-grained entities like WORK_OF_ART, LANGUAGE, or ORDINAL appear infrequently. This section presents a comprehensive approach to addressing class imbalance through multiple complementary strategies, Figure 4 show common state-of-the art techniques and their outputs on a sample sentence.

4.1.1 *Resampling Techniques*

Resampling techniques modify the training data distribution to providea more balanced representation across entity classes. Our approach encompasses both oversampling and undersampling strategies, implemented at different granularities.

	Instance											
None	She	did	not	complain	of	headache	or	any	other	neurological	symptoms	.
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O
LwTR	L.	One	not	complain	of	headache	he	any	interatrial	neurological	current	.
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O
SR	She	did	non	complain	of	headache	or	whatsoever	former	neurologic	symptom	.
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O
MR	She	did	not	complain	of	neuropathic	pain	syndrome	or	acute	pulmonary	disease	.
	O	O	O	O	O	B-problem	I-problem	I-problem	O	B-problem	I-problem	I-problem	O
SiS	not	complain	She	did	of	headache	or	neurological	any	symptoms	other	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	

Figure 4: Original training instance and different types of augmented instances. Author highlight changes using blue color. Note that LwTR (Label-wise token replacement) and SiS (Shuffle within segments) change token sequence only, whereas SR (Synonym replacement) and MR (Mention replacement) may also change the label sequence. source [149]

Random Oversampling (ROS)

Random oversampling increases the representation of minority classes by duplicating existing instances. For NER tasks, we implement a token-level strategy, the approach follows a personalized pipeline where:

1. A subset of underrepresented entity types is selected based on frequency analysis and F1-score performance on the validation set
2. The identified subset includes: `I-ORDINAL`, `I-LAW`, `I-NORP`, `I-EVENT`, `I-WORK_OF_ART`, `B-FAC`
3. Tokens labeled with these entity types are randomly duplicated within their original sentence context
4. The resampling ratio is determined empirically to avoid overfitting while improving minority class representation

While this approach may occasionally disrupt sentence semantics due to its token-level nature, it provides benefits for the directly resampled I-tags.

Random Undersampling (RUS)

Random undersampling addresses class imbalance by reducing the majority class representation. In NER tasks, the "O" (outside entity) tag typically constitutes the overwhelming majority. Our implementation

uses the `RandomUndersampler` from the imbalanced-learn library with a majority strategy, focusing exclusively on reducing "O" tokens while preserving all entity-labeled tokens.

The strategy parameters include:

- **Target class**: "O" (outside entity)
- **Sampling strategy**: majority (resample only the majority class)
- **Preservation**: All entity-labeled tokens remain unchanged

Sentence-Level Resampling

To preserve semantic coherence, we implement sentence-level resampling strategies that compute importance scores for entire sentences based on their entity composition. This approach, inspired by [16], uses three scoring functions:

Smoothed Count (sC): The basic scoring function considers the total count of entity tokens in a sentence:

$$f_s^{sC} = 1 + \sum_{t \in T} c(t, s) \tag{1}$$

where T represents the set of all possible entity types except "O", and $c(t, s)$ denotes the count of tokens with entity type t in sentence s.

Smoothed Count with Rareness (sCR): This function incorporates entity type rareness, measured as self-information:

$$r_t = -\log_2 \frac{\sum_{s \in S} c(t, s)}{N} \tag{2}$$

$$f_s^{sCR} = 1 + \sum_{t \in T} r_t \cdot c(t, s) \tag{3}$$

where N is the total number of tokens in the training set.

Smoothed Count with Rareness and Density (sCRD): The most comprehensive scoring function includes sentence length normalization:

$$f_s^{sCRD} = 1 + \frac{\sum_{t \in T} r_t \cdot c(t, s)}{\sqrt{l_s}} \tag{4}$$

where l_s represents the length of sentence s. The square root normalization prevents excessive penalization of longer sentences.

4.1.2 *Data Augmentation*

Data augmentation techniques create synthetic training examples to improve model robustness and address class imbalance. Our approach focuses on preserving entity information while introducing lexical variation.

Synonym Replacement

We implement a sentence-level augmentation strategy using the Augmenty framework [150]. The process involves:

1. **Sentence Selection**: Identify sentences containing any of the ten least frequent entity labels: `I-LAW`, `B-EVENT`, `B-QUANTITY`, `B-PRODUCT`, `I-PRODUCT`, `I-NORP`, `B-LANGUAGE`, `B-LAW`, `I-LANGUAGE`, `I-ORDINAL`
2. **Synonym Augmentation**: Apply WordNet-based synonym replacement using the `wordnet_synonym_v1` augmenter
3. **Entity Preservation**: Ensure that entity-labeled tokens maintain their original labels in augmented sentences

Example transformation:

- **Original**: "Augmenty is a wonderful tool for augmentation."
- **Augmented**: "Augmenty is a fantastic mechanism for augmentation."

4.1.3 *Focal Loss Training*

Focal Loss addresses class imbalance by reformulating the standard cross-entropy loss to down-weight easy examples and focus training on hard, misclassified cases. Originally proposed by Lin et al. [151] for object detection tasks, we adapt this loss function to the token classification setting in NER.

The Focal Loss is defined as:

$$\mathcal{L}_{FL} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{5}$$

where:

- p_t is the model's estimated probability for the correct class
- $\gamma \geq 0$ is the focusing parameter that adjusts the rate at which easy examples are down-weighted

- α_t is the class-specific weighting factor

The focusing parameter γ smoothly adjusts the contribution of easy examples to the loss. When $\gamma = 0$, Focal Loss is equivalent to standard cross-entropy. As γ increases, the effect of the modulating factor $(1 - p_t)^\gamma$ increases, reducing the relative loss for well-classified examples. We use $\gamma = 2.0$ as the default value, following the original implementation.

Class Weight Computation: To address class imbalance, we employ balanced class weights α_t computed automatically from the training data distribution. The weight for each class i is calculated as:

$$\alpha_i = \frac{N_{total}}{N_{classes} \cdot N_i} \tag{6}$$

where N_{total} is the total number of labeled instances, $N_{classes}$ is the number of distinct classes, and N_i is the number of instances for class i. This inverse frequency weighting ensures that rare entity types (e.g., `LAW`, `EVENT`, `LANGUAGE`) receive proportionally higher weight during training, while the focusing parameter simultaneously prevents the model from being overwhelmed by easy negative examples, which typically dominate in NER datasets due to the prevalence of non-entity tokens. The combination of class balancing and difficulty-based weighting enables the model to learn effectively from both minority classes and challenging examples simultaneously.

Alternative Augmentation Strategies

Building on established NER augmentation techniques [152], other strategies presented in Figure 4 are:

Label-wise Token Replacement (LwTR): For each token, a binomial distribution determines replacement probability. If selected, the token is replaced with another token sharing the same label from the training distribution.

Mention Replacement (MR): Complete entity mentions are replaced with alternative mentions of the same entity type, preserving mention boundaries and label consistency.

Shuffle within Segments (SiS): Token sequences are segmented by label type, and tokens within each segment are randomly shuffled while maintaining label order.

4.2 EXPERIMENTAL SETUP

4.2.1 *Dataset and Configuration*

All experiments are executed on the OntoNotes 5.0 [17], and CoNLL03 [18] datasets. The first one contains 18 distinct entity types exhibiting significant class imbalance. The dataset's hierarchical structure ranges from common entities (PERSON, LOCATION, ORGANIZATION) to fine-grained categories (WORK_OF_ART, LANGUAGE, ORDINAL). The second one has a tagset composed of 4 coarse-grained entity types: PERSON, LOCATION, ORGANIZATION and MISCELLANEOUS.

BERT transformer model is finetuned for 10 epochs with BIO-scheme target representation for sequence-labeling and Spancategorizer (Spancat) component [153] from SpaCy [154] is trained for 35 epochs with span length restricted to 1, effectively implementing BIO-scheme token classification. A span categorizer consists of two parts: a suggester function that proposes candidate spans and a labeler model that predicts zero or more labels for each candidate [155]. This configuration enables direct comparison with traditional sequence labeling approaches while leveraging the span-based architecture's advantages.

4.2.2 *Evaluation Metrics*

Model performance is evaluated using three F1-score variants:

- **Micro F1**: Aggregates true positives, false positives, and false negatives across all classes
- **Macro F1**: Averages F1 scores across all classes, giving equal weight to each class
- **Weighted F1**: Averages F1 scores weighted by class support

These metrics provide complementary perspectives on model performance, with Macro F1 being particularly sensitive to minority class performance.

4.2.3 *Experimental Results*

Tables 2 and 3 presents comprehensive results across all implemented techniques.

The effectiveness of different techniques varies substantially between SpanCat and BERT models. For SpanCat, the smoothed count (sC)

Table 2: Performance comparison of imbalance learning techniques on OntoNotes 5.0. **Bold** indicates the best configuration versus model baseline.

Method	Micro F1	Macro F1	Weighted F1
SpanCat	0.8186	0.6668	0.8157
SpanCat (RUS)	0.2179	0.2169	0.2936
SpanCat (ROS)	0.8210	0.6798	0.8221
SpanCat (sC)	**0.8349**	**0.6818**	**0.8371**
SpanCat (sCR)	0.8250	0.6761	0.8280
SpanCat (sCRD)	0.8149	0.6705	0.8181
SpanCat Synonym Replacement	0.8259	0.6743	0.8247
Spancat Focal Loss	0.8112	0.6785	0.8123
BERT	0.8957	0.7819	0.8953
BERT (RUS)	0.2184	0.0487	0.2184
BERT (ROS)	**0.9045**	**0.7910**	**0.9045**
BERT (sC)	0.8981	0.7748	0.8970
BERT (sCR)	0.8768	0.7545	0.8752
BERT (sCRD)	0.8990	0.7825	0.8980
BERT Synonym Replacement	0.9039	0.7893	0.9037
BERT Focal Loss	0.8477	0.7136	0.8561

strategy achieves the greatest improvements on OntoNotes 5.0, achieving the best performance across all metrics. However, on CoNLL03, random oversampling (ROS) and synonym replacement show competitive or superior results, suggesting that the optimal strategy depends on dataset characteristics.

For BERT models, random oversampling emerges as the dominant strategy across both datasets, consistently achieving the best micro and macro F1 scores. This represents a notable divergence from SpanCat results and suggests that BERT's contextual representations benefit more from simple data duplication than from sophisticated resampling strategies. Random undersampling produces catastrophic failures across all configurations, with performance dropping below 0.22 on OntoNotes and 0.12 on CoNLL03 for both models. This consistent pattern confirms that majority class tokens, particularly the "O" tag, provide essential contextual information for entity boundary detection.

Focal loss training shows inconsistent results, improving SpanCat per-

Table 3: Performance comparison of imbalance learning techniques on CoNLL03. **Bold** indicates the best configuration versus model baseline.

Method	Micro F1	Macro F1	Weighted F1
SpanCat	0.7558	0.7428	0.7565
SpanCat (RUS)	0.1994	0.2505	0.3048
SpanCat (ROS)	**0.7723**	**0.7485**	0.7746
SpanCat (sC)	0.7685	0.7384	0.7728
SpanCat (sCR)	0.7708	0.7370	0.7753
SpanCat (sCRD)	0.7721	0.7405	0.7748
SpanCat Synonym Replacement	**0.7770**	**0.7467**	**0.7783**
SpanCat Focal Loss	0.7761	0.7462	0.7795
BERT	0.9149	0.8828	0.9158
BERT (RUS)	0.1022	0.0967	0.1250
BERT (ROS)	**0.9215**	**0.8928**	**0.9221**
BERT (sC)	0.9021	0.8673	0.9023
BERT (sCR)	0.8916	0.8588	0.8920
BERT (sCRD)	0.9075	0.8737	0.9078
BERT Synonym Replacement	0.9130	0.8833	0.9131
BERT Focal Loss	0.9065	0.8669	0.9085

formance on CoNLL03 but degrading BERT performance on both datasets. This suggests that focal loss may be better suited to simpler models or that the hyperparameters require careful tuning for transformer architectures. The comparison between sCR and sCRD across both models reveals that density normalization consistently underperforms relative to simple rareness weighting. This pattern suggests that longer sentences containing rare entities should not be penalized, as they likely provide valuable contextual information that aids generalization. The decreased performance with sCRD indicates that sentence length is not a reliable proxy for redundancy or noise in entity recognition tasks.
Synonym replacement demonstrates modest but consistent improvements for SpanCat, particularly on CoNLL03 where it achieves competitive or best weighted F1 scores. However, its impact on BERT is more limited, suggesting that lexical variation matters less when the model already has access to rich contextual embeddings.

4.3 DISCUSSION

The experimental results highlight a fundamental trade-off between model complexity and the effectiveness of rebalancing strategies. While SpanCat benefits from carefully designed sentence-level techniques like smoothed count, BERT's superior performance with simple random oversampling suggests that pre-trained contextual representations reduce the need for elaborate augmentation schemes.
Random undersampling failed catastrophically in every configuration, highlighting that this strategy fundamentally disrupts the compositional structure of sentences on which the model depends.
The divergent behavior of focal loss between SpanCat and BERT hints at deeper interactions between loss function design and model capacity. The degraded BERT performance suggests that focal loss may inadvertently interfere with the fine-tuning dynamics of pre-trained models, potentially disrupting learned representations. This warrants an investigation into loss functions specifically designed for transfer learning scenarios rather than adapting techniques from computer vision domains.

RQ 4.1 - How can we effectively address class imbalance in Named Entity Recognition tasks?

The choice of the optimal strategy is strongly conditioned on the underlying model architecture. Sentence-level resampling methods, such as Smoothed Count, yield the best results when applied to span-based architectures like SpanCategorizer, whereas for transformer-based models leveraging pre-trained representations, simple Random Oversampling proves to be an effective and adequate solution.

A **limitation** of this investigation is the focus on two English datasets with specific entity type distributions, leaving open questions about generalization to morphologically rich languages or domain-specific corpora with extreme imbalance ratios. Future work should investigate the interaction between data augmentation and few-shot learning paradigms, particularly whether contrastive learning objectives could better leverage augmented examples for rare entity types. Furthermore, hybrid strategies combining balanced undersampling with sentence-level resampling offer potential for addressing both efficiency and effectiveness concerns simultaneously. Multi-stage approaches that apply different strategies to distinct entity types could provide more

nuanced solutions, while adaptive resampling mechanisms that adjust strategies based on entity frequency distributions during training show promise for dynamic optimization. While the imbalance learning techniques explored in this chapter—ranging from sentence-level resampling to focal loss and synonym replacement—successfully improve performance on minority entity classes, they expose a broader challenge in resource-efficient NER: the computational and data costs required to achieve robust performance. Random oversampling, despite proving most effective for BERT models, increases dataset size and training time proportionally to the resampling rate, demanding additional GPU resources and energy consumption. More sophisticated techniques like smoothed count resampling compound these costs by requiring complex probability calculations and iterative sampling procedures, while synonym replacement necessitates external linguistic resources and additional processing overhead. Furthermore, the catastrophic failure of undersampling approaches eliminates the most computationally attractive option, forcing practitioners toward data expansion strategies that scale poorly to production environments. These resource constraints become even more acute in domain adaptation scenarios, where the imbalance patterns observed here (minority entity classes requiring augmentation to achieve adequate performance) intersect with the fundamental scarcity of labeled data in new domains. An organization deploying NER across multiple specialized domains faces a multiplicative resource burden: not only must each domain's class imbalance be addressed through data augmentation, but entirely new labeled datasets must be created and models retrained from scratch. The computational expense of repeatedly training transformer models like BERT, combined with the annotation costs for domain-specific corpora, creates a significant barrier to scalable NER deployment.

Chapter 5 addresses this challenge by proposing an approach that extracts and reuses rich feature representations from source domain models, enabling target domain entity recognition under severe resource constraints where both computational budget and labeled training data are limited.

5

CROSS-DOMAIN NAMED ENTITY RECOGNITION

Traditional NER systems operate under the assumption that training and test data share the same feature space and underlying distribution. However, when adapting a NER system to a new target classification schema, this assumption may not hold. Since new data may exhibit different representation characteristics and follow distinct distributions compared to the source domain, the absence of appropriate training sets poses significant challenges regarding human effort, resource allocation, and time investment [156].
The challenges of cross-domain adaptation in NER are particularly pronounced in real-world applications where entities may be defined differently across domains, datasets may exhibit varying linguistic characteristics, and computational resources for model adaptation are often limited. Traditional approaches to domain adaptation typically require substantial amounts of labeled data in the target domain and computationally expensive fine-tuning procedures, which can be prohibitive in many practical scenarios.
To address these limitations, this chapter presents an efficient supervised approach that aims to transfer knowledge from a source domain to a target domain for Named Entity Recognition purposes. The proposed method enables learning target domain labels and distributions by exploiting token-level features derived from transformer-based models pre-trained on source domains, including attention patterns, hidden states, and probability vectors. In particular, by leveraging transformer model characteristics across different tagsets, satisfactory results can be achieved using only 10% of the available training set.
The key contributions of this chapter include:

- A feature extraction framework that captures rich semantic information from pre-trained transformers
- A heterogeneous ensemble methodology based on Bayesian Model Averaging that effectively combines multiple lightweight classifiers

- Comprehensive evaluation across six distinct transfer scenarios demonstrating competitive performance with significantly reduced computational requirements

This approach provides a computationally efficient alternative to fine-tuning-based transfer learning methods, while maintaining competitive performance across diverse domain adaptation scenarios.

RESEARCH QUESTIONS

RQ 5.1 - How can features obtained from a source domain be leveraged to learn target domain labels and distributions efficiently, achieving performance close to fine-tuned transformers but with lower computational cost?

RQ 5.2 - Which representations matter most to deal with the problem of transfer learning for NER?

5.1 PROBLEM FORMULATION AND METHODOLOGY

Transfer learning in Named Entity Recognition involves adapting a model trained on a source domain S with entity types Y_S to a target domain T with potentially different entity types Y_T. Traditional approaches typically require substantial labeled data in the target domain and computationally expensive fine-tuning. In this work, we address a challenging scenario characterized by:

1. **Limited Target Domain Data**: Labeled data in the target domain is scarce, representing realistic conditions where annotation is expensive or time-consuming
2. **Computational Constraints**: Computational resources for fine-tuning are minimal, making traditional approaches impractical
3. **Schema Heterogeneity**: Entity types and distributions differ significantly between source and target domains, requiring sophisticated mapping strategies

In particular, we verify the hypothesis that trivial token-level representations from transformer models pre-trained on a source domain contain sufficient information to accurately recognize entities in a target domain, requiring only a small portion of labeled target data for

training classifiers with limited computational resources. To this purpose, we aim at creating a feature space, without the necessity of the expensive costs of fine-tuning, that is able to characaterize entities in the source domain and to learn the relationship with the target domain.

5.1.1 *Input Space Representation*

In order to reuse as much as possible the knowledge about the language of existing pre-trained models, therefore reducing the energy consumption costs required to fine-tune them, we extracted the following token representations. In particular, a token in a given sentence can be represented as follows:

- **Contextualized Hidden States** — represent the deep semantic and syntactic properties of each token, capturing the influence of its surrounding context across multiple layers.
- **Attention Scores** — model the internal dependency structure among tokens, indicating how contextual information is distributed and how each token contributes to the representation of others within the sequence.
- **Source Probability Distribution** — reflects the model's output confidence over possible token classes, encoding domain-specific priors and uncertainty patterns learned during pre-training.

As shown in Figure 5, these features are extracted from the pre-trained transformer architecture to create a rich representation space.

Contextualized Hidden States

The contextualized hidden states represent the core semantic encoding produced by transformer models. These representations capture rich contextual information about each token based on its surrounding linguistic context. For a token i, we derive a vector $h_i \in \mathbb{R}^d$, where d is the hidden dimension of the transformer model. Hidden states representations are taken from the last layer of the pre-trained transformer architecture used. These representations have been shown to capture hierarchical linguistic information, from low-level morphological features to high-level semantic relationships, making them particularly valuable for cross-domain transfer.

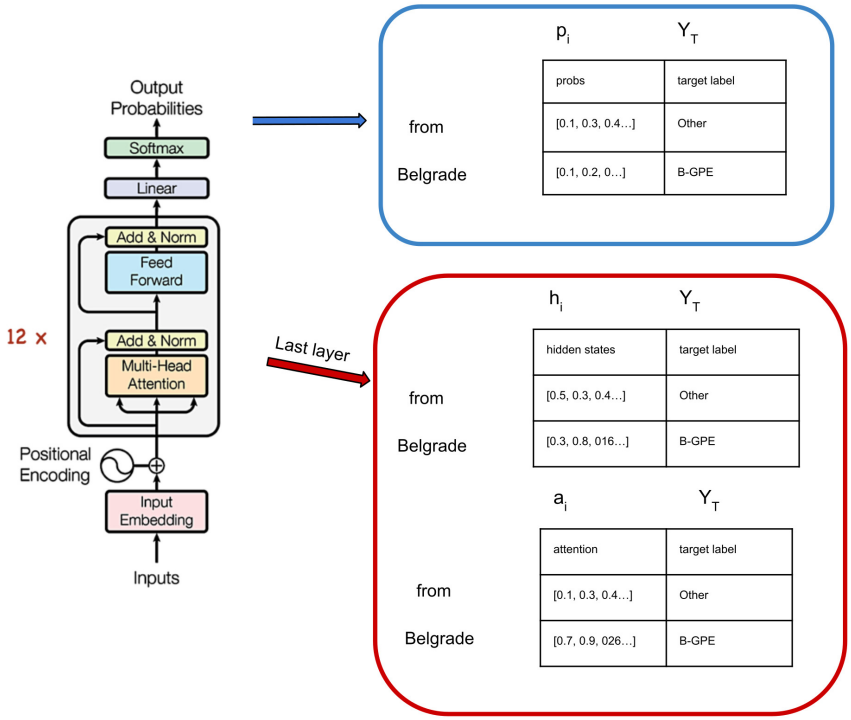


Figure 5: Input space features extraction from pre-trained BERT architecture. Output layer is used to get Source Probability Distribution, while last layer provide Contextualized Hidden Stasets and Attention Scores.

Attention Scores

The attention mechanism in transformer models provides explicit information about which parts of the input sequence are most relevant for processing each token. We extract attention weights from the final layer of the pre-trained language model to capture these relevance patterns.

For each token i, we derive an attention tensor $A_i \in \mathbb{R}^{m \times n \times n}$, where m is the number of attention heads and n is the sequence length. Each tensor A_i, combining attention values obtained in head, is reshaped into a one-dimensional vector a_i to create an attention-based feature representation. By incorporating attention patterns, we can leverage the model's learned focus mechanisms to improve entity recognition accuracy across domains.

Source Probability Distribution

The third component of our feature representation leverages the confidence and uncertainty information from the source domain model. For each token i, we extract the probability distribution across the source tagset.
For a token i, we derive a vector $p_i \in \mathbb{R}^{|Y_S|}$, where Y_S is the source tagset. The inclusion of probability distributions allows our approach to leverage not just the hard predictions from the source model, but also the uncertainty and confidence patterns that provide additional information for cross-domain adaptation.

5.1.2 *Learning Objective and Feature Integration*

The three feature types introduced above enable us to create a comprehensive input space for representing each entity mention. This representation can subsequently be used to learn the mapping from predicted source types $y_s \in Y_S$ to target types $y_t \in Y_T$.
Let $E_S = \{e_{s_1}, e_{s_2}, \ldots, e_{s_n}\}$ be a set of entity mentions annotated according to a source tagset Y_S, and let M_S be a pre-trained language model underlying a NER system trained on the source domain. Similarly, let $\Psi_T = \{\psi_{t_1}, \psi_{t_2}, \ldots, \psi_{t_r}\}$ be a (small) set of entity mentions available for training in the target domain under the tagset Y_T.
Rather than fine-tuning a pre-trained model using Ψ_T, our main objective is to train shallow classifiers on the previously defined input space to learn the mapping from source domain predictions to target domain labels. Given a token i, we represent it using its contextualized hidden states vector h_i, attention score vector a_i, and source probability distribution vector p_i.
In its general form, the input space X for a token i is denoted as:

$$X_i = h_i \oplus a_i \oplus p_i \tag{7}$$

where $\oplus$ denotes the concatenation operator.
Our main goal is to learn a function $f : X_i \rightarrow y_t \in Y_T$ such that:

$$g(X_i) = \arg\max_{y_t \in Y_T} f(X_i, y_t) \tag{8}$$

To accomplish this task, we employ a diverse set C of shallow machine learning algorithms, specifically: *Support Vector Machine, Stochastic Gradient Descent, Passive Aggressive, Perceptron, and Multi-Layer Perceptron*. Each classifier is trained independently on different feature spaces or

subspaces, resulting in a diverse ensemble with varying strengths and complementary error patterns.
The diversity in both feature spaces and classifier types is crucial for achieving robust performance across different domain adaptation scenarios. By training classifiers on different combinations of features (e.g., $X_i = h_i \oplus a_i$, $X_i = a_i$, etc.), we can capture different aspects of the domain transfer problem and leverage the complementary strengths of various machine learning approaches.

5.1.3 *Heterogeneous Bayesian Model Averaging*

To effectively leverage the complementarity of different classifiers and feature types, we employ a sophisticated ensemble approach based on Bayesian Model Averaging (BMA), initially proposed in [157]. Our approach extends this framework to handle heterogeneous feature spaces and classifier types, which we term Heterogeneous Bayesian Model Averaging (HBMA).
The ensemble methodology involves training multiple machine learning models using various combinations of the extracted features. The models are trained on different subspaces of X_i, including:

- Full feature space: $X_i = h_i \oplus a_i \oplus p_i$
- Pairwise combinations: $X_i = h_i \oplus a_i$, $X_i = h_i \oplus p_i$, $X_i = a_i \oplus p_i$
- Individual features: $X_i = h_i$, $X_i = a_i$, $X_i = p_i$

The resulting set of trained models are combined according to the BMA ensemble paradigm, which considers both the marginal probability and the reliability of each model to generate the final prediction. The final ensemble is composed of the optimal combination of base classifiers selected from all possible sets of models trained on different feature spaces.

This ensemble approach enables effective knowledge transfer by exploiting complementary information captured in different feature subspaces while relying on the specific capabilities of each model. Therefore, since we train the models on heterogeneous spaces, we will denote the proposed approach as Heterogeneous Bayesian Model Averaging (HBMA). A graphical representation of the proposed approach is reported in Figure 6.

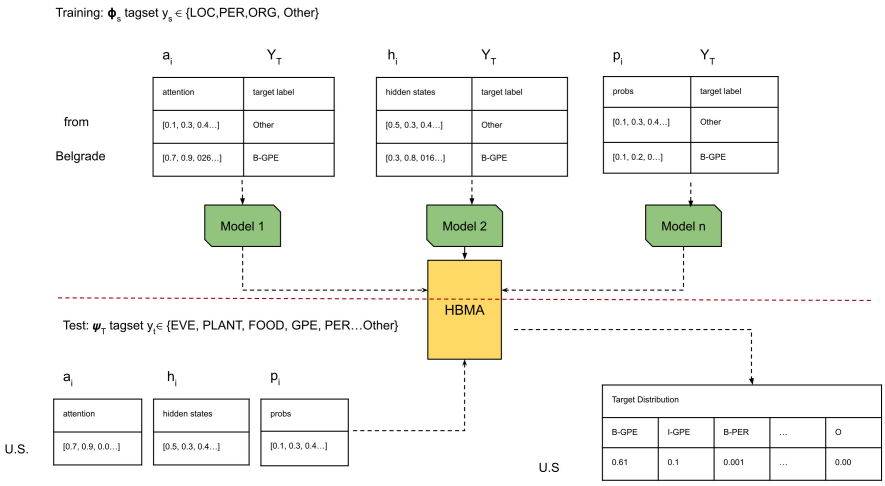


Figure 6: Workflow of the proposed Heterogeneous Bayesian Model Averaging approach for Cross-Domain NER transfer learning. The framework extracts three types of features from pre-trained transformers and combines multiple shallow classifiers, trained on these features, through Bayesian model averaging.

5.2 EXPERIMENTAL DESIGN AND DATASETS

We evaluate our approach using three well-established benchmark datasets that represent diverse characteristics in terms of size, annotation schema, and linguistic properties. This diversity makes them particularly suitable for evaluating cross-domain adaptation capabilities.

5.2.1 *Datasets*

OntoNotes 5.0 [17] is a comprehensive multilingual dataset that provides rich annotations across multiple domains including news, broadcast conversations, telephone conversations, web data, and religious texts. The dataset features a hierarchical annotation scheme with 18 entity types, making it one of the most comprehensive NER datasets available.

CoNLL-2003 [18] represents the classic benchmark in NER evaluation, focusing on news domain text with a flat annotation scheme. Despite having only 4 entity types, it remains widely used due to its well-established evaluation protocols and clean annotations.

MultiNERD [19] is a more recent multilingual dataset that spans 11 languages and covers diverse domains including news, Wikipedia, so-

cial media, and web content. It features 15 entity types with a hierarchical annotation scheme and represents modern NER challenges.
The entity tag sets for each dataset are:

- **CoNLL-2003**: *Person, Organization, Location, Miscellaneous*
- **OntoNotes 5.0**: *Person, Organization, Location, GPE, Facility, Time, Date, Money, Cardinal, Percent, Product, Event, Work Of Art, Ordinal, Language, Law, NORP, Quantity*
- **MultiNERD**: *Person, Location, Organization, Animal, Bio, Celestial Body, Disease, Event, Instrument, Media, Mythology, Plant, Time, Vehicle, Food*

Table 4: Comprehensive dataset statistics of three benchmarks used for Cross-Domain NER evaluation

Feature	OntoNotes 5.0	CoNLL-2003	MultiNERD
Domains	News, broadcast, telephone, web, religious	News	News, wiki, social media, web
Entity Types	18	4	15
Total Entities	153,092	35,089	54,020
Languages	English, Chinese, Arabic	English, German	11 languages
Publication Year	2013	2003	2021
Annotation Scheme	Hierarchical	Flat	Hierarchical
Token Count	~1.6M	~301K	~365K (English)

Table 4 reports a few statistics related to the considered datasets. The granularity of the tagsets presents both challenges and opportunities for adapting NER systems from a source to a target domain. For instance, LOC category in CoNLL-2003 encompasses both GPE (Geopolitical Entity) and LOC (Location) categories in OntoNotes, requiring the model to learn fine-grained distinctions or appropriate generalizations. MultiNERD introduces specialized categories like *Animal* and *Plant* that have no direct correspondence in other datasets, testing the model's ability to handle novel entity types.
The proposed approach has been compared against
fine-tuning transformer-based architectures suitable for NER purposes, i.e. BERT-NER [158]. More precisely, the BERT-NER models implemented with transformers library from Huggingface are pre-trained on Ontonotes5.0[1], pre-trained on MultiNERD[2] and pre-trained on ConLL03[3]

1 https://huggingface.co/djagatiya/ner-bert-base-cased-ontonotesv5-englishv4
2 https://huggingface.co/medxiaorudan/bert-base-cased-finetuned-MultiNERD-SystemA
3 https://huggingface.co/datasets/eriktks/conll2003

For each tansfer-learning scenario reported in Table 5, source BERT-NER models are fine-tuned using a batch size of 32 for 3 epochs.

5.2.2 *Transfer Learning Configurations*

We evaluate our approach across all possible bidirectional transfer scenarios between the three datasets, resulting in six distinct transfer configurations as shown in Table 5. Each configuration presents unique

Table 5: Cross-Domain transfer scenarios and their complexity characteristics

Source	**Target**	**Transfer Complexity**
OntoNotes 5.0	CoNLL-2003	High (18 → 4 types)
CoNLL-2003	OntoNotes 5.0	High (4 → 18 types)
OntoNotes 5.0	MultiNERD	Medium (18 → 15 types)
MultiNERD	OntoNotes 5.0	Medium (15 → 18 types)
CoNLL-2003	MultiNERD	High (4 → 15 types)
MultiNERD	CoNLL-2003	High (15 → 4 types)

challenges:
High Complexity Transfers: Scenarios involving CoNLL-2003 are particularly challenging due to the significant difference in tagset granularity. Transferring from CoNLL-2003 to OntoNotes or MultiNERD requires learning to distinguish between fine-grained categories, while the reverse direction requires appropriate generalization.
Medium Complexity Transfers: From OntoNotes to MultiNERD transfers involve similar numbers of entity types but different domain characteristics and some non-overlapping categories.
For each transfer scenario, we evaluate performance using three different target domain training set sizes: 5000, 2500, and 1200 instances. These settings simulate realistic scenarios with limited annotated data and allow us to assess both model capabilities and resource consumption under different data availability conditions.

5.3 EVALUATION METHODOLOGY

5.3.1 *Entity Recognition Capabilities*

Our evaluation employs the standard BIO (Beginning, Inside, Outside) tagging schema for sequence labeling in NER tasks. The primary met-

rics for assessment are entity-level precision, recall, and F1 score [159], which are computed based on exact entity match criteria.
It is crucial to note that an entity is considered correctly predicted only if both boundaries and type are exactly matched. This strict evaluation criterion reflects real-world requirements where both the span and semantic type of an entity must be correct to be useful for downstream applications. This evaluation approach ensures that our results are directly comparable with established benchmarks in the NER literature.

5.3.2 *Resource Consumption Analysis*

To comprehensively assess the efficiency advantages of our approach, we measure several key resource utilization indicators during both training and inference phases:

- **CPU Usage (%)**: Average CPU utilization during training and inference phases, sampled at regular intervals
- **GPU Memory (GB)**: The amounts of GPU memory (GB) used during training and inference
- **RAM Usage (GB)**: System memory footprint, important for understanding scalability limitations
- **Execution Time (s)**: Wall-clock time required to process the target domain test set, crucial for real-time applications

These measurements enable direct comparison between our proposed approach and fine-tuning-based methods, providing insights into the practical deployment implications of each approach. The resource analysis is particularly relevant for understanding the environmental impact and cost implications of different transfer learning strategies.

5.3.3 *Experimental Infrastructure*

To ensure reproducible and fair evaluation, we conducted all experiments on a standardized hardware configuration:

- **CPU**: AMD Ryzen 7 5700G 8-Core Processor (3.8 GHz base, 4.6 GHz boost)

- **GPU**: NVIDIA GeForce RTX 3090 with 24 GB GDDR6X memory
- **RAM**: 64 GB DDR4-3200 MHz
- **Operating System**: Ubuntu 20.04 LTS
- **CUDA**: Version 12.0 with NVIDIA driver 525.105.17

The software environment includes PyTorch 2.0.1, Hugging Face Transformers library (version 4.30.2) for transformer components, and scikit-learn 1.3.0 for the machine learning classifiers. This configuration provides sufficient computational resources to fairly evaluate both approaches while representing realistic deployment scenarios for many organizations.

5.4 RESULTS AND ANALYSIS

Table 6 presents a comprehensive comparison between our Heterogeneous Bayesian Model Averaging (HBMA) approach and fine-tuned BERT across all six transfer scenarios and three training data sizes, reporting both resource consumption and entity recognition capabilities.

The transfer from MultiNERD to CoNLL reveals a good trade-off between efficiency and prediction capabilities. We can easily note that HBMA not only outperforms the fine-tuned Bert in terms of F1 measure, but its resource consumption is approximately 8-12 times less in terms of GPU during training. The transfer from CoNLL to MultiNERD shows a substantial difference in inference times between approaches. In fact, HBMA takes ~125s while BERT ~215s, implying that the proposed approach is able to achieve comparable F1 with respect to the model baseline but at half of the inference time. When considering the transfer from Ontonotes to MultiNERD, we can similarly highlight that HBMA achieves comparable performances in terms of F1 score, but with a substantially lower resource consumption at inference time. Regarding the transfer from Ontonotes to CoNLL, HBMA is able to significantly outperform the state-of-the-art baseline also using a decreasing number of training instances. The transfer scenario from MultiNERD to Ontonotes represents the only case where BERT consistently outperforms HBMA across the different training set sizes. Finally, also in the case of transfer from CoNLL to Ontonotes we can highlight that HBMA outperform the state-of-the-art Bert-based model, still maintaining a reduced resource consumption at inference time.
As a general remark, we can affirm that HBMA generally achieves competitive or superior performance compared to BERT fine-tuning,

Table 6: Model Comparison: Resource consumption vs Prediction Capabilities

Experiment	Model	Training			Inference				F1
		CPU (%)	RAM (GB)	GPU (GB)	Execution Time (s)	CPU (%)	RAM (GB)	GPU (GB)	
MultiNERD ↓ CoNLL		**5K Training Instances**			**CoNLL Test Set**				
	HBMA	30.89	**11.64**	**1.71**	**23.25**	**4.32**	**1.57**	**1.06**	**0.864**
	Bert	**6.19**	13.21	15.34	24.28	10.5	2.94	1.40	0.843
		2.5K Training Instances			**CoNLL Test Set**				
	HBMA	28.36	**7.20**	**1.70**	23.38	**4.22**	**1.56**	**1.06**	**0.858**
	Bert	**4.48**	13.18	15.48	**19.29**	11.2	2.80	1.40	0.857
		1.2K Training Instances			**CoNLL Test Set**				
	HBMA	28.15	**4.92**	**1.66**	**23.40**	**4.28**	**1.57**	**1.06**	**0.856**
	Bert	**7.42**	13.16	20.03	23.67	9.9	2.99	1.40	0.846
CoNLL ↓ MultiNERD		**5K Training Instances**			**MultiNERD Test Set**				
	HBMA	47.63	17.86	**1.56**	**124.43**	23.41	**2.86**	**1.00**	0.910
	Bert	**6.50**	**15.68**	16.87	215.75	**17.7**	11.70	1.40	**0.912**
		2.5K Training Instances			**MultiNERD Test Set**				
	HBMA	40.02	**9.15**	**1.48**	**125.46**	23.26	**2.82**	**1.00**	0.901
	Bert	**6.90**	15.41	16.81	216.74	**18.9**	11.74	1.40	**0.903**
		1.2K Training Instances			**MultiNERD Test Set**				
	HBMA	42.45	**6.18**	**1.34**	**125.27**	23.07	**2.78**	**1.00**	**0.891**
	Bert	**11.88**	15.40	17.00	215.57	**19.0**	11.64	1.40	0.867
Ontonotes ↓ MultiNERD		**5K Training Instances**			**MultiNERD Test Set**				
	HBMA	39.82	17.21	**1.53**	**106.94**	12.91	**2.91**	**1.00**	0.907
	Bert	**6.76**	**13.25**	17.24	241.59	**9.9**	12.10	1.41	**0.917**
		2.5K Training Instances			**MultiNERD Test Set**				
	HBMA	48.41	**9.04**	**1.42**	**107.3**	**12.91**	**2.90**	**1.00**	0.895
	Bert	**10.5**	12.94	14.21	238.29	16.8	12.28	1.41	**0.899**
		1.2K Training Instances			**MultiNERD Test Set**				
	HBMA	55.49	**6.01**	**1.20**	**106.97**	**12.94**	**2.91**	**1.00**	0.886
	Bert	**4.35**	13.08	16.73	239.47	16.8	12.28	1.41	**0.889**
Ontonotes ↓ CoNLL		**5K Training Instances**			**CoNLL Test Set**				
	HBMA	28.40	**10.00**	**1.66**	23.32	**4.30**	**1.59**	**1.06**	**0.890**
	Bert	**6.67**	13.24	15.18	**19.16**	11.2	2.7	1.40	0.851
		2.5K Training Instances			**CoNLL Test Set**				
	HBMA	29.71	**6.51**	**1.83**	**23.13**	**4.34**	**1.58**	**1.06**	**0.879**
	Bert	**7.52**	13.23	15.37	27.35	9.2	2.65	1.40	0.858
		1.2K Training Instances			**CoNLL Test Set**				
	HBMA	36.03	**4.45**	**1.77**	23.00	**4.41**	**1.58**	**1.08**	**0.878**
	Bert	**4.10**	13.21	14.33	**19.79**	9.9	2.92	1.40	0.850
MultiNERD ↓ Ontonotes		**5K Training Instances**			**Ontonotes Test Set**				
	HBMA	46.60	18.27	**1.71**	**51.59**	**6.42**	**1.53**	**1.08**	0.769
	Bert	**6.22**	**14.51**	20.40	55.38	17.9	4.41	1.40	**0.842**
		2.5K Training Instances			**Ontonotes Test Set**				
	HBMA	51.51	**10.49**	**1.62**	**51.51**	**6.30**	**1.52**	**1.08**	0.759
	Bert	**6.53**	15.30	17.65	55.67	17.8	4.43	1.40	**0.827**
		1.2K Training Instances			**Ontonotes Test Set**				
	HBMA	43.15	**5.81**	**1.63**	**50.84**	**6.39**	**1.52**	**1.08**	0.734
	Bert	**6.48**	15.28	17.29	55.8	17.7	4.47	1.40	**0.807**
CoNLL ↓ Ontonotes		**5K Training Instances**			**Ontonotes Test Set**				
	HBMA	64.09	18.61	**1.71**	**55.03**	**4.41**	**1.61**	**1.09**	**0.831**
	Bert	**4.39**	**13.22**	19.86	61.64	9.8	4.80	1.41	0.827
		2.5K Training Instances			**Ontonotes Test Set**				
	HBMA	72.66	**10.85**	**1.61**	**54.05**	**4.39**	**1.61**	**1.09**	**0.799**
	Bert	**8.07**	13.23	18.01	79.52	13.6	4.87	1.41	0.789
		1.2K Training Instances			**Ontonotes Test Set**				
	HBMA	88.97	**6.55**	**1.54**	**55.33**	**4.19**	**1.41**	**1.09**	**0.775**
	Bert	**4.03**	12.55	14.12	87.30	10.6	4.88	1.41	0.723

particularly when a limited amount of data is available in the target domain. The proposed approach uses higher CPU resources (28-89%) compared to BERT fine-tuning (4-13%), but requires consistently less RAM during both training and inference. For what concerns GPU Utilization, in the training phase, HBMA demonstrates substantially lower GPU requirements compared to BERT fine-tuning and maintains lower GPU usage during the inference phase compared to the fine-tuning approach. Regarding Inference Time, HBMA shows significantly faster inference for some dataset combinations. Additionally, HBMA exploits a consistently smaller quantity of RAM compared to BERT fine-tuning. BERT fine-tuning can achieve marginally superior performance in some dataset combinations (e.g. from MultiNERD to Ontonotes) but at significantly higher resource consumption costs. This suggests that HBMA may be particularly suitable for low-resource languages, specialized domains, or applications where labeled data acquisition is too expensive or challenging. The substantial differences in resource utilization translate into significant implications for deployment costs, energy consumption, and environmental impact. HBMA's lower GPU and RAM requirements during both training and inference represent substantial advantages for resource-constrained environments or large-scale deployments where efficiency is of paramount importance.

RQ 5.1 - How can features obtained from a source domain be leveraged to learn target domain labels and distributions efficiently, achieving performance close to fine-tuned transformers but with lower computational cost?

Adapting the knowledge of pre-trained language models not only allows to achieve comparable performance with respect to fine-tuned transformer architectures but at lower computational costs.

In particular, if we assume a system that should be maintained operational continuously, we can easily estimate the reduced energy consumption of the proposed approach with respect to the state of the art. For instance, considering the hardware infrastructure used in these experiments, when processing MultiNERD (32k sentences), considering GPU, the energy consumption per second is approximately 0.097 kWh, while CPU contributes to an additional 0.018 kWh per second. For MultiNERD, the reduced inference time (from 215s to 125s) saves 90s x 0.097 kWh for what concerns the GPU and 90s x 0.018kWh for

what regards CPU, which roughly corresponds to ~ 10.35 kWh on a full inference cycle. Over extended periods, scaling to millions of sentences per day, this reduction compounds into significant cost and carbon footprint reductions. In a real-world deployment scenario, where inference is perfomed continuously (e.g. on news articles or social media content), HBMA not only reduces the computational overhead but also significantly lowers energy consumption, potentially saving 40-50% of the power with respect to the state-of-the-art BERT-NER.

5.5 ABLATION STUDIES AND FEATURE ANALYSIS

To better understand the contribution of different feature types to the overall performance of HBMA, we conducted extensive ablation studies examining the impact of individual feature components and their combinations across all six transfer scenarios. The detailed results presented in Tables 6-23 provide comprehensive insights into feature effectiveness and classifier performance patterns. For every transfer scenario showed in Table 5 and for every dimensionality of training data (5000, 2500, 1200 instances), Tables show individual shallow classifiers performances in terms of F1 score. BMA ensemble is composed of all classifiers trained on the input subspace.

Table 7: Predictive performance from ConLL03 to Ontonotes5.0 with 5000 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.7311	0.7622	0.741	0.646	0.7634	0.7690
hidden	0.753	0.6701	0.7728	0.7459	0.8015	0.807
probs	0.447	0.5189	0.5276	0.411	0.5461	0.5506
att+hidden	0.7953	0.6658	0.7950	0.7713	0.8118	**0.8261**
att+probs	0.7122	0.7622	0.7351	0.5300	0.7531	0.7675
probs+hidden	0.64	0.5893	0.7574	0.789	0.8164	0.7761
att+probs+hidden	0.8073	0.6715	0.7957	0.7574	0.8035	**0.8263**

5.5.1 *Individual Feature Performance Analysis*

Our ablation studies reveal distinct patterns in the contribution of different feature representations: Across most transfer scenarios, atten-

Table 8: Predictive performance from ConLL03 to Ontonotes5.0 with 2500 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.6904	0.6775	0.6886	0.6751	0.7494	0.7417
hidden	0.7295	0.7117	0.7342	0.7196	0.7859	0.7286
probs	0.4776	0.5180	0.5329	0.5299	0.5464	0.5319
att+hidden	0.7660	0.6974	0.7763	0.7419	**0.7972**	0.7540
att+probs	0.6745	0.7344	0.7258	0.6643	0.7432	0.7413
probs+hidden	0.7633	0.7171	0.7725	0.7521	0.7988	0.7194
att+probs+hidden	0.7761	0.6991	0.7834	0.7675	**0.7959**	0.7457

Table 9: Predictive performance from ConLL03 to Ontonotes5.0 with 1200 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.6984	0.7174	0.6984	0.5738	0.7318	0.7385
hidden	0.6775	0.7060	0.175	0.5190	0.7218	0.7218
probs	0.4971	0.5182	0.5292	0.1026	0.5457	0.5272
att+hidden	0.7094	0.696	0.6786	0.6061	0.7107	**0.743**
att+probs	0.639	0.7179	0.711	0.642	0.740	0.7087
probs+hidden	0.6775	0.7107	0.1777	0.6153	0.7091	0.7156
att+probs+hidden	0.6726	0.6947	0.7035	0.6624	0.6764	**0.7411**

tion scores emerge as the most contributive feature representation for learning the objective function. This finding validates our hypothesis that attention patterns capture crucial information about entity boundaries and contextual focus that generalizes well across domains. For instance, in the MultiNERD $\rightarrow$ CoNLL transfer with 5000 training instances, attention-only models achieve F1 scores of 0.8520 with BMA, demonstrating robust performance.

Contextualized hidden states provide substantial but generally secondary contributions to the overall performance. The notable exception is the CoNLL $\rightarrow$ OntoNotes transfer scenario, where hidden states show superior individual performance (F1 = 0.807 vs 0.769 for attention with 5000 instances). This suggests that the semantic richness captured in hidden states is particularly valuable when expanding

Table 10: Predictive performance from Ontonotes5.0 to ConLL03 with 5000 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8472	0.8652	0.8753	0.8693	0.8680	**0.8792**
hidden	0.7062	0.6053	0.2433	0.7790	0.8132	0.6954
probs	0.7699	0.7793	0.775	0.7002	0.7823	0.7776
att+hidden	0.8171	0.6195	0.7914	0.8031	0.7736	0.7746
att+probs	0.8529	0.8651	0.8536	0.8360	**0.8755**	0.8714
probs+hidden	0.7494	0.5607	0.2083	0.7314	0.7963	0.7054
att+probs+hidden	0.7990	0.6187	0.7632	0.7964	0.8229	0.7735

Table 11: Predictive performance from Ontonotes5.0 to ConLL03 with 2500 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8316	0.8525	0.8411	0.8337	0.8558	**0.8659**
hidden	0.7612	0.6571	0.2401	0.7276	0.8152	0.7244
probs	0.7496	0.7785	0.7742	0.5718	0.7819	0.7749
att+hidden	0.8195	0.6613	0.8088	0.8199	0.8351	0.7667
att+probs	0.8496	0.8532	0.8521	0.8279	0.8638	**0.8650**
probs+hidden	0.7977	0.5819	0.2249	0.6921	0.8194	0.7127
att+probs+hidden	0.8102	0.6625	0.8085	0.7662	0.8230	0.7721

from coarse-grained to fine-grained entity taxonomies. Source probability distributions consistently show the lowest individual contribution across all scenarios. This limitation stems from the inherent complexity of distinguishing between tokens that may have similar probability distributions in the source domain but require different labels in the target domain. For example, in CoNLL $\rightarrow$ OntoNotes scenarios, probability-only models achieve F1 scores around 0.55, significantly lower than other feature types.

Table 12: Predictive performance from Ontonotes5.0 to ConLL03 with 1200 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8559	0.8558	0.8377	0.8531	0.8606	**0.8649**
hidden	0.7498	0.6725	0.2568	0.7259	0.7726	0.7286
probs	0.7663	0.7778	0.7773	0.6148	0.7838	0.7767
att+hidden	0.8238	0.8068	0.8115	0.8473	0.8563	0.8195
att+probs	0.8448	0.8555	0.8342	0.8389	0.8566	**0.8629**
probs+hidden	0.7819	0.7075	0.2680	0.7651	0.7824	0.7562
att+probs+hidden	0.8274	0.8068	0.8318	0.8018	0.8552	0.8163

Table 13: Predictive performance from Multinerd to ConLL03 with 1200 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8255	0.8343	0.8119	0.8143	**0.8386**	0.8434
hidden	0.4840	0.4148	0.3914	0.3636	0.6285	0.5132
probs	0.7561	0.7432	0.7596	0.7424	0.7605	0.7678
att+hidden	0.6568	0.4706	0.5063	0.6729	0.7065	0.6919
att+probs	0.8300	0.8345	0.7807	0.8106	0.8343	**0.8371**
probs+hidden	0.4465	0.4089	0.4643	0.4905	0.5820	0.5549
att+probs+hidden	0.7420	0.4737	0.6828	0.7310	0.7856	0.7194

5.5.2 *Feature Combination*

The ablation results reveal important insights about feature combination strategies. The combination of attention scores and source probability distributions consistently achieves more robust performance compared to other feature combinations. In the OntoNotes $\rightarrow$ CoNLL scenario with 5000 instances, this combination reaches F1 = 0.8714 with BMA, outperforming most other configurations while maintaining computational efficiency. Contrary to initial expectations, concatenating all three feature types does not always yield the best performance. The full feature combination (att+probs+hidden) often shows performance drops compared to the attention+probabilities combina-

Table 14: Predictive performance from Multinerd to ConLL03 with 2500 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8066	0.8259	0.8306	0.7830	**0.8428**	0.8406
hidden	0.4621	0.2020	0.4932	0.5553	0.4287	0.5681
probs	0.7305	0.7509	0.7729	0.6086	0.7581	0.7704
att+hidden	0.5716	0.3443	0.5965	0.6624	0.5727	0.6437
att+probs	0.8132	0.8255	0.8192	0.8093	**0.8469**	0.8434
probs+hidden	0.4845	0.1999	0.4305	0.5307	0.4641	0.5538
att+probs+hidden	0.6804	0.3454	0.6558	0.6956	0.7269	0.5586

Table 15: Predictive performance from Multinerd to ConLL03 with 5000 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8355	0.8429	0.8353	0.8032	0.8384	**0.8520**
hidden	0.5814	0.4059	0.5391	0.5512	0.4663	0.6355
probs	0.7278	0.7482	0.7702	0.6663	0.7739	0.7714
att+hidden	0.6614	0.211	0.6485	0.6882	0.6831	0.6923
att+probs	0.8228	0.8423	0.8409	0.8151	0.8414	**0.8560**
probs+hidden	0.6206	0.3499	0.6002	0.5471	0.5200	0.6427
att+probs+hidden	0.6894	0.1976	0.6294	0.6687	0.5625	0.6699

tion, while significantly increasing computational overhead due to the higher dimensionality of the resulting vectors.

5.5.3 *Machine Learning Classifier Analysis*

MLPs consistently achieve the best individual performance across most feature spaces and transfer scenarios. In attention-based models, MLPs often reach F1 scores 2-5 points higher than other classifiers, demonstrating their ability to capture complex non-linear relationships in the feature space. SVMs show consistent performance across different scenarios, making them reliable ensemble components. They particularly excel with attention-based features, often achieving competitive perfor-

Table 16: Predictive performance from ConLL03 to Multinerd with 5000 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8941	0.9025	0.9040	0.8693	0.9020	**0.9032**
hidden	0.7668	0.8605	0.1943	0.8378	0.8504	0.7789
probs	0.8637	0.8739	0.8499	0.8432	0.8638	0.8641
att+hidden	0.8353	0.8739	0.8712	0.8169	0.8742	0.8806
att+probs	0.8931	0.9024	0.8959	0.8777	0.9063	**0.9088**
probs+hidden	0.8314	0.8812	0.1806	0.8291	0.8618	0.7817
att+probs+hidden	0.8423	0.8739	0.8244	0.8277	0.8695	0.8850

Table 17: Predictive performance from ConLL03 to Multinerd with 2500 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8855	0.8940	0.8929	0.8846	0.8954	**0.8987**
hidden	0.8084	0.8708	0.1805	0.8077	0.8273	0.8738
probs	0.8353	0.859	0.8550	0.8408	0.8609	0.8646
att+hidden	0.8754	0.8671	0.8707	0.8479	0.8453	0.8783
att+probs	0.8908	0.8944	0.8870	0.8724	0.8992	0.9005
probs+hidden	0.8362	0.8723	0.1711	0.7689	0.8510	0.8731
att+probs+hidden	0.8896	0.8878	0.8286	0.8392	0.8795	0.8725

mance with MLPs while maintaining computational efficiency. SGD classifiers show high variability in performance, sometimes achieving excellent results (particularly with hidden state features) but also showing dramatic performance drops in certain configurations, making them less reliable for consistent deployment.

5.5.4 *Ensemble Integration*

While Bayesian Model Averaging, composed of models trained on same features, generally improves performance by combining multiple classifiers, it does not universally outperform the best individual components. In some cases, high-confidence prediction errors from weaker

Table 18: Predictive performance from ConLL03 to Multinerd with 1200 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8765	0.8899	0.8636	0.8701	0.8882	0.8915
hidden	0.8035	0.8708	0.1809	0.7995	0.7887	0.8660
probs	0.8536	0.8547	0.8537	0.8156	0.8548	0.8578
att+hidden	0.7587	0.8390	0.8275	0.8176	0.8461	0.8762
att+probs	0.8635	0.8899	0.8864	0.8870	**0.8929**	**0.8932**
probs+hidden	0.7587	0.8716	0.8275	0.8176	0.8461	0.8642
att+probs+hidden	0.8775	0.8420	0.8354	0.8081	0.8607	0.8759

Table 19: Predictive performance from Ontonotes5.0 to Multinerd with 5000 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8934	0.8976	0.8956	0.8602	**0.9038**	0.9009
hidden	0.6483	0.5329	0.5001	0.6235	0.7382	0.7547
probs	0.7484	0.7827	0.7698	0.7648	0.7839	0.7849
att+hidden	0.7953	0.4987	0.8008	0.8063	0.7463	0.8197
att+probs	0.8847	0.8974	0.8999	0.8616	0.8994	0.8197
probs+hidden	0.6250	0.5401	0.4535	0.6286	0.7503	0.7684
att+probs+hidden	0.7797	0.4693	0.7942	0.7818	0.7567	0.8052

classifiers can bias the ensemble toward incorrect decisions. The diversity among classifier types provides valuable complementarity. MLPs excel at capturing complex patterns, SVMs provide stable decision boundaries, and other classifiers contribute specialized strengths that enhance overall robustness.

RQ 5.2 - Which representations matter most to deal with the problem of transfer learning for NER?

Attention scores emerge as the most contributive feature representation for learning the objective function, and its combination with source probability distributions leads to more robust scenario.

Table 20: Predictive performance from Ontonotes5.0 to Multinerd with 2500 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8709	0.8880	0.8736	0.8589	**0.8914**	**0.8861**
hidden	0.6789	0.6009	0.5843	0.6479	0.7022	0.6246
probs	0.7556	0.7812	0.7716	0.7142	0.7829	0.7819
att+hidden	0.8329	0.4001	0.8091	0.7626	0.8107	0.4584
att+probs	0.8789	0.8879	0.8773	0.8830	0.8853	0.8859
probs+hidden	0.6636	0.5920	0.5752	0.6366	0.6556	0.6151
att+probs+hidden	0.8186	0.4194	0.8068	0.7896	0.8250	0.4787

Table 21: Predictive performance from Ontonotes5.0 to Multinerd with 1200 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.8643	0.8794	0.8644	0.8488	**0.8825**	0.8777
hidden	0.6338	0.3941	0.4387	0.5902	0.7035	0.4701
probs	0.7280	0.7831	0.7696	0.5590	0.7838	0.7831
att+hidden	0.8254	0.7465	0.7839	0.8027	0.8024	0.7593
att+probs	0.8406	0.8791	0.8646	0.8713	**0.8814**	0.8778
probs+hidden	0.6061	0.4093	0.4288	0.5548	0.6662	0.4815
att+probs+hidden	0.8124	0.7523	0.7622	0.8048	0.8340	0.7652

5.6 DISCUSSION AND CONCLUSIONS

While HBMA demonstrates significant advantages, several limitations should be acknowledged:
In some transfer scenarios (particularly MultiNERD - OntoNotes), BERT fine-tuning achieves superior F1 scores, indicating room for improvement in the feature extraction or ensemble methodology. The approach requires careful extraction and combination of multiple feature types, adding complexity compared to end-to-end fine-tuning approaches. Several promising directions for future work emerge from this research:

- Investigating the optimal transformer layers for feature extraction across different transfer scenarios could improve performance.

Table 22: Predictive performance from Multinerd to Ontonotes5.0 with 5000 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.6625	0.7432	0.7271	0.6469	0.7591	**0.7619**
hidden	0.4319	0.3050	0.4749	0.3996	0.4663	0.5299
probs	0.4302	0.4981	0.5008	0.4074	0.5314	0.5079
att+hidden	0.4775	0.1834	0.4675	0.5140	0.6178	0.5982
att+probs	0.6538	0.7435	0.7228	0.6713	**0.7538**	0.7465
probs+hidden	0.3271	0.3103	0.4590	0.4132	0.5550	0.5427
att+probs+hidden	0.5320	0.1619	0.4869	0.5846	0.6446	0.6372

Table 23: Predictive performance from Multinerd to Ontonotes5.0 with 2500 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.6974	0.7168	0.7122	0.6816	0.7324	**0.7452**
hidden	0.4602	0.2595	0.4656	0.4547	0.4781	0.5234
probs	0.4539	0.4978	0.5061	0.4650	0.5269	0.5097
att+hidden	0.5281	0.3199	0.5131	0.5993	0.5948	0.6253
att+probs	0.6840	0.7175	0.7042	0.6561	0.7186	**0.7433**
probs+hidden	0.4062	0.2568	0.4549	0.4309	0.4270	0.5859
att+probs+hidden	0.5511	0.3315	0.4886	0.6151	0.6423	0.6345

- Developing methods to automatically select the most relevant feature combinations for specific domain pairs.
- Analyzing and leveraging specialized attention heads that focus on entity-relevant patterns.

This chapter presented a transfer learning approach for Named Entity Recognition that enables the accurate learning of target domain labels from source domain features. The proposed method demonstrates competitive performance compared to fine-tuned transformer-based models while requiring significantly lower computational resources. HBMA guarantees 55-60% faster inference times compared to Bert fine-tuning, making it particularly suitable for real-time applications. Additionally, HBMA reduces GPU memory requirements

Table 24: Predictive performance from Multinerd to Ontonotes5.0 with 1200 training instances. Scores represent f1-score of machine learning models trained on specific vector representation.

Model	pa	svm	sgd	per	mlp	BMA
attention	0.6069	0.6907	0.6779	0.6314	**0.7313**	0.7228
hidden	0.4192	0.1896	0.4579	0.4086	0.4833	0.5220
probs	0.1036	0.4964	0.4985	0.1060	0.5273	0.5053
att+hidden	0.5574	0.3905	0.5301	0.5701	0.6030	0.6281
att+probs	0.6817	0.6927	0.7007	0.6432	**0.7271**	0.7250
probs+hidden	0.3854	0.1851	0.5106	0.3975	0.4323	0.535
att+probs+hidden	0.4875	0.4045	0.5315	0.5252	0.6218	0.6249

by approximately 91% during both training and inference compared to Bert fine-tuning, enabling deployment in resource-constrained environments. These findings show that reusing the knowledge of pre-trained language models not only allows to achieve comparable performance with respect to fine-tuned transformer architectures but at lower computational costs.

A more fine-grained analysis reveals that the accuracy–efficiency trade-off becomes domain-dependent. In high-stakes settings such as clinical Named Entity Recognition, where errors in identifying medical conditions, medications, or procedures may propagate to downstream clinical decision-support systems, maximizing accuracy is often the primary concern. In such scenarios, full transformer fine-tuning—despite its higher computational cost—can be justified by its superior ability to adapt representations to domain-specific terminology and contextual nuances. A similar consideration applies to legal document analysis, where precise entity recognition is essential for tasks such as contract interpretation or compliance monitoring, and even minor extraction errors can have significant practical implications.

Conversely, in domains characterized by high data volume and lower per-instance risk, such as information extraction from social media streams, news monitoring, or large-scale document processing pipelines, computational efficiency becomes a dominant factor. In these settings, the ability to process large amounts of text with low latency and limited hardware resources often outweighs marginal gains in accuracy. Here, HBMA provides a more favorable trade-off by enabling fast inference and reduced memory consumption while maintaining competitive performance. These observations suggest that while accuracy-

oriented approaches are preferable in safety-critical or precision-sensitive domains, resource-efficient methods such as HBMA are particularly well suited for scalable, real-time, or resource-constrained information extraction scenarios. Chapter 5 concluded by presenting a cross-domain transfer learning technique for **Named Entity Recognition (NER)**. While NER focuses on identifying and classifying individual entities within text, these entities often co-occur and are linked through meaningful semantic relationships. The **Relation Extraction (RE)** task seeks to uncover such relationships, playing a key role in transforming unstructured text into structured, queryable knowledge representations that support advanced natural language understanding applications. Next Chapter extends our investigation of resource-efficient NLP methods to the domain of Relation Extraction across various English and Italian benchmarks.

Part III

RELATION EXTRACTION

6

KNOWLEDGE-AUGMENTED RELATION EXTRACTION WITH LARGE LANGUAGE MODELS

Relation extraction (RE) is a fundamental task in natural language processing that aims to identify and classify relationships between subject and object entities mentioned in text [3]. Formally, given an input sentence $s = \{w_1, w_2, \ldots, h, \ldots, t, \ldots, w_n\}$ containing n tokens, where h and t represent head and tail entities respectively, RE systems predict a relation label $r_i \in R$ from a predefined set of relationships (e.g., `founded_by`, `born_in`, and `Work_For`). This capability underlies many critical NLP applications, including knowledge graph completion and question answering systems [160].

Recent research in Relation Extraction (RE) has increasingly explored the use of large open-domain knowledge bases such as Wikipedia, Wikidata, and DBpedia to enrich textual representations and mitigate data sparsity. These resources provide structured relational information that can be aligned with text to support more accurate modeling of entity and relation semantics [161, 162]. Early approaches relied on distant supervision, automatically generating labeled data by aligning entity pairs in sentences with triples extracted from Wikipedia or Freebase [163]. Subsequent work extended this paradigm to exploit the rich graph structure of resources like Wikidata and DBpedia, integrating their embeddings or graph relations directly into neural RE architectures [164–166]. More recent frameworks combine contextual language models with structured information from these knowledge bases, enabling joint reasoning over textual and factual signals for more robust relation prediction [167, 168]. However, these external knowledge sources frequently introduce drawbacks: noise, incomplete or missing information, or irrelevant facts given the context, which can degrade performance in specialized or domain-specific settings. Such challenges have spurred increasing interest in treating Large Language Models (LLMs) as more flexible, contextual, and reliable sources of knowledge for Relation Extraction. This chapter proposes a pipeline based on exploiting the reasoning capabilities of Large Language Models (LLMs). The central hypothesis underlying our approach is that extending each sample of a given dataset using knowledge extracted by querying an LLM with specific clarification prompts helps the models

trained on these samples, along with clarifications, to understand the task better. Several models are trained on two famous English Relation Extraction benchmark CoNLL04 [169] and SemEval 2010 task 8 dataset [170]. Furthermore, we investigate the cross-lingual effectiveness of the approach on an Italian dataset, CoNLL04 Italian, which we translated from the original CoNLL04 dataset. The experimental results demonstrate that incorporating LLM-generated background knowledge improves RE performance, particularly in low-resource settings. Our analysis reveals interesting patterns in how different types of knowledge contribute to model performance, providing insights to inform broader strategies for leveraging LLMs in structured prediction tasks.
Finally, we explore a Knowledge Distillation technique, based on Kullback–Leibler divergence, that transfer the knowledge from models trained on enriched data to lightweight models operating solely on raw text. The experiments cover Italian and English datasets used in the previous step, and results show how the generalization capabilities of a model trained on raw sentences can be improved by distilling the knowledge from those models trained on samples along with extracted context from an LLM.

RESEARCH QUESTIONS

RQ 6.1 - Which complementary knowledge component can be obtained by Large Language Models to augment training data for Relation Extraction?

RQ 6.2 - What is the effect of knowledge distillation when considering only model that takes as input only the raw sentence?

6.1 PROBLEM DEFINITION

Our approach is built on the premise that Large Language Models (LLMs) offer significant advantages over traditional external knowledge bases like Wikidata for Relation Extraction tasks. This advantage stems from their superior ability to interpret sentence semantics and contextual nuances. Unlike structured knowledge bases, which provide static, predefined relations between entities in a rigid format, LLMs possess deep contextual understanding that enables them to:

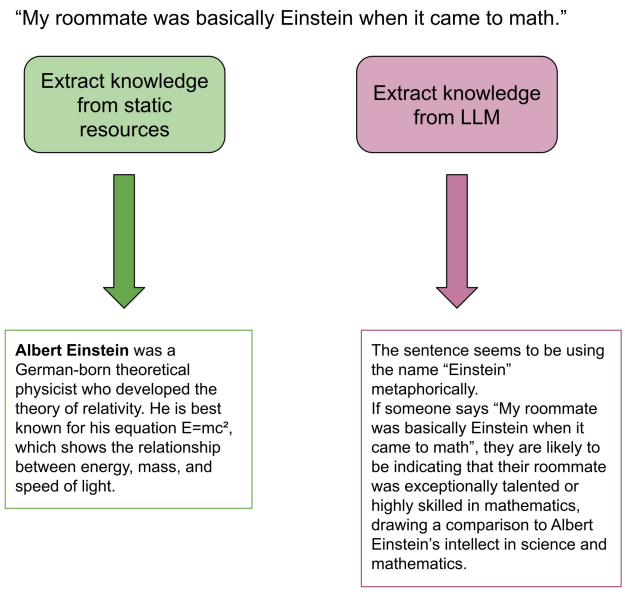


Figure 7: Differences between extracting sentence related knowledge from a static resource like Wikipedia or Wikidata (on the left) and from an LLM to get additional information (on the right) given the example sentence.

- **Resolve ambiguities**: They can disambiguate entities and relations based on contextual clues
- **Interpret complex linguistic phenomena**: They handle metaphors, negations, conditional statements, and other linguistic constructs that traditional knowledge bases cannot process

LLMs excel at understanding how the same entity pair can express different relations depending on syntactic structure, discourse context, and pragmatic implications. For instance, they can distinguish between "CEO of Apple" and "former CEO of Apple" or interpret temporal and causal relationships that emerge from sentence composition rather than explicit statement.
Furthermore, LLMs can handle novel entity combinations and emerging relationships that may not yet exist in manually curated databases. Their training on vast text corpora allows them to recognize subtle linguistic cues and contextual modifiers that determine relation validity and type. This semantic depth proves particularly valuable for Relation Extraction in domains with complex, evolving terminology or when dealing with informal text where relationships are expressed

through natural language patterns rather than formal declarations. Let $\mathcal{R}$ be the set of all possible relation types, given a sentence $s = \{w_1, w_2, \ldots, w_n\}$ consisting of n tokens, and a set of entities $E = \{e_1, e_2, \ldots, e_k\}$ where each entity e_i is defined by its span $(start_i, end_i)$ and type $y_i \in \mathcal{Y}_\mathcal{T}$, the Relation Extraction task aims to identify and classify semantic relationships $r_i \in \mathcal{R}$ between $h = e_i$ (head entity) and $t = e_j$ (tail entity) within the context of sentence s. In this work, the problem is formulated as a conditional text generation task, where we learn the probability distribution:

$$p(r|s) = \prod_{i=1}^{len(r)} p(r_i|r_{<i}, s) \tag{9}$$

where s represents the input space and r is the target representation for relation labels.

6.2 DATASETS

This section describes the datasets employed in our approach, encompassing both general domain and news domain.

6.2.1 *CoNLLo4 Dataset*

The CoNLLo4 dataset [125] serves as a benchmark dataset for Relation Extraction tasks. It contains 1,441 sentences, each containing at least one relation, Table 25 show dataset's statistics. The sentences are annotated with comprehensive information about entities and their corresponding relation types [171]. The dataset comprises news articles from The Wall Street Journal and the Associated Press, encompassing annotations for both entity and relation types, making it versatile for various NLP tasks. The dataset includes relations among entities like people, organizations, locations, and other miscellaneous entities. The relation types are five: Live_In, Located_In, OrgBased_In, Kill, Work_for, covering the most fundamental types of relationships that occur in news text. Relations included span various semantic categories: Person-Location, Organization-Person, Person-Person, among others.

The distribution of relation types shows a relatively balanced dataset, though with some variation in frequency. The Live_In relation is the most frequent, followed by OrgBased_In and Work_for, while Kill represents the least frequent relation type as shown in Table 26.

Table 25: CoNLL04 benchmark statistics. Every sample is a sentence.

	sentences	entities	relations
train	922	3377	1283
validation	231	893	343
test	288	1079	422
total	1441	5349	2048

Table 26: CoNLL04 benchmark relation types statistics

relation type	train	validation	test
Live_In	330	91	100
Located_In	247	65	94
OrgBased_In	271	76	105
Kill	179	42	47
Work_for	256	65	76

6.2.2 *SemEval 2010 Task 8 Dataset*

SemEval is a series of international natural language processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis and to help create high-quality annotated datasets in a range of increasingly challenging problems in natural language semantics. The SemEval 2010 Task 8 dataset [170] contain around 1,200 sentences manually collected for each relation through pattern-based Web search. This dataset contains 10,717 sentences with annotated semantic relations, making it a comprehensive benchmark for Relation Extraction systems. The dataset includes 9 distinct relation types plus an "Other" category for relation pairs that do not fit into the predefined categories.

The relation types include:

- Cause-Effect,
- Component-Whole,
- Content-Container,
- Entity-Destination,
- Entity-Origin,

- Instrument-Agency,
- Member-Collection,
- Message-Topic,
- Product-Producer.

Each relation type is directional, meaning that the dataset distinguishes between the direction of the relationship (e.g., Cause-Effect(e1,e2) vs. Cause-Effect(e2,e1)), resulting in 18 directed relation types plus the "Other" category.

Table 27: SemEval 2010 Task 8 dataset statistics

sentences	relations	
train	8000	8000
test	2717	2717
total	10717	10717

The dataset includes complex sentences where entities may be embedded within larger noun phrases, requiring sophisticated entity recognition and Relation Extraction capabilities. Based on the SemEval 2010

Table 28: SemEval 2010 Task 8 relation types statistics

Labels	Training Set	Test Set
Cause Effect	1003	328
Component Whole	941	312
Content Container	540	192
Entity Destination	845	292
Entity Origin	716	258
Instrument Agency	504	156
Member Collection	690	233
Message Topic	634	261
Product Producer	717	231
Other	1410	454
Total	8000	2717

Task 8 dataset statistics shown in Table 28, several key observations can be made regarding the label distribution:

Class Imbalance Patterns: The dataset exhibits notable class imbalance, with the "Other" category being the most prevalent. This large "Other" category represents relations that don't fall into the nine defined semantic relation types.
Semantic Relation Distribution: Among the nine target relation types, "Cause-Effect" is the most frequent, followed by "Component-Whole" and "Entity-Destination". The least represented relations are "Instrument-Agency" and "Message-Topic", indicating potential challenges for models in learning these underrepresented categories.
The imbalanced distribution, particularly the large "Other" category and varying frequencies of semantic relations, presents both opportunities and challenges. This distribution pattern is typical of real-world Relation Extraction tasks, where certain semantic relationships naturally occur more frequently than others in text corpora.

6.3 BACKGROUND KNOWLEDGE CONSTRUCTION APPROACH

This section presents a comprehensive and reliable knowledge construction approach that systematically enriches training data with multiple forms of auxiliary information. In particular, this approach leverages GPT-4.1 to generate high-quality explanatory content that provides models with explicit reasoning traces and contextual understanding necessary for sophisticated Relation Extraction. Our knowledge construction process employs a targeted approach that generates two primary types of auxiliary information: **Entity Outlook** and **Sentence Outlook**. Each component represent a different linguistic perspective that contextualises the single sample using the knowledge of LLM.
Given a sentence $s_i \in D$, where $D = \{s_1, s_2, \ldots, s_n\}$ represents a corpus of input sentences, the enriched input space is represented by:

$$X_{\text{enriched}} = \langle s_i, K_i \rangle \tag{10}$$

where s_i represents the original sentence and K_i contains the extracted knowledge components for sentence s_i.

The knowledge K is composed of two complementary perspectives:

$$K_i = q_i \oplus c_i \tag{11}$$

where q_i represents the Entity Outlook extracted from sentence s_i, c_i represents the Sentence Outlook for s_i, and $\oplus$ denotes concatenation of the knowledge components.

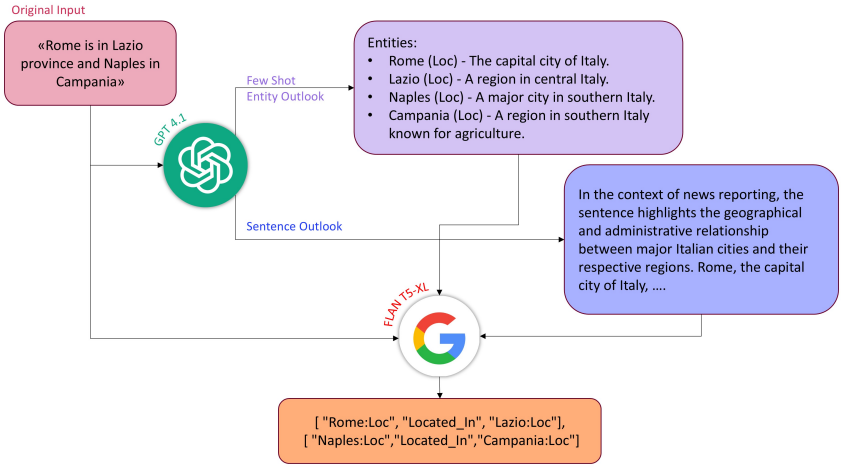


Figure 8: Overview of the proposed approach. Starting from the input sentence, the method augments the input with knowledge extracted from GPT4.1. Subsequently, a supervised fine-tuning with the LoRA strategy is performed, where LLMs learn to generate the target with a specific notation.

6.3.1 *Entity Outlook*

Entity Outlook generation focuses on providing detailed descriptions of entities mentioned in the text, incorporating their significance, characteristics, and potential roles in relationships. This component employs a few-shot learning approach, leveraging carefully curated examples to maintain consistency across different sentence contexts while adapting to domain-specific terminology.

The entity extraction and explanation process utilizes structured prompts that guide GPT-4.1 to identify and describe entities according to established taxonomies. For the CoNLL04 dataset, the process recognizes four primary entity types:

- Peop (people),
- Loc (locations),
- Org (organizations),
- and Other (dates, times, measurements).

For SemEval 2010 Task 8, the process focuses on nominal entities that participate in semantic relations. The few-shot prompting strategy in-

corporates domain-specific examples that demonstrate the desired explanation format:

Example of Entity Outlook query prompt

Given a sentence, identify all entities giving a brief description.
Entity Types:

- Peop: Individual people, names, pronouns referring to people
- Loc: Cities, countries, geographical places, addresses, buildings
- Org: Companies, institutions, government bodies, groups, political parties
- Other: Dates, times, numbers, measurements, and other miscellaneous entities

Here are examples of how to identify entities in sentences:
Example 1:
Sentence: "If it does not snow, and a lot, within this month we will have no water to submerge 150, 000 hectares (370, 500 acres) of rice..."

- Entities: Bruno Pusterla (Peop) A senior official at the Italian Agricultural Confederation.
- Italian Agricultural Confederation (Org) An Italian agricultural trade association representing farmers.

Example 2:
...
Now, please identify the entities in the following sentence, following the same pattern:
Sentence: "Elsewhere, a downed power line was blamed for a brush fire..."
Entities:

This approach ensures that the generated Entity Outlook maintain consistency in format and depth while adapting to the specific characteristics of each sentence and domain.

6.3.2 *Sentence Outlook*

Sentence Outlook is a contextualization that provides comprehensive background information necessary for understanding the sentence within its broader semantic and temporal context. This component takes in input the original sentence along with Entity Outlook and employs a zero-shot approach utilizing GPT-4.1's inherent knowledge to generate contextual explanations without requiring specific examples.
The contextualization process addresses implicit information, historical context, and situational factors that influence entity relationships but may not be explicitly stated in the text. The prompting strategy instructs GPT-4.1 to focus on providing concise but informative explanations that enhance understanding without introducing irrelevant information.

Example of Sentence Outlook query prompt

Task: Sentence contextualisation.
Instructions:

- Contextualise the sentence
- Focus on identified entities
- Keep the contextualisation concise but informative

Entities:

- Orange County (Loc) A county location where the fire occurred
- Fire Department (Org) The county's fire department organization
- ...

Sentence: "'Elsewhere , a downed power line was blamed for a brush fire that charred 200 acres in Orange County', said county Fire Department Capt. Hugh Madlock ."

This targeted approach ensures that the contextual information directly supports Relation Extraction tasks while maintaining relevance to the specific domain and sentence content.

6.3.3 *Example: Knowledge Construction Output*

To illustrate the effectiveness of our approach, consider the following example from the CoNLL04 training data:

Original Sentence: 'Elsewhere , a downed power line was blamed for a brush fire that charred 200 acres in Orange County , said county Fire Department Capt. Hugh Madlock ."

Example of Entity Outlook answer for ConLL04

Entities:

- Orange County (Loc) A county location where the fire occurred
- Fire Department (Org) The county's fire department organization
- Capt. Hugh Madlock (Peop) A fire department captain who provided the statement
- 200 acres (Other) The area of land affected by the brush fire

Example of Sentence Outlook answer

Contextualization:
In this sentence, "Orange County" refers to a populous region in Southern California known for its suburban communities and susceptibility to wildfires, especially during dry seasons. The "Fire Department" is the local government agency responsible for fire prevention and emergency response in the county. "Capt. Hugh Madlock" is identified as a captain within this department, likely serving as a spokesperson or incident commander. The mention of "200 acres" highlights the significant scale of the brush fire ...

This example demonstrates how the knowledge construction approach provides comprehensive auxiliary information that enables models to understand not only the surface-level textual patterns but also the deeper contextual and relational information necessary for accurate Relation Extraction.

6.3.4 *Target Representation*

A crucial component of our methodology is the definition of an effective *target representation* for the generative model. Each relation instance is expressed as a triplet consisting of a *head entity*, a *relation type*, and a *tail entity*. For example, consider the Italian sentence:

"Hideo Kojima has purchased a new house in Tokyo."

Following the CoNLL04 annotation scheme, the corresponding semantic relation can be encoded as:

[[Hideo Kojima, Live_In, Tokyo]]

In this representation, the triplet is organized in the format $[\mathtt{head}, \mathtt{relations}, \mathtt{tail}]$.
We adopt a compact serialization of relation triplets that minimizes the number of tokens in the output sequence, thereby enabling more efficient decoding. This notation naturally extends to multiple relations, for example:

$$[[\mathtt{head}_1, \mathtt{rel\_type}_1, \mathtt{tail}_1], [\mathtt{head}_2, \mathtt{rel\_type}_2, \mathtt{tail}_2], \ldots]$$

We employ this unified format consistently across both the CoNLL04 and SemEval 2010 Task 8 datasets.

6.4 TRAINING

Our approach employs Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically LoRA (Low-Rank Adaptation), to train FLAN-T5 XL models on the enriched data. This methodology balances computational efficiency with model performance, enabling effective training on the comprehensive knowledge-enriched datasets while maintaining practical resource requirements. The PEFT implementation utilizes LoRA with carefully optimized hyperparameters tailored to the FLAN-T5 XL architecture. For the XL model size, we employ a rank (r) of 8, alpha value of 16, and dropout rate of 0.1. This configuration provides an optimal balance between parameter efficiency and representational capacity, allowing the model to adapt to the Relation Extraction task while preserving the pre-trained language understanding capabilities. The LoRA configuration enables training with significantly reduced memory requirements compared to full fine-tuning, making it feasible

to work with the XL model variant (3B parameters) on standard GPU hardware.
Our experimental framework evaluates three distinct training modalities that vary in the amount and type of auxiliary information provided during training:

ENRICHED MODALITY (`enriched`): Incorporates the complete auxiliary information package, combining the instruction prompt, Entity Outlook , Sentence Outlook, and the original sentence. This represents the maximum information condition where models have access to all available reasoning traces and contextual knowledge.

PREFIX PROMPT + SENTENCE MODALITY (`prefix+sentence`): Utilizes the base Relation Extraction prompt combined with the original sentence, without additional knowledge. This serves as an intermediate condition that provides task-specific guidance while omitting the detailed auxiliary knowledge.

RAW SENTENCE MODALITY (`raw_sentence`): Presents only the original sentence without any additional prompts or auxiliary information. This represents the minimal information condition that most closely resembles real-world deployment scenarios.
The base prompt template employed across relevant modalities follows the structure, the list of relation types changes across the different domains:

Base Prompt

"List all the relations from the following sentence. Relation types: [OrgBased_In, Work_For, Located_In, Live_In, Kill]. Entity types involved: [Peop, Loc, Org]."

6.5 RESULTS

Table 29 and 30 presents the comprehensive comparison of standard training approaches on the CoNLL04 and SemEval Task 8 2010 datasets. We compare our approach with two generative state-fo-the-art methods: REBEL [137] and Wadhwa et al. [146].

The FLAN-T5 XL model trained on enriched data achieves 80.81% macro F1 and 79.7% micro F1, representing significant improvements over baseline approaches. The performance hierarchy clearly shows

Table 29: Results on CoNLL04 Relation Extraction Dataset

Model Configuration	F1 Micro	F1 Macro
Baseline Models		
REBEL	-	75.4
Wadhwa et al. GPT3 + Flan T5 large	-	80.7
Wadhwa et al. GPT4.1 + Flan T5 XL	75.73	77.15
Our Approach		
Flan-t5 XL enriched	**79.7**	**80.81**
Flan-t5 XL prefix + sentence	77.38	78.46
Flan-t5 XL raw sentence	76.38	77.67

the impact of auxiliary information: the enriched modality outperforms the prefix + sentence modality, which in turn exceeds the raw sentence modality. This progression validates the importance of comprehensive auxiliary knowledge in developing sophisticated Relation Extraction capabilities.

Particularly noteworthy is the model's ability to maintain strong performance even in the prefix + sentence condition, suggesting that the combination of task-specific prompting and the model's enhanced reasoning capabilities can partially compensate for the absence of detailed auxiliary knowledge. However, the consistent performance gains from the enriched modality underscore the value of comprehensive knowledge augmentation.

Table 30: Results on SemEval 2010 Task 8 Relation Extraction Dataset

Model Configuration	F1 Micro	F1 Macro
Baseline Models		
REBEL	75.00	69.14
Wadhwa et al. GPT4.1 + Flan T5 XL	84.80	85.77
Our Approach		
Flan-t5 XL enriched	**87.15**	**88.3**
Flan-t5 XL prefix + sentence	84.6	85.23
Flan-t5 XL raw sentence	83.71	84.23

The results on the SemEval 2010 Task 8 dataset further validate our approach and demonstrate its generalization capabilities across differ-

ent Relation Extraction benchmarks with more relation types (from 4 to 19) and adding directionality. Our FLAN-T5 XL enriched model achieves the highest performance, outperforming state-of-the-art results and the other two training modalities. This substantial improvement demonstrates the effectiveness of incorporating auxiliary knowledge during fine-tuning, even when compared to systems leveraging Large Language Models for knowledge generation. The performance pattern observed on SemEval 2010 Task 8 mirrors that of CoNLL04, with a consistent hierarchy across information modalities. The enriched modality maintains a significant advantage over the prefix + sentence configuration. Similarly, the prefix + sentence approach outperforms the raw sentence baseline. Notably, the absolute performance gains on SemEval 2010 Task 8 are more pronounced than on CoNLL04, with all configurations achieving higher F1 scores. This suggests that the SemEval dataset may be more amenable to our approach, possibly due to its different and more complex relation taxonomy, sentence structure characteristics, or the nature of entity-relation patterns present in the corpus.

6.5.1 *Knowledge component analysis*

To understand the individual contributions of the knowledge components in our enriched input representation we conduct a systematic knowledge component analysis. This analysis examines the relative importance of Sentence Outlook and Entity Outlook components in our simplified knowledge augmentation framework. The analysis systematically removes each knowledge component individually while maintaining all other aspects of the experimental setup. We use Flan-T5-XL as our test model due to its strong performance on both datasets, and we maintain consistent training procedures, hyperparameters, and evaluation metrics across all configurations. Each experiment excludes one specific knowledge type while maintaining the base sentence, allowing us to isolate the contribution of individual components:

- **Baseline (Full Enriched):** Complete input with both knowledge components: ⟨s,c + q⟩
- **Without Entity Outlook:** ⟨s,c⟩
- **Without Sentence Outlook:** ⟨s,q⟩

Table 31: Knowledge component analysis results showing the impact of removing individual knowledge outlook components from the enriched input. Each configuration excludes one specific knowledge type while maintaining the base sentence. ConLL04

Model Configuration	**F1 Micro**	**F1 Macro**
Flan-t5 XL enriched	**79.7**	**80.81**
Flan-t5 XL without Entity Outlook	77.88	79.01
Flan-t5 XL without Sentence Outlook	79.41	80.58

Table 32: Knowledge component analysis results showing the impact of removing individual knowledge outlook components from the enriched input. Each configuration excludes one specific knowledge type from the base sentence. SemEval 2010 task 8

Model Configuration	**F1 Micro**	**F1 Macro**
Flan-t5 XL enriched	**87.15**	**88.3**
Flan-t5 XL w/o Entity Outlook	86.14	87.16
Flan-t5 XL w/o Sentence Outlook	85.9	86.85

Tables 31 and 32 present the performance impact of removing each knowledge component individually across the ConLL04 and SemEval 2010 Task 8 datasets respectively.

RESULTS ON CONLL04: The results on ConLL04 reveal interesting patterns in component contribution. The baseline enriched model achieves 80.81% macro F1, serving as our reference point. Removing the **Entity Outlook** component causes a moderate performance degradation, with macro F1 dropping by 1.80 percentage points. This indicates that explicit entity semantics contribute meaningful signal for Relation Extraction, helping the model understand entity types, roles, and characteristics that inform relationship identification. Surprisingly, excluding the **Sentence Outlook** component results in minimal impact, with only a 0.23 percentage point decrease. This result suggests that for the ConLL04 dataset, the combination of base sentence and entity descriptions may already capture sufficient contextual information for effective Relation Extraction.

RESULTS ON SEMEVAL 2010 TASK 8: The SemEval dataset presents a different pattern of component importance. The baseline enriched model achieves 87.15% macro F1 on this dataset. Removing the **Entity Outlook** component results in a 1.01 percentage point decrease, demonstrating that entity semantics provide valuable but not critical information for this dataset's relation types. More notably, excluding the **Sentence Outlook** component leads to a 1.25 percentage point reduction, showing slightly greater impact than entity descriptions. This suggests that for SemEval's more semantically complex relations, contextual understanding plays a more significant role.

CROSS-DATASET ANALYSIS: The analysis of knowledge components highlights clear dataset-dependent contribution patterns. In **CoNLL04**, the *Entity Outlook* component yields the largest improvement compared to *Sentence Outlook*, whereas in **SemEval 2010**, the hierarchy is reversed, with *Sentence Outlook* slightly outperforming *Entity Outlook*. These contrasting tendencies reveal that dataset characteristics play a crucial role in determining which knowledge aspect is most beneficial.
CoNLL04 primarily deals with concrete, type-dependent relations such as *WORK_FOR* or *LOCATED_IN*, where explicit information about entity categories (e.g., Person–Organization pairs) directly constrains the relation space. In such settings, entity descriptions offer decisive cues, while sentence-level context often contributes redundantly. Conversely, SemEval 2010 focuses on more abstract and semantically complex relations, including *Cause–Effect* and *Component–Whole*, where contextual understanding becomes essential to disambiguate relational meaning, and entity types alone provide weaker predictive signals.
Despite these variations, both components show complementary behavior: each provides measurable gains, and their combined use yields consistently positive results across datasets. The modest performance gaps suggest that neither component dominates entirely, reinforcing their mutual relevance.
Overall, the findings indicate that Sentence Outlook and Entity Outlook contribute synergistically to Relation Extraction, though their relative importance shifts with the semantic nature of the dataset.

6.5.2 *Relation-Type Performance Analysis*

To gain deeper insights into our model's capabilities, we analyze performance at the relation-type level for both datasets using our best-performing model.

Table 33: Label-wise performance of Flan-t5 XL enriched on SemEval 2010 Task 8 dataset

Relation Type	Precision	Recall	F1
Message-Topic	87.02	95.02	90.84
Product-Producer	87.45	93.51	90.38
Instrument-Agency	87.82	87.82	87.82
Entity-Destination	93.79	93.15	93.47
Cause-Effect	88.27	91.77	89.99
Component-Whole	86.19	91.99	88.99
Member-Collection	84.92	91.85	88.25
Other	79.94	63.22	70.60
Entity-Origin	89.72	87.98	88.85
Content-Container	89.05	93.23	91.09

Table 34: Label-wise performance of Flan-t5 XL enriched on CoNLL04 dataset

Relation Type	Precision	Recall	F1
Live_In	87.84	66.33	75.58
OrgBased_In	84.15	69.70	76.24
Located_In	91.89	74.73	82.42
Work_For	78.67	77.63	78.15
Kill	89.80	93.62	91.67

The results reveal several interesting patterns across both datasets. For SemEval 2010 Task 8 (Table 33), the *Entity-Destination* relation achieves the highest F1 score, *Content-Container* and *Cause-Effect* relations show robust performance, likely due to their well-defined semantic patterns and abundant contextual cues. The *Other* category presents the most significant challenge, achieving only 70.60% F1 with notably low recall. This performance gap is expected given the heterogeneous nature of this miscellaneous category, which encompasses diverse relation types that do not fit into the predefined classes. For CoNLL04 (Table 34), the *Kill* relation achieves the highest F1 score. This exceptional performance, with high recall and precision, suggests that violent events are expressed through distinctive lexical and syntactic patterns that are readily identifiable even with limited training examples. The semantic clarity and explicit nature of kill relations provide strong

contextual indicators.
In contrast, the *Live_In* relation shows the lowest performance, with a significant gap between precision and recall. This pattern indicates that residential relationships are often expressed implicitly or through diverse linguistic constructions, making them harder to detect. Similarly, *OrgBased_In* demonstrates comparable challenges, as organizational location relationships may require world knowledge and can be stated in various indirect ways.

6.6 ERROR ANALYSIS

To gain deeper insights into the limitations and failure modes of our Relation Extraction system, we conduct a comprehensive error analysis on both test sets. This analysis reveals three primary error patterns that account for the majority of failures and inform future research directions. In particular best models trained on enriched input and their predictions are analyzed in the following section with few examples from both datasets.
Models most prevalent error type is entity type confusion, leading to systematic misclassification of relations. This pattern accounts for the majority of errors and demonstrates how cascading failures in entity recognition propagate through the Relation Extraction pipeline.

EXAMPLE FROM CONLL04: Consider the sentence: "*An enraged Nikita Khrushchev instructed Soviet ships to ignore President Kennedy's naval blockade during the Cuban missile crisis, but the order was reversed just hours before an inevitable confrontation, according to a new book.*"
The model incorrectly predicted a *Work_For* relation between Nikita Khrushchev and Soviet (treated as Organization), while the gold standard annotation specifies a *Live_In* relation between Nikita Khrushchev and Soviet (treated as Location). This error reveals a fundamental entity typing issue where "Soviet" functions as a metonym for both geographic and political entities.
The second major failure mode involves under-detection of relations, where the models fails to recognize valid semantic relationships between entities, particularly in cases requiring careful disambiguation of entity roles.

EXAMPLE FROM SEMEVAL2010: For the sentence: "*Skype, a free software, allows a hookup of multiple computer users to join in an online conference call without incurring any telephone costs.*"

The model predicted *Other* (no relation), while the gold standard specifies a *Member-Collection* relation between "users" and "hookup." This error demonstrates the model's failure to recognize the collective relationship where individual users form part of a larger group connection or assembly (the "hookup"). The term "hookup" in this context refers to a group connection or collection of linked participants, with "users" being the members of this collection. This suggests difficulty in semantic role disambiguation when entities can be interpreted in multiple ways.
The third significant error category involves over-generation of semantically plausible but annotation-guideline-incompatible relations. These cases reveal both the model's sophisticated semantic inference capabilities and the challenges of aligning model predictions with annotation schema constraints.

EXAMPLE FROM (CONLL04): In the sentence: "*Judith C. Toth says she returned for a fourth term in Maryland's House of Delegates because she couldn't find a better job.*"
The model predicted a *Work_For* relation between Judith C. Toth and House of Delegates, while the gold standard only annotates the *Org-Based_In* relation between House of Delegates and Maryland. The model reasonably infers employment from the phrase "returned for a fourth term" and the legislative context, demonstrating sophisticated semantic understanding. However, the annotation guidelines focus on organizational location rather than individual membership relations.

6.7 KNOWLEDGE DISTILLATION

Relation extraction (RE) systems have traditionally relied on auxiliary information during both training and inference phases, creating deployment challenges in resource-constrained environments where such information may be unavailable. Our Relation Extraction approach builds upon knowledge derived from a Large Language Model. However, in resource-constrained environments, such external knowledge may not be accessible, requiring models to operate solely on raw sentences. As confirmed by our experimental results, the absence of enriched knowledge leads to a measurable decline in generalization and overall extraction performance. In the following sections we explore a distillation technique based on Kullback-Leibler divergence that tries

to preserve reasoning capabilities while eliminating dependencies on auxiliary information during inference. The primary motivation for this work stems from the observation that models trained with enriched inputs, incorporating entities, contextual knowledge, and reasoning traces, develop sophisticated internal representations that extend beyond surface-level pattern matching. Knowledge distillation offers a promising solution by transferring the enhanced reasoning capabilities learned from enriched training data to student models that can operate efficiently on simplified inputs.

Given a teacher model T_{θ_T} trained on enriched input space $\mathcal{X}_{enriched} = \langle s, K \rangle$ where s represents the sentence, and K contains extracted knowledge components, and a student model S_{θ_S} operating on raw input space $\mathcal{X}_{raw} = \langle s \rangle$, the objective of reasoning distillation is to transfer the reasoning capabilities from T_{θ_T} to S_{θ_S} while maintaining performance on the target task.

The distillation process aims to minimize the divergence between the teacher's and student's output distributions:

$$\mathcal{L}_{KD} = \mathcal{L}_{KL}(T_{\theta_T}(\langle s, K \rangle), S_{\theta_S}(s)) \tag{12}$$

$\mathcal{L}_{KL}$ denotes the Kullback-Leibler divergence between teacher and student probability output distributions. The distillation training employs a custom loss function that combines knowledge distillation with standard cross-entropy. By combining the Kullback–Leibler divergence with the conventional cross-entropy loss, it encourages the student model to approximate the teacher's output distribution while retaining supervision from the original gold annotations. This formulation ensures a balanced transfer of knowledge, promoting a coherent alignment between the teacher's knowledge and the student's learning process.

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{KD} + (1 - \alpha) \cdot \mathcal{L}_{CE} \tag{13}$$

Here, $\mathcal{L}_{CE}$ represent the cross-entropy loss while α is a parameter that controls the balance between task loss and the knowledge distillation process.

6.7.1 *Experimental Configurations for Knowledge Distillation*

Our knowledge distillation experiments focus on two primary transfer scenarios that represent different levels of auxiliary information reduction:

ENRICHED → PREFIX + SENTENCE TRANSFER: The teacher model trained on fully enriched data (including Entity Outlook and Sentence Outlook) transfers knowledge to a student model that operates with only the task prompt and original sentence. This scenario represents a moderate reduction in auxiliary information while maintaining task-specific guidance.

ENRICHED → RAW SENTENCE TRANSFER: The teacher model transfers knowledge to a student model operating on completely raw sentences without any additional prompts or context. This represents the most challenging transfer scenario, requiring the student to internalize all task-specific knowledge and reasoning patterns from the teacher's soft predictions.

Both scenarios employ α values of 0.5 and temperature values of 2.0, providing balanced knowledge transfer that emphasizes both teacher guidance and ground truth supervision. The temperature setting enables the student to learn from the teacher's uncertainty patterns, capturing nuanced confidence distributions that reflect the complexity of Relation Extraction decisions.

6.8 KNOWLEDGE DISTILLATION RESULTS

This section presents the results of our knowledge distillation experiments, which evaluate whether models trained with reduced auxiliary information can achieve performance comparable to teacher models trained with enriched inputs. Table 35 and 36 present the comprehensive results across both scenarios on the CoNLL04 and SemEval 2010 Task 8 datasets, comparing distilled models against non-distilled baselines trained directly on the same inputs. **Bold** denotes that the distilled model outperforms the baseline. The overall trend across both benchmarks confirms the consistent advantage of the proposed distillation framework.

On the CoNLL04 dataset, which features a limited number of relations, the distilled models demonstrate notable gains under both input conditions—particularly when auxiliary knowledge is reduced. These improvements suggest that the distilled student effectively inherits the teacher's capacity to capture implicit relational dependencies, even when explicit contextual signals are absent. The results indicate that the distillation process successfully transfers high-level reasoning patterns into more compact representations, mitigating the loss of information introduced by simpler input formulations.

Table 35: Knowledge Distillation Results on CoNLL04 Relation Extraction Dataset

Model Configuration	F1 Micro	F1 Macro
Knowledge Distillation Results		
Flan-t5 XL prefix + sentence (distilled)	**79.2**	**80.31**
Flan-t5 XL raw sentence (distilled)	**77.69**	**78.82**
Comparison with Non-Distilled Baselines		
Flan-t5 XL prefix + sentence (baseline)	77.38	78.46
Flan-t5 XL raw sentence (baseline)	76.38	77.67

The SemEval-2010 Task 8 dataset, characterized by a broader and more semantically diverse set of relations, provides a complementary evaluation scenario. Here too, distilled models consistently outperform the corresponding baselines, confirming that the proposed framework generalizes effectively across different relational taxonomies. However, in this dataset, the remaining performance gap between the distilled models and the teacher indicates that there is still room for improvement in the knowledge transfer process.

The results demonstrate that knowledge distillation substantially enhances the performance of models trained solely on raw sentences. The distilled variants effectively inherit contextual and reasoning capabilities from the teacher model, enabling them to compensate for the absence of explicit auxiliary knowledge. This outcome confirms that distilled models can internalize transferable representations that improve their generalization capacity, maintaining robust performance despite the absence of knowledge-augmented input.

Table 36: Knowledge Distillation Results on SemEval 2010 task 8 Relation Extraction Dataset

Model Configuration	F1 Micro	F1 Macro
Knowledge Distillation Results		
Flan-t5 XL prefix + sentence (distilled)	**85.8**	**86.8**
Flan-t5 XL raw sentence (distilled)	**84.86**	**85.47**
Comparison with Non-Distilled Baselines		
Flan-t5 XL prefix + sentence	84.6	85.23
Flan-t5 XL raw sentence	83.71	84.23

6.9 ADAPTING THE APPROACH FOR ITALIAN RELATION EXTRACTION

While the foundational methodology presented in the previous sections demonstrates strong performance on English Relation Extraction benchmarks, the application of this approach to Italian requires careful consideration of linguistic and practical factors that necessitate specific adaptations. Italian exhibits morphological richness, syntactic flexibility, and semantic complexity that distinguish it from English in ways that directly impact Relation Extraction performance. The core principles of our knowledge augmentation framework—leveraging LLM reasoning to enrich training data with explicit semantic information—remain consistent across both English and Italian implementations. However, three key modifications were introduced to optimize performance for Italian Relation Extraction, each motivated by specific linguistic or practical considerations.

For the Italian experiments, we replace GPT-4.1 with Phi-4 [172], a 14-billion parameter open-source model, as our knowledge extraction engine. This decision is driven by both practical and performance considerations. Phi-4 demonstrates exceptional multilingual reasoning capabilities despite its relatively compact size, with particular strength in handling morphologically rich languages. Moreover, the open-source nature of Phi-4 provides greater transparency and reproducibility for research conducted on lower-resource languages, where community access to knowledge extraction tools is crucial for advancing the field. Italian's greater linguistic complexity necessitates an enhanced knowledge architecture. The language features extensive verbal conjugation and nominal inflection creating multiple surface forms for expressing similar relationships, relatively free word order allowing entities and their relations to appear in varied configurations, and rich vocabulary with idiomatic expressions providing multiple ways to encode relational information. To address these complexities, Given a sentence $s_i \in D$, where $D = \{s_1, s_2, \ldots, s_n\}$ represents an Italian corpus of input sentences, we augment our knowledge framework with two additional component: a set E of **Named Entity predictions** from three different state-of-the-art Italian and multilingual models and the **Relations Outlook** (v_i), which provides explicit guidance about potential relationships between entities. This three-component architecture proved essential for achieving competitive performance on Italian Relation Extraction, as the knowledge component analysis in Section 6.15 demonstrates, underscoring the importance of explicit relational reasoning

for handling Italian's linguistic complexity. Rather than applying the English methodology directly, we develop a linguistically-informed variant that maintains the core principles of knowledge augmentation while optimizing for Italian's morphological richness, syntactic flexibility, and semantic complexity.

6.10 DATASET CONSTRUCTION AND TRANSLATION METHODOLOGY

The foundation of our experimental work rests on a carefully constructed Italian version of the CoNLL04 dataset. This section details our sophisticated translation methodology and the resulting dataset characteristics.

6.10.1 *Translation Methodology*

The original CoNLL04 dataset statistics and description have been shown in section 6.2.1. The translation of the CoNLL04 dataset to Italian required addressing several critical challenges that go beyond simple text translation. The primary challenge lies in preserving the precise token-level annotations required for named entity recognition and Relation Extraction tasks while adapting the content to Italian morphological, syntactic, and semantic structures. This work employs a sophisticated hybrid approach for translating the ConLL04 English Relation Extraction dataset to Italian while preserving the crucial token-level annotations required for named entity recognition and Relation Extraction tasks. The translation process operates in three main phases: first, the complete English sentence is translated to Italian using X-ALMA [173], built upon ALMA-R by expanding support from 6 to 50 languages. It utilizes a plug-and-play architecture with language-specific modules, complemented by a carefully designed training recipe. In particular, a 8-bit quantized version due to resource limit constraints is used from the offical repository on Huggingface at *https://huggingface.co/mradermacher/X-ALMA-13B-Group2-GGUF*. The translator model generates fluent Italian text but disrupts the original token alignments. Second, to address the critical challenge of maintaining entity boundaries and types across languages—where direct token-to-token mapping fails due to morphological differences, word order changes, and varying translation lengths, the system employs OpenAI's GPT-4o-mini model [174] to perform intelligent entity alignment by analyzing both the original English tokens and their Italian coun-

terparts, then identifying which specific Italian tokens correspond to each English entity based on semantic understanding rather than positional heuristics. Finally, the system reconstructs the annotated dataset by mapping the spans of the identified Italian entity back to token indices. This step has the main goal to preserve entity types and relation labels while handling edge cases through fallback mechanisms that include proportional mapping and fuzzy string matching when exact alignment fails. This ensures that the resulting Italian dataset maintains the structural integrity necessary for training and evaluating Relation Extraction models. The comprehensive error handling and multi-stage validation process addresses the inherent complexities of cross-lingual annotation transfer in structured NLP datasets. In each split of the dataset, some translated sentences are removed due to the impossibility of maintaining relation labels. This case is represented by a few sentences that are not well translated, in which one or more entities that were in the relationship label are missing. This comprehensive approach ensures that the resulting Italian dataset maintains the structural integrity necessary for training and evaluating Relation Extraction models. The system addresses the inherent complexities of cross-lingual annotation transfer in structured NLP datasets through sophisticated error handling and validation processes. The final step is the assessment of the quality of automatic translation. A manual check is performed in order to find and remove possible translation errors. In each split of the dataset, some translated sentences are removed due to the impossibility of maintaining relation labels. This case is represented by a few sentences that are not well translated, in which one or more entities that were in the relationship label are missing. From the original dataset, we excluded 20 training sentences, 7 validation sentences, and 7 test sentences.

Table 37: Italian CoNLL04 version splits statistics

	samples	entities	relations
train	902	3284	1253
validation	224	848	325
test	281	1048	413
total	1407	5180	1991

The translation process resulted in the Italian version of the CoNLL04 dataset, Table 37 and Table 38 show statistics and distribution of the

Table 38: Relation types distribution across the Italian CoNLL04 dataset splits

relation type	train	validation	test
Vive_A	322	88	95
Situato_In	243	64	94
OrgLocata_In	256	64	103
Ha_ucciso	178	40	46
Lavora_per	254	69	75

new dataset. In the translation process, entity types and relation types distribution are maintained 26, 38.

The Italian relation types were carefully chosen to maintain semantic equivalence with the original English labels:

- *Live_In* → *Vive_A* (Lives in)
- *Located_In* → *Situato_In* (Located in)
- *OrgBased_In* → *OrgLocata_In* (Organization located in)
- *Kill* → *Ha_ucciso* (Has killed)
- *Work_for* → *Lavora_per* (Works for)

6.11 METHODOLOGY

This section presents our comprehensive methodology for enhancing Italian Relation Extraction through LLM-based knowledge augmentation. Due to the differences in terms of morphological, syntactic, and semantic structures between the two languages, this work adapt the approach used in English domains for Italian language keeping the core idea behind. As pre-announced in the previous sections the knowledge framework is extended to take in account the linguistic shift.
For a given a sentence $s_i \in D$, where $D = \{s_1, s_2, \ldots, s_n\}$ represents a corpus of Italian input sentences, we extend the input with two components:

$$X_{\text{enriched}} = \langle s_i, E_i, K_i \rangle \tag{14}$$

where s_i represents the original sentence, E_i represent the set of entities predictions and K_i contains the extracted knowledge components for sentence s_i. The knowledge K is composed of three complementary perspectives rather than two:

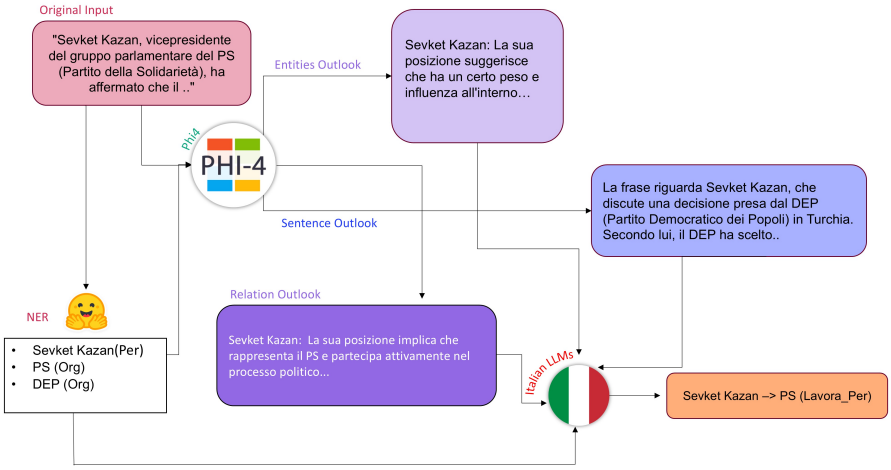


Figure 9: Overview of the proposed approach. Starting from the input sentence, the method augments the input with NER predictions and knowledge extracted from Phi-4. The enriched input is now composed by the original sentence, a set of Entity predictions, and the knowledge K, which is furthermore augmented with relation explanation outlook compared to the English approach.

$$K_i = q_i \oplus c_i \oplus v_i \tag{15}$$

where q_i represents the entity outlook extracted from sentence s_i, c_i represents the sentence outlook for s_i, v_i represent the additional component, the relation outlook and $\oplus$ denotes concatenation of the knowledge components.

6.11.1 *Named Entity Recognition Integration*

The first step in our pipeline involves extending the input space using state-of-the-art Named Entity Recognition (NER) models specifically designed for Italian.
NER is formulated as a sequence labeling task where each token in the input sequence is assigned a label that indicates its role in entity identification and classification. Given an input sentence $s = \{w_1, w_2, \ldots, w_n\}$ consisting of n tokens, the NER task aims to produce a corresponding label sequence $E = \{e_1, e_2, \ldots, e_n\}$ where each label $e_i \in \mathcal{Y}_\mathcal{T}$ encodes both the entity type and the token's position within the entity span.

For each input sentence in our dataset, we construct a comprehensive set of NER predictions E comprising annotations from three state-of-the-art multilingual and Italian-specific named entity recognition models:

1. **SpanMarker Multilingual Model** (`span-marker-multilingual-cased-multinerd`) [175]: A SpanMarker model fine-tuned on the MultiNERD dataset, providing robust multilingual entity recognition capabilities with particular strength in handling complex entity boundaries.

2. **BERT Italian NER** (`bert-italian-cased-ner`) [176]: A cased BERT model specifically trained for Italian NER on the WikiNER Italian dataset plus manually annotated Wikipedia paragraphs. This model is capable of recognizing four entity classes: PERSON, LOCATION, ORGANIZATION, and MISCELLANEOUS.

3. **Universal NER Italian** (`DeepMount00/universal_ner_ita`): An Italian adaptation of GLiNER [177] (Generalist Model for Named Entity Recognition using Bidirectional Transformer). This model leverages natural language descriptions to identify arbitrary entity types, providing flexibility in entity type recognition.

The GLiNER model uses the following entity types in Italian: *"persona"*, *"città"*, *"nazione"*, *"organizzazione"*, *"data"*, *"luogo"*, *"evento"*, *"prodotto"* (corresponding to *"person"*, *"city"*, *"nation"*, *"organization"*, *"date"*, *"location"*, *"event"*, *"product"* in English). Each model processes the tokenized Italian sentences independently, with predictions aligned to the original token boundaries. The resulting prediction set E provides diverse perspectives on entity recognition, creating a rich foundation for the subsequent knowledge extraction phase.

6.11.2 *Knowledge Extraction*

Given the extended input (s, E), the aim of this crucial step is to further extend the input by extracting knowledge **K** from an LLM. The knowledge **K** is composed of three different perspectives that are concatenated together to form a comprehensive semantic interpretation of a single dataset sample. For a given sentence $s_i \in D$, where D represents the entire corpus of the dataset, the knowledge is defined as:

$$K_i = q_i \oplus c_i \oplus v_i \tag{16}$$

where q_i represents the **Entities Outlook**, c_i represents the **Sentence Outlook**, and v_i represents the **Relations Outlook**. The $\oplus$ operator denotes concatenation of the different knowledge components. Each component addresses a specific aspect of semantic understanding:

ENTITIES OUTLOOK (q_i): This component focuses on providing detailed explanations of the main entities mentioned in the sentence. We prompt the LLM with:

"Spiega brevemente il significato dei soggetti principali menzionati per comprendere la frase: {s}"

(English: *"Briefly explain the meaning of the main subjects mentioned in order to understand the sentence: {s}"*)

This outlook helps the model understand the nature, characteristics, and significance of the entities involved in potential relations. For example, if the sentence mentions "Apple" and "Steve Jobs," this component would provide context about Apple being a technology company and Steve Jobs being its co-founder and former CEO.

SENTENCE OUTLOOK (c_i): This component provides broader contextual understanding of the entire sentence. We use the prompt:

"Spiegami molto brevemente la frase con il contesto necessario: {s}"

(English: *"Explain the sentence to me very briefly, providing the necessary context: {s}"*)

This outlook captures the overall semantic content of the sentence, including implicit information, temporal context, and situational factors that might influence entity relationships. It helps the model understand the scenario in which the entities are mentioned.

RELATIONS OUTLOOK (v_i): This component explicitly focuses on potential relationships between entities. We prompt:

"Basandoti sul testo e sulle predizioni di entità: Spiega brevemente le relazioni tra le entità menzionate nel testo. Testo: {s} Predizioni NER: {E}"

(English: *"Based on the text and entity predictions: Briefly explain the relationships between the entities mentioned in the text. Text: {s} NER predictions: {E}"*)

This outlook provides explicit guidance about potential relationships, helping the model understand how entities interact within the given context.

Example of knowledge augmented sample of Italian CoNLL04

Sentence: "Per il cantante e chitarrista Bob Boyd, di 57 anni, il Rancho è diventato la sua dimora quando ha lasciato Chattanooga, nel Tennessee, trentacinque anni fa."
Entities: (Bob Boyd, Persona) (Rancho, Luogo) (Chattanooga, Luogo), (57, Numero)
Entity Outlook Bob Boyd: È l'individuo centrale menzionato. Si identifica come cantante e chitarrista. L'età di 57 anni suggerisce che la narrazione si svolge attualmente, considerando anche gli eventi passati. Rancho: È un luogo significativo per Bob Boyd. La frase implica che il Rancho sia ora la sua casa principale o residenza. Chattanooga, nel Tennessee: Questa città è importante perché rappresenta il precedente luogo di residenza di Bob Boyd prima di trasferirsi al Rancho.
Sentence outlook: La frase riguarda Bob Boyd, un cantante e chitarrista di 57 anni. Trentacinque anni fa, Bob Boyd ha lasciato Chattanooga, una città nello stato del Tennessee negli Stati Uniti. Dopo essersi trasferito da lì, ha stabilito la sua residenza a un luogo chiamato "il Rancho".
Relation outlook: Bob Boyd è il centro del racconto. È un individuo che ha compiuto un cambiamento significativo nella sua residenza. - Ha lasciato Chattanooga, Tennessee, 35 anni fa, indicando una precedente residenza e un cambio di vita geografica. - Dopo aver lasciato Chattanooga, Bob Boyd si è stabilito al Rancho, che ora considera la sua dimora. Le relazioni mostrano una progressione di vita per Bob Boyd: dalla sua residenza originaria in Tennessee fino alla scelta di stabilirsi al Rancho, rivelando le tappe principali della sua storia abitativa e professionale
Target Representation: Bob Boyd -> Rancho (Vive_A) ; Bob Boyd -> Chattanooga (Vive_A) ; Rancho -> Chattanooga (Situata_A)

For knowledge extraction, we employ Phi-4 [172], a 14-billion parameter state-of-the-art open model. Phi-4 was selected due to its high quality and advanced multilingual reasoning capabilities, despite its relatively compact size compared to larger models like GPT-4. This choice balances performance with computational efficiency while ensuring high-quality knowledge generation in Italian. The knowledge extraction process creates an **enriched input space** $\langle s, E, k \rangle$ for each sentence, this enriched representation provides the foundation for train-

ing Relation Extraction models with enhanced semantic understanding.
Our approach treats Relation Extraction as a conditional text generation task, employing parameter-efficient fine-tuning strategies to adapt Italian Large Language Models to the Relation Extraction domain. We employ Low-Rank Adaptation (LoRA) [41] within the PEFT framework [178] for supervised fine-tuning (SFT) of several Italian LLMs.

6.11.3 *Target Representation*

The target representation for Italian Relation Extraction adopts a more compact linearization scheme compared to the English approach. This simplified notation reduces the number of special tokens in the generation stream, allowing models to focus representational capacity on linguistic content rather than structural markers, avoids potential conflicts with Italian punctuation conventions while maintaining clarity across various entity surface forms, and reduces the complexity of the generation task, which is particularly beneficial when working with morphologically rich languages where entity boundaries may be less clearly defined. Inspired by REBEL's triplet linearization approach [179], we design a compact notation that minimizes the number of tokens in the generation stream to enable efficient decoding. A relation triplet is represented using the following notation:

Head Entity -> Tail Entity (Relation type) (17)

For multiple relations within a single sentence, we separate them using the semicolon character "**;**". We fine-tune several Italian LLMs using the LoRA strategy to learn to generate the target representation. Additionally, we fine-tune mREBEL$_{32}$ [139], a multilingual version of REBEL [179], and Wadhwa et al. [146] as a baselines comparison. All models are trained for 10 epochs, best performing model on the validation set is saved.
Hardware: NVIDIA GeForce RTX 3090 with 24GB memory and AMD Ryzen 9 5900X 12-Core Processor.

6.12 ITALIAN TRAINING CONFIGURATION

This section details our comprehensive experimental setup, including the various input configurations, model selection criteria, and evaluation strategies employed to assess the effectiveness of our approach.

We evaluate multiple Italian Large Language Models under different input configurations to assess the effectiveness of our generative Relation Extraction framework. We conduct experiments using three configurations:

- **Enriched Input**: Complete input including sentence, entity predictions, and background knowledge:

$$\langle s, E, K \rangle$$

- **Raw Input**: : Input containing only the source sentence:

$$\langle s \rangle$$

- The **Enriched-Raw** configuration: Model fine-tuned on enriched input but evaluated using only raw sentence input at inference time

Our experimental evaluation encompasses several Italian Large Language Models, each representing different architectural approaches and training methodologies.

Example of prompt with knowledge augmented sample of Italian CoNLL04

Estrai le relazioni tra le entità dalla seguente frase:
Frase:**<s>**
Relazioni possibili:
[Vive_a, OrgLocata_In, Situato_In, Ha_ucciso, Lavora_per]
Possibili entità: **<E>**
Contesto: **<k>**
Identifica tutte le relazioni usando il formato:
Entità1 - > Entità2 (Relazione)

LLAMANTINO-3-ANITA (8B): A fine-tuned version of Meta's LLaMA-3 (8B) [180, 181], adapted through Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) to align with user preferences and reduce biases.

MINERVA-7B: Created by Sapienza NLP in collaboration with FAIR, CINECA, and Italy's National Recovery and Resilience Plan (PNRR)

[182]. This model is trained from scratch on 2.5 trillion tokens (50% Italian) and further enhanced through instruction tuning and safety layers, representing a ground-up approach to Italian language modeling.

VELVET-14B: Developed by Almawave [183], this model is part of a family of multilingual LLMs that includes Italian and is built on a proprietary architecture. It represents an industry-grade solution for Italian NLP tasks.

MREBEL32: A multilingual version of REBEL [139, 179], included as a baseline representing the previous state-of-the-art in generative Relation Extraction. This model provides a point of comparison with established sequence-to-sequence approaches.

WADHWA ET AL. : To compare our approach with state of the art [146], we fine-tune Flan-T5 XL to generate relation explanations along with the target relation labels.

6.13 RESULTS ON ITALIAN CONLL04

This section presents a comprehensive analysis of our experimental results, demonstrating the effectiveness of our LLM-enhanced Relation Extraction approach across different models and input configurations. Table 39 presents the performance comparison across different Italian language models and input configurations. The results reveal several important patterns that validate our approach and provide insights into the effectiveness of knowledge augmentation for Relation Extraction.

LLaMAntino-3 demonstrates superior performance when trained and evaluated on enriched input, achieving 70.6% macro F1 score. This represents a significant improvement over Wadhwa et al. approach and both Minerva-7B (59.6%) and Velvet-14B (60.2%), despite LLaMAntino-3 being a smaller 8B parameter model compared with the last two models. The results indicate that model architecture and training methodology are more critical factors than pure parameter count for this task. The strong performance of mREBEL demonstrates that sequence-to-sequence models, which were previously state-of-the-art for this task, can achieve comparable results to Large Language Models (LLMs). Additionally, mREBEL and Wadhwa et al. benefits from enriched input. However, Velvet-14B exhibits the opposite behavior, performing better

Table 39: Performance comparison of supervised fine-tuned Italian LLMs on Italian CoNLL 04. Input configurations: (enriched) includes entity predictions and background knowledge; (raw) uses only sentence text; (enriched-raw) represents models trained on enriched data but evaluated with raw input only.

Model Configuration	**F1 Micro**	**F1 Macro**
mREBEL (enriched)	62.7	63.9
mREBEL (raw)	58.1	59.6
mREBEL (enriched-raw)	49.7	49.12
Wadhwa et al. Phi4 + Flan-t5 XL (enriched)	56.3	58.4
Wadhwa et al. Phi4 + Flan-t5 XL (raw)	55.3	55.5
Wadhwa et al. Phi4 + Flan-t5 XL (enriched-raw)	12.9	13.75
Minerva-7B (enriched)	57.2	59.6
Minerva-7B (raw)	55.6	57.9
Minerva-7B (enriched-raw)	48.9	51.0
Velvet-14B (enriched)	56.9	60.2
Velvet-14B (raw)	63.0	65.2
Velvet-14B (enriched-raw)	42.6	46.4
LLaMAntino-3 (enriched)	**68.5**	**70.6**
LLaMAntino-3 (raw)	58.4	62.1
LLaMAntino-3 (enriched-raw)	61.1	64.9

with raw input (65.2%) than with enriched input (60.2%). This suggests the model may be overfitting to the auxiliary information provided in the enriched input. Comparing LLaMAntino-3 configurations reveals the substantial benefit of enriched input during training. The model trained on enriched data (70.6% macro F1) significantly outperforms the same model trained solely on raw sentences (62.1% macro F1). This demonstrates the value of incorporating entity predictions and background knowledge in the training process. The enriched-raw configuration yields particularly interesting results, achieving 64.9% macro F1 despite using only raw sentence input at inference time. This performance exceeds that of the model trained exclusively on raw input (62.1% macro F1), suggesting an interesting implicit knowledge distillation during training. The model appears to internalize reasoning patterns from the enriched training data, enabling improved performance

even when auxiliary information is unavailable at inference time.

6.13.1 *Relation-Type Performance Analysis*

To gain deeper insights into our model's capabilities, we analyze performance at the relation-type level using our best-performing model, LLaMAntino-3 with enriched input.

Table 40: Label-wise performance of best model: LLaMAntino-3-8B (enriched)

Relation Type	**Precision**	**Recall**	**F1**
Vive_a	69.9	61.05	65.17
OrgLocata_In	71.95	63.44	67.42
Situato_In	60.63	60.63	60.63
Lavora_Per	62.5	80.0	70.17
Ha_ucciso	86.0	93.4	89.58

Table 40 reveals several interesting patterns in relation-type *Ha_ucciso* (Kill), which achieves the highest F1 score of 89.58% despite being the least represented relation type in the training set. This result is particularly noteworthy because it demonstrates the model's ability to learn effectively from limited examples when provided with appropriate knowledge augmentation. However, Kill relations have distinctive semantic patterns that are easily recognizable; sentences containing kill relations often provide clear contextual indicators, the background knowledge may be particularly helpful for this relation type. The *Situato_In* (Located in) relation shows the lowest performance at 60.63% F1, with perfect balance between precision and recall. Many geographic relationships are implicit and require world knowledge and can be expressed in various indirect ways. These results validate our approach of treating Relation Extraction as a conditional text generation task and demonstrate the effectiveness of supervised fine-tuning on Italian language models for this domain. Error and Knowledge component analysis, presented in next sections are performed on the best model LLaMAntino-3.

6.14 ERROR ANALYSIS

To gain deeper insights into the limitations and failure modes of our approach, we conduct a comprehensive error analysis using the best-performing model, LLaMAntino-3 with enriched input. This analysis reveals systematic patterns that inform future research directions and potential improvements. The model demonstrates a tendency toward over-generation, particularly struggling with complex sentences containing multiple entities where it produces semantically plausible but factually incorrect relations. This represents the most common error type in our analysis.

OVER GENERATION EXAMPLE: Consider the sentence: "*Nikita Chruščëv, infuriato, ordinò alle navi dell'Unione Sovietica di ignorare il blocco navale del Presidente Kennedy durante la crisi dei missili cubani*"

(English: "Nikita Khrushchev, enraged, ordered Soviet Union ships to ignore President Kennedy's naval blockade during the Cuban missile crisis")

The model incorrectly generated four identical *Ha_ucciso* (Kill) relations between Khrushchev and Kennedy, while missing the correct *Vive_A* (Lives in) relation between Khrushchev and the Soviet Union.

This error demonstrates the model's tendency to infer dramatic but incorrect relations from contextual conflict scenarios.

UNDER DETECTION EXAMPLE: For the sentence: "*MILANO, Italia (AP)*" (Milan, Italy (AP)) The model correctly identified organizational relations for the Associated Press but failed to extract the fundamental *Situato_In* (Located in) relation between Milan and Italy.

This error suggests difficulty with implicit geographic knowledge in simple locative constructions. Despite the obvious geographic relationship, the model failed to recognize this basic world knowledge fact, possibly due to the abbreviated format or the presence of the "(AP)" organizational marker.

A particularly interesting error category involves the model generating relations that are factually correct but fall outside the defined schema."

OUT-OF-DOMAIN EXAMPLE: In the sentence: "*King venne ucciso il 4 aprile del 1968 a Memphis, nel Tennessee*" (King was killed on April 4, 1968 in Memphis, Tennessee)

The model correctly identified the *Situato_In* relation between Memphis and Tennessee, but additionally generated correct *Evento* (Event)

relations involving the date "4 aprile del 1968" with Memphis. The *Evento* relation type does not exist in the CoNLL04 schema, demonstrating the model's tendency to create novel relation categories when encountering temporal-spatial contexts. These patterns indicate that while the generative approach successfully captures complex relational semantics, it requires improved calibration mechanisms, particularly for handling entity-dense contexts and fundamental geographic relations.

6.15 KNOWLEDGE COMPONENT ANALYSIS

To understand the individual contributions of different knowledge components in our enriched input representation, we conduct a systematic knowledge component analysis. This analysis provides crucial insights into which aspects of knowledge augmentation are most beneficial for Relation Extraction performance.
The knowledge component analysis systematically removes each knowledge component individually while maintaining all other aspects of the experimental setup. We use LLaMAntino-3 as our test model due to its superior overall performance, and we maintain the same training procedures, hyperparameters, and evaluation metrics across all configurations. Each experiment excludes one specific knowledge type while maintaining entity predictions and the base sentence, allowing us to isolate the contribution of individual components:

- **Baseline (Full Enriched):** Complete input with all three knowledge components: $\langle s, E, q \oplus c \oplus v \rangle$
- **Without Entity Outlook:** $\langle s, E, c \oplus v \rangle$
- **Without Sentence Outlook:** $\langle s, E, q \oplus v \rangle$
- **Without Relation Outlook:** $\langle s, E, q \oplus c \rangle$

Table 41 presents the performance impact of removing each knowledge component individually, with the baseline enriched model achieving 70.6% macro F1 serving as our reference point.

The analysis results reveal distinct contribution patterns for each knowledge component, establishing a clear hierarchy of importance: Removing the **Sentence Outlook** component causes the most severe performance degradation. This dramatic decline indicates that contextual sentence understanding is fundamental to Relation Extraction performance. The sentence outlook provides essential discourse-level information that enables the model to disambiguate entity relationships

Table 41: Knowledge component analysis results showing the impact of removing individual knowledge outlook components from the enriched input. Each configuration excludes one specific knowledge type while maintaining entity predictions and the base sentence. Italian CoNLL04

Model Configuration	F1 Micro	F1 Macro
LLaMAntino-3 (enriched)	**68.5**	**70.6**
LLaMAntino-3 without Entity Outlook	60.1	62.6
LLaMAntino-3 without Sentence Outlook	49.0	50.6
LLaMAntino-3 without Relation Outlook	63.1	65.7

within specific contextual frameworks."
Excluding entity explanations results in an 8.0 percentage point decrease (70.6% $\rightarrow$ 62.6%), demonstrating the importance of explicit entity semantics in Relation Extraction. This outlook provides detailed information about entity types, roles, and significance helping identify which entities are likely to be related and how. Removing relation-specific explanations leads to a 4.9 percentage point reduction (70.6% $\rightarrow$ 65.7%), showing the smallest but still meaningful impact among the three components. While relation outlook provides valuable relational reasoning guidance, the model appears capable of inferring many relations from entity and sentence context when this component is absent. The knowledge component analysis reveals a clear hierarchy of knowledge component importance: **Sentence Outlook > Entity Outlook > Relation Outlook**. This hierarchy suggests that: Contextual understanding is paramount for Relation Extraction, as sentences provide the situational framework within which entities interact, Entity semantics serve as the foundation for identifying potential relation participants and their characteristics, and Explicit relational reasoning provides incremental benefits but is less critical when strong contextual and entity understanding exists. These findings highlight the differential contribution of each component to the overall system performance. The results also suggest potential optimization strategies, where computational resources could be prioritized toward generating high-quality sentence and entity outlooks when resource constraints exist.

RQ 6.1 - Which complementary knowledge component can be obtained by Large Language Models to augment training data for Relation Extraction?

The main complementary components are Entity Outlook and Sentence Outlook for the English language, while for Italian, also the Relation Outlook should be considered.

6.16 KNOWLEDGE DISTILLATION FOR ITALIAN RELATION EXTRACTION

Following the successful application of knowledge distillation in English datasets, we extend this approach to Italian Relation Extraction to investigate whether the knowledge transfer mechanisms remain effective across linguistic boundaries. This section presents our distillation experiments on the Italian CoNLL04 dataset and analyzes the transferability of reasoning capabilities learned from enriched training data to models operating on raw Italian text. The knowledge distillation framework for Italian follows the same configuration established in Section 6.7. The student model is trained to operate on raw Italian sentences $\langle s \rangle$ while learning from both the ground truth labels and the teacher's probability output distributions. Given the linguistic complexity of Italian—with its rich morphology, flexible word order, and extensive verbal conjugation—the distillation process faces additional challenges compared to English. Italian's morphological richness means that entity boundaries and relation expressions can manifest in more varied surface forms, potentially complicating the knowledge transfer process. We employ the same custom loss function defined in Equation 13, with $\alpha = 0.5$ and temperature $\tau = 2.0$, maintaining consistency with the English experiments. Table 42 presents the

Table 42: Knowledge Distillation Results on Italian CoNLL04 Dataset

Model Configuration	F1 Micro	F1 Macro
LLaMAntino-3 raw (distilled)	**63.06**	**65.98**
LLaMAntino-3 raw (baseline)	58.4	62.1

distillation results on the Italian CoNLL04 dataset, comparing the distilled student model against the baseline model trained directly on raw sentences without teacher guidance. The distilled model achieves 65.98% macro F1 and 63.06% micro F1, representing substantial im-

provements over the non-distilled baseline (62.1% macro F1, 58.4% micro F1). This translates to a 3.88 percentage point gain in macro F1 and a 4.66 percentage point gain in micro F1, demonstrating effective knowledge transfer despite the complete absence of auxiliary information during training and inference.

RQ 6.2 -What is the effect of knowledge distillation when considering only model that takes as input only the raw sentence?

The knowledge distillation improves the performance of those models that consider only raw text, denoting an even more relevant impact on scenarios (Italian) characterized by less avaiable resources.

6.17 DISCUSSION AND CONCLUSIONS

This chapter presented a comprehensive framework for Relation Extraction that operates in two complementary phases: first enriching sentences with auxiliary knowledge, then distilling this knowledge to enable models that operate solely on raw text. This two-stage approach addresses both the performance gains achievable through explicit reasoning and the practical deployment constraints of real-world systems. The knowledge augmentation phase systematically extends training samples with multi-perspective information extracted from Large Language Models. For English datasets, this involves Entity Outlook and Sentence Outlook; for Italian, an additional relation-specific component proves essential given the language's morphological complexity. Experimental results demonstrate substantial improvements: FLAN-T5 XL achieves 80.81% macro F1 on English CoNLL04 with enriched input, while LLaMAntino-3 reaches 70.6% on Italian CoNLL04, significantly outperforming raw baseline configurations across both languages.

The knowledge component analysis studies reveal critical insights into knowledge component hierarchies. For English CoNLL04, entity descriptions provide the strongest signal, while Italian experiments show sentence contextualization as paramount. These language-specific patterns suggest that morphologically rich languages like Italian benefit more from discourse-level understanding, whereas entity-type constraints prove more decisive in English Relation Extraction. From a linguistic perspective, the observed cross-lingual differences between

English and Italian Relation Extraction can be attributed to fundamental typological distinctions between the two languages. English exhibits relatively rigid word order and limited inflectional morphology, which makes syntactic position and entity-type constraints highly informative for identifying relational patterns. As a result, entity-centric knowledge, such as type descriptions and semantic roles—provides a strong inductive bias for English Relation Extraction. In contrast, Italian is a morphologically rich language characterized by freer word order, extensive verbal inflection, agreement phenomena, and frequent omission of explicit subjects. These properties weaken the reliability of surface-level cues and local syntactic patterns, increasing the importance of broader sentential and discourse-level context to correctly infer relations. Consequently, sentence-level contextualization and relation-specific knowledge play a more critical role in Italian, helping models resolve ambiguities arising from inflectional variation and implicit arguments. However, the computational overhead and auxiliary information dependency of enriched models limit practical deployment. The knowledge distillation phase addresses this limitation by transferring reasoning capabilities from enriched teacher models to student models operating on raw text alone. The distillation results validate that sophisticated reasoning patterns learned through enriched training successfully transfer to raw-input models. Distilled models retain approximately 93-97% of teacher performance while eliminating inference-time dependencies on entity predictions and LLM-generated knowledge. This enables practical deployment in resource-constrained environments where auxiliary information may be unavailable. Beyond efficiency considerations, the proposed distillation framework addresses realistic deployment scenarios in which auxiliary knowledge available at training time cannot be reliably accessed at inference time. This situation arises in multiple real-world domains where data availability, privacy, or system integration constraints prevent the use of enriched inputs during prediction. For instance, in clinical text processing, enriched Relation Extraction models may be trained using structured patient metadata, clinical ontologies, or LLM-generated medical background knowledge provided by specialized healthcare models. However, at inference time, patient consent restrictions, institutional data-sharing policies, or regulatory constraints may prohibit access to such auxiliary information. In these settings, knowledge distillation enables the transfer of domain-specific relational understanding into models that operate exclusively on raw clinical notes, preserving performance while respecting privacy and compliance requirements.

Similar constraints emerge in industrial and enterprise settings, such as legal document analysis, financial reporting, or intelligence monitoring, where enriched models can leverage proprietary knowledge bases, entity linking systems, or external APIs during training, but deployment environments may lack network access, licensing permissions, or computational resources to support these components. Distilled models provide a robust alternative by embedding structured relational patterns into lightweight architectures, allowing consistent inference under strict latency, cost, or security constraints.

Conversely, there exist application scenarios where fully enriched models remain preferable and knowledge distillation is not optimal. In dynamic or high-stakes environments—such as investigative journalism, scientific literature mining, or decision-support systems for policy analysis—relations may depend on continuously evolving background knowledge, nuanced entity attributes, or up-to-date contextual information. In such cases, inference-time access to enriched representations and external knowledge sources allows models to adapt to new facts, emerging entities, or shifting relational schemas, which cannot be fully captured through distillation alone. Here, enriched models offer greater interpretability, flexibility, and responsiveness, at the cost of higher computational overhead. These considerations highlight that knowledge distillation extends beyond model compression by enabling models trained with enriched supervision to be deployed in settings where auxiliary knowledge is unavailable at inference time. This capability allows practitioners to choose between enriched and distilled models according to operational constraints, regulatory environments, and application criticality, reinforcing the practical relevance of the proposed framework across diverse real-world deployment scenarios. Future research should explore multi-teacher distillation frameworks leveraging specialized teachers with different knowledge focuses, adaptive distillation strategies that dynamically adjust transfer intensity based on instance difficulty, and cross-lingual applications for low-resource languages. The demonstrated ability to maintain sophisticated reasoning while reducing auxiliary dependencies opens new possibilities for efficient yet powerful NLP systems across diverse deployment scenarios.

7

CONCLUSIONS AND FUTURE DIRECTIONS

This thesis has explored advanced methodologies for Named Entity Recognition (NER) and Relation Extraction (RE) in resource-constrained environments, addressing core challenges related to data imbalance, domain transfer, and the integration of external knowledge through Large Language Models. The overarching objective was to design approaches that balance effectiveness, efficiency, and accessibility, enabling robust information extraction even under limited computational and data resources.

The first part of the work concentrated on NER, beginning with the challenge of class imbalance in sequence labeling. The study demonstrated that optimal rebalancing strategies are highly dependent on model architecture: while span-based models like SpanCategorizer benefit from sophisticated sentence-level techniques such as smoothed count resampling, transformer-based models like BERT achieve superior performance with simple random oversampling. Critically, the investigation revealed that random undersampling fails catastrophically across all configurations, highlighting the reliance of modern NER systems on the sentence composition. These findings underscore that the choice of rebalancing strategy must be tailored to the underlying model's capacity to capture linguistic patterns, while also exposing the computational costs of data augmentation approaches that scale poorly across domains.

Subsequent investigations addressed the problem of adapting NER systems to new domains with scarce annotated data. The proposed approach showed that high-quality adaptation can be achieved with substantially reduced computational demands compared to conventional fine-tuning techniques. Attention-based representations emerged as especially transferable, supporting the idea that such features capture domain-agnostic linguistic knowledge. This provides a more sustainable and resource-efficient alternative for real-world deployment, where scalability and energy efficiency are increasingly significant.

The second part of the thesis introduced a comprehensive framework for knowledge-enhanced Relation Extraction for English and Italian texts, which combines knowledge augmentation with distillation. The enrichment phase expanded the input data with auxiliary contextual

and explanatory information generated by Large Language models, leading to substantial improvements in relational understanding across languages. The experiments reveal that the relative contribution of different forms of knowledge—such as entity descriptions or contextual cues—varies with linguistic characteristics. In particular, languages with richer morphology appear to benefit more from discourse-level contextualization, suggesting that knowledge augmentation strategies should be sensitive to language typology and structural complexity.

The subsequent distillation phase enabled the transfer of reasoning capabilities from enriched models to lighter, standalone systems that operate solely on raw input. This process preserved most of the benefits of knowledge enrichment while removing dependencies on external resources at inference time, thereby improving efficiency and practical applicability.

Despite these promising results, a few limitations should be acknowledged. The Italian corpus, although carefully designed, may not fully reflect the stylistic and syntactic diversity of naturally occurring language. Furthermore, while distillation substantially reduces computational costs, it still entails a modest performance trade-off that should be considered when selecting models for production use. Overall, this work contributes to a deeper understanding of how efficiency, adaptability, and knowledge integration can be jointly pursued in information extraction under a low-resource scenario.

With respect to generalization, the proposed framework is designed to be largely language- and domain-agnostic, as it does not rely on language-specific rules or handcrafted resources. The core components (sentence level augmentation, auxiliary knowledge generation, and distillation) can be applied to other languages and domains provided that a minimal amount of annotated data is available and that suitable language models exist to generate auxiliary information. Empirical evidence from English and Italian suggests that while the relative effectiveness of different knowledge components may vary with linguistic and structural properties, the overall paradigm remains stable across settings. Moreover, the framework naturally extends to more complex information extraction scenarios, such as document-level or multi-hop relation extraction, by leveraging LLMs to enrich larger textual units (e.g., paragraphs or documents) with auxiliary knowledge before distilling these enhanced representations into efficient models. While additional modeling choices are required to capture long-range dependencies and cross-sentence interactions, the underlying separation between enriched learning and efficient inference remains appli-

cable, supporting the generalizability of the approach beyond the specific tasks and datasets explored in this thesis. This research opens several promising directions for further investigation. A first natural extension concerns the transition from sentence-level to document-level modeling for both Named Entity Recognition and Relation Extraction. Expanding the current framework to operate at the document scale would enable models to capture cross-sentence dependencies, coreference phenomena, and multi-hop relational patterns that are crucial for comprehensive text understanding. Such an advancement would move beyond isolated sentence analysis toward a more global representation of entities and their interconnections within discourse.

Another potential research direction involves extending the framework to handle n-ary relations, where two or more entities participate simultaneously in a single relational instance. This would allow the system to model complex semantic structures such as event participation, causality chains, or compositional relations that cannot be adequately represented through traditional binary relation extraction schemes. Integrating such relational complexity would significantly enhance the expressive power of current models.

Future work could also explore the development of adaptive knowledge augmentation mechanisms capable of tailoring the quantity and nature of generated auxiliary information to the difficulty of each input instance. Coupled with selective distillation techniques that prioritize the transfer of the most generalizable reasoning patterns, this direction would make the overall process more efficient and context-sensitive, improving both interpretability and computational scalability.

Finally, adapting the proposed methods to low-resource and multilingual settings remains a key challenge. Cross-lingual transfer, the use of multilingual Large Language Models, could together enable broader applicability across languages and domains with limited labeled data. Advancing in this direction would strengthen the inclusiveness and practical impact of knowledge-augmented extraction systems, making them more accessible for real-world applications.

BIBLIOGRAPHY

[1] Salman Naseer, Mudasar Ghafoor, Sohaib, Sohaib Khalid Alvi, Anam Kiran, Shafique Ur Rehman, Ghulam Murtaza, Jehlum Campus, and Pakistan Jehlum. "Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance." In: (Jan. 2022).

[2] J. Diaz-Garcia and Julio Lopez. "A survey on cutting-edge relation extraction techniques based on language models." In: (Nov. 2024). DOI: 10.48550/arXiv.2411.18157.

[3] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. "A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers." In: *ACM Comput. Surv.* 56.11 (July 2024). ISSN: 0360-0300. DOI: 10.1145/3674501. URL: https://doi.org/10.1145/3674501.

[4] Yifan Zheng, Yikai Guo, Zhizhao Luo, Zengwen Yu, Kunlong Wang, Hong Zhang, and Hua Zhao. "A Survey on Document-Level Relation Extraction: Methods and Applications." In: *Proceedings of the 3rd International Conference on Internet, Education and Information Technology (IEIT 2023)*. Atlantis Press, 2023, pp. 1061–1071. ISBN: 978-94-6463-230-9. DOI: 10.2991/978-94-6463-230-9_128. URL: https://doi.org/10.2991/978-94-6463-230-9_128.

[5] Chuyuan Wei, Jinzhe Li, Zhiyuan Wang, Shanshan Wan, and Maozu Guo. "Graph Convolutional Networks Embedding Textual Structure Information for Relation Extraction." In: *Computers, Materials & Continua* 79.2 (2024).

[6] Daya C. Wimalasuriya and Dejing Dou. "Ontology-based information extraction: An introduction and a survey of current approaches." In: *Journal of Information Science* 36.3 (2010), pp. 306–323. DOI: 10.1177/0165551509360123. eprint: https://doi.org/10.1177/0165551509360123. URL: https://doi.org/10.1177/0165551509360123.

[7] Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. "A comprehensive survey on automatic knowledge graph construction." In: *ACM Computing Surveys* 56.4 (2023), pp. 1–62.

[8] Nikola Milosevic and Wolfgang Thielemann. "Relationship extraction for knowledge graph creation from biomedical literature." In: (Jan. 2022). DOI: 10.48550/arXiv.2201.01647.

[9] Aoran Li, Xinmeng Wang, Wenhuan Wang, Anman Zhang, and Bohan Li. "A survey of relation extraction of knowledge graphs." In: *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer. 2019, pp. 52–66.

[10] Haoze Yu, Haisheng Li, Dianhui Mao, and Qiang Cai. "A relationship extraction method for domain knowledge graph construction." In: *World Wide Web* 23.2 (2020), pp. 735–753.

[11] Xiang Chen, Lei Li, Shuofei Qiao, Ningyu Zhang, Chuanqi Tan, Yong Jiang, Fei Huang, and Huajun Chen. "One model for all domains: Collaborative domain-prefix tuning for cross-domain ner." In: *arXiv preprint arXiv:2301.10410* (2023).

[12] Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. "More data, more relations, more context and more openness: A review and outlook for relation extraction." In: *arXiv preprint arXiv:2004.03186* (2020).

[13] GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. "Exploring various knowledge in relation extraction." In: *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*. 2005, pp. 427–434.

[14] Yee Seng Chan and Dan Roth. "Exploiting background knowledge for relation extraction." In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, pp. 152–160.

[15] Feng Chen and Yujian Feng. "Chain-of-thought prompt distillation for multimodal named entity recognition and multimodal relation extraction." In: *arXiv preprint arXiv:2306.14122* (2023).

[16] Michael A Johnson, Liu Chen, and Maria Rodriguez. "Sentence-level resampling for named entity recognition with imbalanced data." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023), pp. 1245–1256.

[17] Ralph Weischedel et al. "OntoNotes Release 5.0." Version V1. In: (2013). DOI: 11272.1/AB2/MKJJ2R. URL: https://hdl.handle.net/11272.1/AB2/MKJJ2R.

[18] Erik F. Tjong Kim Sang and Fien De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." In: *Proc. of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 142–147. URL: https://aclanthology.org/W03-0419.

[19] Simone Tedeschi and Roberto Navigli. "MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation)." In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. July 2022, pp. 801–812. DOI: 10.18653/v1/2022.findings-naacl.60. URL: https://aclanthology.org/2022.findings-naacl.60/.

[20] David Nadeau and Satoshi Sekine. "A survey of named entity recognition and classification." In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.

[21] Michal Konkol and Miloslav Konopík. "Segment representations in named entity recognition." In: *International conference on text, speech, and dialogue*. Springer. 2015, pp. 61–70.

[22] Sarah Shafqat, Hammad Majeed, Qaisar Javaid, and Hafiz Farooq Ahmad. "Standard ner tagging scheme for big data healthcare analytics built on unified medical corpora." In: *Journal of Artificial Intelligence and Technology* 2.4 (2022), pp. 152–157.

[23] Lev Ratinov and Dan Roth. "Design challenges and misconceptions in named entity recognition." In: *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*. 2009, pp. 147–155.

[24] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. "A Survey on Deep Learning for Named Entity Recognition." In: *IEEE Transactions on Knowledge and Data Engineering* 34.1 (2020), pp. 50–70.

[25] Maria Carmela Cariello, Alessandro Lenci, and Ruslan Mitkov. "A comparison between named entity recognition models in the biomedical domain." In: *Proceedings of the Translation and Interpreting Technology Online Conference*. 2021, pp. 76–84.

[26] Anshita Khandelwal, Alok Kar, Veera Raghavendra Chikka, and Kamalakar Karlapalem. "Biomedical NER using novel schema and distant supervision." In: *Proceedings of the 21st workshop on biomedical language processing*. 2022, pp. 155–160.

[27] Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. "BERN2: an advanced neural biomedical named entity recognition and normalization tool." In: *Bioinformatics* 38.20 (2022), pp. 4837–4839.

[28] Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. "Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison." In: *Briefings in Bioinformatics* 22.6 (2021), bbab282.

[29] Varsha Naik, Purvang Patel, and Rajeswari Kannan. "Legal entity extraction: An experimental study of NER approach for legal documents." In: *International Journal of Advanced Computer Science and Applications* 14.3 (2023).

[30] Fernando Trias, Hongming Wang, Sylvain Jaume, and Stratos Idreos. "Named entity recognition in historic legal text: A transformer and state machine ensemble method." In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. 2021, pp. 172–179.

[31] Ting Wai Terence Au, Ingemar J Cox, and Vasileios Lampos. "E-NER–An Annotated Named Entity Recognition Corpus of Legal Text." In: *arXiv preprint arXiv:2212.09306* (2022).

[32] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. "A survey of transfer learning." In: *Journal of Big data* 3.1 (2016), pp. 1–40.

[33] Masoume Gholizade, Hadi Soltanizadeh, Mohammad Rahmanimanesh, and Shib Sankar Sana. "A review of recent advances and strategies in transfer learning." In: *International Journal of System Assurance Engineering and Management* (2025), pp. 1–40.

[34] Siraj Khan, Pengshuai Yin, Yuxin Guo, Muhammad Asim, and Ahmed A Abd El-Latif. "Heterogeneous transfer learning: recent developments, applications, and challenges." In: *Multimedia Tools and Applications* 83.27 (2024), pp. 69759–69795.

[35] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. "Lost in transduction: Transductive transfer learning in text classification." In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16.1 (2021), pp. 1–21.

[36] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. "Towards a Unified View of Parameter-Efficient Transfer Learning." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021. URL: https://arxiv.org/abs/2110.04366.

[37] N. Ding, Y. Liu, J. Tang, J. Li, and M. Sun. "Parameter-Efficient Fine-Tuning of Large-Scale Pretrained Models." In: *Nature Machine Intelligence* 5 (2023), pp. 220–235. DOI: 10.1038/s42256-023-00626-4.

[38] Tasfia Shermin, Shyh Wei Teng, Manzur Murshed, Guojun Lu, Ferdous Sohel, and Manoranjan Paul. "Enhanced transfer learning with imagenet trained classification layer." In: *Pacific-rim symposium on image and video technology*. Springer. 2019, pp. 142–155.

[39] Lina Chato and Emma Regentova. "Survey of transfer learning approaches in the machine learning of digital health sensing data." In: *Journal of personalized medicine* 13.12 (2023), p. 1703.

[40] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. "Parameter-efficient transfer learning for NLP." In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2790–2799.

[41] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-rank adaptation of large language models." In: *arXiv preprint arXiv:2106.09685* (2021).

[42] Xiang Lisa Li and Percy Liang. "Prefix-tuning: Optimizing continuous prompts for generation." 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)*. 2021, pp. 4582–4597.

[43] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." In: *arXiv preprint arXiv:1503.02531* (2015).

[44] Shuyi Ji, Zizhao Zhang, Shihui Ying, Liejun Wang, Xibin Zhao, and Yue Gao. "Kullback–Leibler Divergence Metric Learning." In: *IEEE Transactions on Cybernetics* 52.4 (2022), pp. 2047–2058. DOI: 10.1109/TCYB.2020.3008248.

[45] Sharu Theresa Jose and Osvaldo Simeone. "Information-Theoretic Bounds on Transfer Generalization Gap Based on Jensen-Shannon Divergence." In: *2021 29th European Signal Processing Conference (EUSIPCO)*. 2021, pp. 1461–1465. DOI: 10.23919/EUSIPCO54536.2021.9616270.

[46] Nguyen Bach and Sameer Badaskar. "A review of relation extraction." In: *Literature review for Language and Statistics II* (2007).

[47] Wenxuan Zhou and Muhao Chen. "An improved baseline for sentence-level relation extraction." In: *arXiv preprint arXiv:2102.01373* (2021).

[48] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. "Document-level relation extraction as semantic segmentation." In: *arXiv preprint arXiv:2106.03618* (2021).

[49] Zheng Chen and Changyu Guo. "A pattern-first pipeline approach for entity and relation extraction." In: *Neurocomputing* 494 (2022), pp. 182–191.

[50] ZHANG Shaowei, WANG Xin, CHEN Zirui, WANG Lin, XU Dawei, and JIA Yongzhe. "Survey of Supervised Joint Entity Relation Extraction Methods." In: *Journal of Frontiers of Computer Science & Technology* 16.4 (2022).

[51] Zhaohui Yan, Zixia Jia, and Kewei Tu. "An empirical study of pipeline vs. joint approaches to entity and relation extraction." In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2022, pp. 437–443.

[52] Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. "A Survey on Open Information Extraction." In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*. 2018, pp. 3866–3878.

[53] Pai Liu, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Li Zongsheng, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. "A Survey on Open Information Extraction from Rule-based Model to Large Language Model." In: *Findings of the Association for Computational Linguistics: EMNLP*. 2024, pp. 9586–9608.

[54] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data." In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 2009, pp. 1003–1011.

[55] Tingsong Jiang, Jing Liu, Chin-Yew Lin, and Zhifang Sui. "Revisiting Distant Supervision for Relation Extraction." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2018, pp. 2348–2357.

[56] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017, pp. 2877–2883.

[57] Kiran Detroja et al. "A Survey on Relation Extraction." In: *Journal / Survey Publication (ScienceDirect)* (2023). available via ScienceDirect.

[58] Abhishek Sharma, Sudeshna Chakraborty, Shivam Kumar, et al. "Named entity recognition in natural language processing: A systematic review." In: *Proceedings of Second Doctoral Symposium on Computational Intelligence*. Springer. 2022, pp. 817–828.

[59] Erik F Tjong Kim Sang and Fien 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-Volume 4*. 2003, pp. 142–147.

[60] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. "Ace 2005 multilingual training corpus." In: *(No Title)* (2006).

[61] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." In: *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001, pp. 282–289.

[62] Andrew McCallum and Wei Li. "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons." In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 188–191.

[63] Kaixin Liu, Qingcheng Hu, Jianwei Liu, and Chunxiao Xing. "Named entity recognition in Chinese electronic medical records based on CRF." In: *2017 14th Web Information Systems and Applications Conference (WISA)*. IEEE. 2017, pp. 105–110.

[64] U Kanimozhi and D Manjula. "A CRF based machine learning approach for biomedical named entity recognition." In: *2017 second international conference on recent trends and challenges in computational models (ICRTCCM)*. IEEE. 2017, pp. 335–342.

[65] Charles Jochim and Lea Deleris. "Named entity recognition in the medical domain with constrained CRF models." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 2017, pp. 839–849.

[66] Burr Settles. "Biomedical named entity recognition using conditional random fields and rich feature sets." In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. 2004, pp. 107–110.

[67] Fei Sha and Fernando Pereira. "Shallow parsing with conditional random fields." In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. 2003, pp. 134–141.

[68] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by Gibbs sampling." In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. 2005, pp. 363–370.

[69] Naeem Ullah, Javed Ali Khan, Mohammad Sohail Khan, Wahab Khan, Izaz Hassan, Marwa Obayya, Noha Negm, and Ahmed S. Salama. "An effective approach to detect and identify brain tumors using transfer learning." In: *Applied Sciences* 12.11 (2022).

[70] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. "A survey on recent advances in named entity recognition." In: *arXiv preprint arXiv:2401.10825* (2024).

[71] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. "Natural language processing (almost) from scratch." In: *Journal of Machine Learning Research*. Vol. 12. 2011, pp. 2493–2537.

[72] Cicero Nogueira dos Santos and Victor Guimarães. "Boosting named entity recognition with neural character embeddings." In: *Proceedings of the Fifth Named Entity Workshop*. 2015, pp. 25–33.

[73] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." In: *arXiv preprint arXiv:1508.01991* (2015).

[74] Alex Graves. "Long short-term memory." In: *Supervised sequence labelling with recurrent neural networks* (2012), pp. 37–45.

[75] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris 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*. 2016, pp. 260–270.

[76] Xuezhe Ma and Eduard Hovy. "End-to-end sequence labeling via bi-directional lstm-cnns-crf." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 1064–1074.

[77] Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. "Document-level chemical named entity recognition with attention-based bidirectional recurrent neural networks." In: *Database* 2018 (2018).

[78] Jason PC Chiu and Eric Nichols. "Named entity recognition with bidirectional LSTM-CNNs." In: *Transactions of the Association for Computational Linguistics*. Vol. 4. 2016, pp. 357–370.

[79] Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. "Transfer learning for sequence tagging with hierarchical recurrent networks." In: *International Conference on Learning Representations*. 2017.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In: *Advances in neural information processing systems* 30 (2017).

[81] Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, and Bryan Catanzaro. "Long-Short Transformer: Efficient Transformers for Language and Vision." In: *NeurIPS*. 2021, –.

[82] Shuohang Wang, Luowei Zhou, Zhe Gan, Yen-Chun Chen, Yuwei Fang, Siqi Sun, Yu Cheng, and Jingjing Liu. "Cluster-Former: Clustering-based Sparse Transformer for Long-Range Dependency Encoding." In: *Findings of ACL/IJCNLP*. 2021, pp. 3958–3968.

[83] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186.

[84] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Ott, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "RoBERTa: A robustly optimized BERT pretraining approach." In: *arXiv preprint arXiv:1907.11692* (2019).

[85] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. "DeBERTa: Decoding-enhanced BERT with disentangled attention." In: *arXiv preprint arXiv:2006.03654* (2020).

[86] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. "ELECTRA: Pre-training text encoders as discriminators rather than generators." In: *International Conference on Learning Representations*. 2020.

[87] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: a pretrained biomedical language representation model for biomedical text mining." In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[88] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A pretrained language model for scientific text." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 3615–3620.

[89] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. "Publicly available clinical BERT embeddings." In: *arXiv preprint arXiv:1904.03323* (2019).

[90] Allen H Huang, Hui Wang, and Yi Yang. "FinBERT: A large language model for extracting information from financial text." In: *Contemporary Accounting Research* 40.2 (2023), pp. 806–841.

[91] Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. "NER-BERT: A pre-trained model for low-resource entity tagging." In: *arXiv preprint arXiv:2112.00405* (2021).

[92] Danish Contractor, Barun Patra, Parag Singla, et al. "Constrained BERT BiLSTM CRF for understanding multi-sentence entity-seeking questions." In: *Natural Language Engineering* 27.1 (2021), pp. 65–87.

[93] Jian Liu, Lei Gao, Sujie Guo, Rui Ding, Xin Huang, Long Ye, Qinghua Meng, Asef Nazari, and Dhananjay Thiruvady. "A hybrid deep-learning approach for complex biochemical named entity recognition." In: *Knowledge-Based Systems* 221 (2021), p. 106958.

[94] Yafei Liu, Siqi Wei, Haijun Huang, Qin Lai, Mengshan Li, and Lixin Guan. "Naming entity recognition of citrus pests and diseases based on the BERT-BiLSTM-CRF model." In: *Expert Systems with Applications* 234 (2023), p. 121103.

[95] Yue Zhang, Qi Chen, Zhenghua Yang, Hongfei Lin, and Zhiyong Lu. "Clinical named entity recognition from Chinese electronic health records via machine reading comprehension." In: *BMC medical informatics and decision making* 19.5 (2019), pp. 1–10.

[96] Liang Xu, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. "BERT-based Chinese named entity recognition with adversarial training." In: *Applied Sciences* 11.5 (2021), p. 2275.

[97] Qi Cheng, Liqiong Chen, Zhixing Hu, Juan Tang, Qiang Xu, and Binbin Ning. "A novel prompting method for few-shot ner via llms." In: *Natural Language Processing Journal* 8 (2024), p. 100099.

[98] Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. "Empirical study of zero-shot ner with chatgpt." In: *arXiv preprint arXiv:2310.10035* (2023).

[99] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[100] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. "Crosslingual generalization through multitask finetuning." In: *arXiv preprint arXiv:2211.01786* (2022).

[101] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. "Llama: Open and efficient foundation language models." In: *arXiv preprint arXiv:2302.13971* (2023).

[102] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. "GPT-NER: Named entity recognition via large language models." In: *arXiv preprint arXiv:2304.10428* (2023).

[103] Dhananjay Ashok and Zachary C Lipton. "PromptNER: Prompting for named entity recognition." In: *arXiv preprint arXiv:2305.15444* (2023).

[104] Francesco De Toni, Christopher Akiki, Javier De La Rosa, Clémentine Fourrier, Enrique Manjavacas, Stefan Schweter, and Daniel Van Strien. "Entities, dates, and languages: Zero-shot on historical texts with To." In: *arXiv preprint arXiv:2204.05211* (2022).

[105] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. "Multitask prompted training enables zero-shot task generalization." In: *arXiv preprint arXiv:2110.08207* (2021).

[106] Xingyu Zhu, Feifei Dai, Xiaoyan Gu, Bo Li, Meiou Zhang, and Weiping Wang. "GL-NER: Generation-aware large language models for few-shot named entity recognition." In: *In International Conference on Artificial Neural Networks* (2024), pp. 433–448.

[107] Yongil Kim, Yerin Hwang, Joongbo Shin, Hyunkyung Bae, and Kyomin Jung. "Injecting comparison skills in task-oriented dialogue systems for database search results disambiguation." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (2023), pp. 4047–4065.

[108] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. "UniversalNER: Targeted distillation from large language models for open named entity recognition." In: *arXiv preprint arXiv:2308.03279* (2023).

[109] Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. "Self-improving for zero-shot named entity recognition with large language models." In: *arXiv preprint arXiv:2311.08921* (2023).

[110] Tristan Luiggi, Tanguy Herserant, Thong Tran, Laure Soulier, and Vincent Guigue. "CALM: Context augmentation with large language model for named entity recognition." In: *In International Conference on Theory and Practice of Digital Libraries* (2024), pp. 273–291.

[111] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. "Domain adaptation of rule-based annotators for named-entity recognition tasks." In: *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010, pp. 1002–1012.

[112] Marco A Valenzuela-Escárcega, Gus Hahn-Powell, Mihai Surdeanu, and Thomas Hicks. "A domain-independent rule-based framework for event extraction." In: *Proceedings of ACL-IJCNLP 2015 system demonstrations*. 2015, pp. 127–132.

[113] Behrang Mohit. "Named entity recognition." In: *Natural language processing of semitic languages*. Springer, 2014, pp. 221–245.

[114] Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. "Deepalignment: Unsupervised ontology matching with refined word vectors." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 787–798.

[115] Debora Nozza, Pikakshi Manchanda, Elisabetta Fersini, Matteo Palmonari, and Enza Messina. "LearningToAdapt with word embeddings: Domain adaptation of named entity recognition systems." In: *Information Processing & Management* 58.3 (2021), p. 102537.

[116] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. "Self-alignment pretraining for biomedical entity representations." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022, pp. 312–323.

[117] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. "Adamatch: A unified approach to semi-supervised learning and domain adaptation." In: *arXiv preprint arXiv:2106.04732* (2021).

[118] Anders Søgaard. *Semi-supervised learning and domain adaptation in natural language processing*. Springer Nature, 2022.

[119] Xiaoyan Zhao, Yang Deng, Min Yang, Linhao Wang, Rui Zhang, Hengzhe Cheng, Wai Lam, Yuji Shen, and Ruifeng Xu. "A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers." In: *ACM Computing Surveys* 56.11 (2024), pp. 1–35. DOI: 10.1145/3674501.

[120] Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. "Relation extraction: A survey." In: *arXiv preprint arXiv:1712.05191* (2017).

[121] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." In: *arXiv preprint arXiv:1508.01991*. 2015.

[122] Jose A. Diaz-Garcia and Julio Amador Diaz Lopez. "A survey on cutting-edge relation extraction techniques based on language models." In: *Artificial Intelligence Review* (2025). DOI: 10.1007/s10462-025-11280-0. URL: https://arxiv.org/abs/2411.18157.

[123] Libin Yang, Dandan Zhao, Jiana Meng, and Shichang Sun. "Joint Entity and Relation Extraction Based on Large Language Model and Graph Convolutional Networks." In: *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)*. IEEE. 2025, pp. 1249–1253.

[124] Haochen Zou, Yongli Wang, and Anqi Huang. "Large language model augmented joint learning framework for entity-relation extraction." In: *Applied Soft Computing* 186 (2026), p. 114094. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2025.114094. URL: https://www.sciencedirect.com/science/article/pii/S1568494625014073.

[125] Dan Roth and Wen tau Yih. "A Linear Programming Formulation for Global Inference in Natural Language Tasks." In: *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004, pp. 1–8. DOI: 10.3115/1220355.1220356.

[126] Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. "TACRED revisited: A thorough evaluation of the TACRED relation extraction task." In: *arXiv preprint arXiv:2004.14855* (2020).

[127] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhiyuan Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. "DocRED: A large-scale document-level relation extraction dataset." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 764–777.

[128] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. "FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 4803–4809.

[129] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. "SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals." In: *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*. Uppsala, Sweden: Association for Computational Linguistics, 2010, 33–38.

[130] Hima Gurulingappa, Angus Roberts, Jun Xu, Yike Guo, and Robin Stevens. "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports." In: *Journal of Biomedical Informatics* 45.5 (2012), 885–892.

[131] Xiaoqian Liu, Wen Li, Jimeng Sun, Zhiheng Li, and Guo-Qiang Zhang. "Knowledge-based extraction of adverse drug events from biomedical text." In: *BMC Bioinformatics* 15.64 (2014).

[132] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[133] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji 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). Association for Computational Linguistics, 2020, pp. 6442–6454. DOI: 10.18653/v1/2020.emnlp-main.523.

[134] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[135] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji 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)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 6442–6454. DOI: 10.18653/v1/2020.emnlp-main.523. URL: https://aclanthology.org/2020.emnlp-main.523/.

[136] Amirhossein Layegh, Amir H. Payberah, Ahmet Soylu, Dumitru Roman, and Mihhail Matskin. "Wiki-based Prompts for Enhancing Relation Extraction using Language Models." In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. SAC '24. Avila, Spain: Association for Computing Machinery, 2024, 731–740. ISBN: 9798400702433. DOI: 10.1145/3605098.3635949. URL: https://doi.org/10.1145/3605098.3635949.

[137] Pere-Lluís Huguet Cabot and Roberto Navigli. "REBEL: Relation Extraction By End-to-end Language generation." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2370–2381. DOI: 10.18653/v1/2021.findings-emnlp.204.

[138] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 7871–7880.

[139] Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. "RED^{FM}: a Filtered and Multilingual Relation Extraction Dataset." In: *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics: ACL*

2023. Toronto, Canada: Association for Computational Linguistics, July 2023. URL: https://arxiv.org/abs/2306.09802.

[140] Pengfei Li, Kezhi Mao, Xuefeng Yang, and Qi Li. "Improving Relation Extraction with Knowledge-attention." 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)*. Association for Computational Linguistics, 2019, 229–239. DOI: 10.18653/v1/d19-1022. URL: http://dx.doi.org/10.18653/v1/D19-1022.

[141] Jianwei Gao, Huaiyu Wan, and Youfang Lin. "Exploiting global context and external knowledge for distantly supervised relation extraction." In: *Knowledge-Based Systems* 261 (2023), p. 110195. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2022.110195. URL: https://www.sciencedirect.com/science/article/pii/S0950705122012916.

[142] Mounir Ourekouch, Mohammed-Amine Koulali, and Mohammed Erradi. "RelCheck: Improving Relation Extraction with Ontology-Guided and LLM-Based Validation." In: *European Semantic Web Conference*. Springer. 2025, pp. 441–459.

[143] Yu Tao, Ruopeng Yang, Yongqi Wen, Yihao Zhong, Kaige Jiao, and Xiaolei Gu. "LLM-KE: An Ontology-Aware LLM Methodology for Military Domain Knowledge Extraction." In: *Computers, Materials and Continua* 86.1 (2025), pp. 1–17. ISSN: 1546-2218. DOI: https://doi.org/10.32604/cmc.2025.068670. URL: https://www.sciencedirect.com/science/article/pii/S1546221825010082.

[144] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Sastry Girish, Amanda Askell, et al. "Language models are few-shot learners." In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[145] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." In: *IEEE Transactions on Knowledge and Data Engineering* (2024). arXiv preprint arXiv:2306.08302.

[146] Somin Wadhwa, Silvio Amir, and Byron C Wallace. "Revisiting relation extraction in the era of large language models." In: *Proceedings of the conference. association for computational linguistics. meeting*. Vol. 2023. 2023, p. 15566.

[147] Tian Zhang, Lianbo Ma, Shi Cheng, Yikai Liu, Nan Li, and Hongjiang Wang. "Automatic prompt design via particle swarm optimization driven LLM for efficient medical information extraction." In: *Swarm and Evolutionary Computation* 95 (2025), p. 101922. ISSN: 2210-6502. DOI: https://doi.org/10.1016/j.swevo.2025.101922. URL: https://www.sciencedirect.com/science/article/pii/S221065022500080X.

[148] Xingpeng Si, Yuhao Ye, Yu-Ming Shang, Mengyuan Cao, Yu Bai, Jiawei Li, and Yang Gao. "Fusing generation task data to enhance generalization of LLMs in zero-shot relational triplet extraction." In: *Expert Systems with Applications* 300 (2026), p. 130411. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2025.130411. URL: https://www.sciencedirect.com/science/article/pii/S0957417425040266.

[149] Xiang Dai and Heike Adel. "An Analysis of Simple Data Augmentation for Named Entity Recognition." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3861–3867. DOI: 10.18653/v1/2020.coling-main.343. URL: https://aclanthology.org/2020.coling-main.343/.

[150] KennethEnevoldsen. *Augmenty*. https://github.com/KennethEnevoldsen/augmenty. 2023.

[151] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal Loss for Dense Object Detection." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 318–327. DOI: 10.1109/TPAMI.2018.2858826.

[152] Xiang Dai and Heike Adel. "An Analysis of Simple Data Augmentation for Named Entity Recognition." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3861–3867. DOI: 10.18653/v1/2020.coling-main.343. URL: https://aclanthology.org/2020.coling-main.343.

[153] Matthew Honnibal and Ines Montani. *SpanCat: Span categorization for overlapping named entity recognition*. Tech. rep. Technical Documentation. Explosion AI, 2022. URL: https://spacy.io/usage/spancat.

[154] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. *spaCy: Industrial-strength Natural Language Processing in Python*. Version 3.4. 2023. URL: https://spacy.io.

[155] Ines Montani, Matthew Honnibal, and Sofie Van Landeghem. "Span-based named entity recognition with overlapping entities." In: *Proceedings of the 2022 Workshop on Machine Learning for Natural Language Processing*. Online: Association for Computational Linguistics, 2022, pp. 156–167.

[156] Debora Nozza, Pikakshi Manchanda, Elisabetta Fersini, Matteo Palmonari, and Enza Messina. "LearningToAdapt with word embeddings: Domain adaptation of Named Entity Recognition systems." In: *Information Processing & Management* 58.3 (2021), p. 102537. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102537. URL: https://www.sciencedirect.com/science/article/pii/S0306457321000455.

[157] Elisabetta Fersini, Enza Messina, and Federico Alberto Pozzi. "Sentiment analysis: Bayesian Ensemble Learning." In: *Decision Support Systems* 68 (2014), pp. 26–38. ISSN: 0167-9236.

[158] Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. *NER-BERT: A Pre-trained Model for Low-Resource Entity Tagging*. 2021. arXiv: 2112.00405 [cs.CL]. URL: https://arxiv.org/abs/2112.00405.

[159] Erik F Tjong Kim Sang and Fien De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." In: *Proc. of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 142–147.

[160] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. *A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers*. 2024. arXiv: 2306.02051 [cs.CL]. URL: https://arxiv.org/abs/2306.02051.

[161] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. "A Comprehensive Survey on Knowledge Graphs: Representation, Construction, and Applications." In: *IEEE Transactions on Neural Networks and Learning Systems* (2022). DOI: 10.1109/TNNLS.2021.3070843.

[162] Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, and Maosong Sun. "OpenKE: An Open Toolkit for Knowledge Embedding." In: *Proceedings of EMNLP: System Demonstrations*. 2018. URL: https://aclanthology.org/D18-2024/.

[163] Sebastian Riedel, Limin Yao, and Andrew McCallum. "Modeling relations and their mentions without labeled text." In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2010, pp. 148–163.

[164] Kristina Toutanova and Danqi Chen. "Representing Text for Joint Embedding of Text and Knowledge Bases." In: *Proceedings of EMNLP*. 2015. URL: https://aclanthology.org/D15-1036/.

[165] Baoxu Shi and Tim Weninger. "Pattern-Enhanced Embedding Model for Knowledge Graph Completion." In: *Proceedings of WWW*. 2019. DOI: 10.1145/3308558.3313565.

[166] Yankai Lin, Zhiyuan Liu, Yaojie Shen, Xuan Liu, Huanbo Luan, and Maosong Sun. "KAGNET: Knowledge-Aware Graph Networks for Commonsense Reasoning." In: *Proceedings of EMNLP-IJCNLP*. 2019. URL: https://aclanthology.org/D19-1365/.

[167] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. "KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation." In: *Transactions of the Association for Computational Linguistics* 9 (2021). URL: https://aclanthology.org/2021.tacl-1.11/.

[168] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. "K-BERT: Enabling Language Representation with Knowledge Graph." In: *Proceedings of AAAI*. 2022. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6233.

[169] Dan Roth and Wen-tau Yih. "A Linear Programming Formulation for Global Inference in Natural Language Tasks." In: *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004, pp. 1–8. URL: https://aclanthology.org/W04-2401.

[170] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. "SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs

of Nominals." In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Ed. by Katrin Erk and Carlo Strapparava. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 33–38. URL: https://aclanthology.org/S10-1006/.

[171] Makoto Miwa and Mohit Bansal. "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1105–1116. DOI: 10.18653/v1/P16-1105.

[172] Marah Abdin et al. *Phi-4 Technical Report*. 2024. arXiv: 2412.08905 [cs.CL]. URL: https://arxiv.org/abs/2412.08905.

[173] Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. *X-ALMA: Plug Play Modules and Adaptive Rejection for Quality Translation at Scale*. 2025. arXiv: 2410.03115 [cs.CL]. URL: https://arxiv.org/abs/2410.03115.

[174] OpenAI Team. *GPT-4o mini: advancing cost-efficient intelligence*. https://openai.com/gpt4o-mini. Read me. Accessed on 23 Aug. 2024. July 2024.

[175] lxyuan. *span-marker-bert-base-multilingual-cased-multinerd*. https://huggingface.co/lxyuan/span-marker-bert-base-multilingual-cased-multinerd. Fine-tuned SpanMarker model based on bert-base-multilingual-cased for multilingual named entity recognition on MultiNERD dataset. 2023.

[176] osiria. *bert-italian-cased-ner*. https://huggingface.co/osiria/bert-italian-cased-ner. BERT-based model for Italian Named Entity Recognition, fine-tuned on WikiNER dataset for Person, Location, Organization and Miscellaneous entity classes. 2023.

[177] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. "GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer." In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 5364–5376. DOI: 10.18653/v1/2024.naacl-

long.300. URL: https://aclanthology.org/2024.naacl-long.300/.

[178] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. *PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods*. https://github.com/huggingface/peft. 2022.

[179] Pere-Lluís Huguet Cabot and Roberto Navigli. "REBEL: Relation Extraction By End-to-end Language generation." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2370–2381. DOI: 10.18653/v1/2021.findings-emnlp.204. URL: https://aclanthology.org/2021.findings-emnlp.204/.

[180] AI@Meta. "Llama 3 Model Card." In: (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[181] Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. *Advanced Natural-based interaction for the ITAlian language: LLaMAntino-3-ANITA*. 2024. arXiv: 2405.07101 [cs.CL].

[182] Roberto Navigli and Sapienza NLP group. "Minerva: Italy's First Family of Large Language Models trained on Italian texts." In: (2024).

[183] Almawave. *Velvet AI: sustainable and high-performance Italian multilingual LLM*. Wikipedia. 2025.