Università
degli Studi
di Padova

Dipartimento
di Ingegneria
dell'Informazione

Master Thesis in ICT FOR INTERNET AND MULTIMEDIA

Applying Natural Language Processing Techniques for Sentence-Level Relation Extraction: Analysis and Performance Evaluation

Master Candidate

Merve Tuncer

Student ID 2040508

Supervisor

Prof. Giorgo Satta

University of Padova

Co-supervisor

Giovanni Angelini

expert.ai

Academic Year
2022/2023

I dedicated this to my lovely mum, dad
to my supportive, sisters
to my love, Davide
and to my caring friends, Sevval, Esther, Gulus

Abstract

This study explores the field of sentence-level relation extraction in the context of natural language processing (NLP) applications. We have analyzed many approaches, including document-level relation extraction, in the goal of creating a reliable model for this purpose. This study clarified the difficulties associated with entity coreference resolution as well as the subtle capture of global context in large textual sources. We also assessed the effectiveness of current sentence-level relation extraction methods. The TACRED dataset provided the main source of information for our research, which also made use of the BERT (Bidirectional Encoder Representations from Transformers) model's impressive capabilities.

The goal was to carefully examine how the Long Short-Term Memory (LSTM) and BERT model performed on the TACRED dataset and assess its precision in extracting relationships between entities embedded within sentences. This project provided insightful information on the relative performance of the LSTM and BERT models in the context of sentence-level relation extraction, which helped to clarify the relative advantages and disadvantages of each model.

In order to gain a more comprehensive understanding of the state of the art in this subject, our research also examined the content of literature and research papers addressing sentence and document-level connection extraction strategies. These sources expanded the depth and scope of our research by providing methodology insights and serving as benchmarks for comparison with our own findings.

Contents

List of Figures

List of Tables

List of Acronyms

NLP Natural Language Processing

LSTM Long Short-Term Memory

BERT Bidirectional Encoder Representation from Transformers

DocRED Document-Level Relation Extraction Dataset

DocRE Document-Level Relatiov Extraction

NER Named Entity Recognition

ACE Automatic Content Extraction

EDR Entity Detection and Recognition

EMD Entity Mention Detection

RDR Relation Detection and Recognition

RDM Relation Mention Detection

SOTA State of the Art

RE Relation Extraction

LM Language Model

LMs Language Models

CRF Conditional Random Field

CHE Candidate Head Entity

MTE Multiple Tail Entities

ERNIE Enhanced Language Representation with Informative Entities

CNN Convolutional Neural Network

CNNs Convolutional Neural Networks

AFL Adaptive Focal Loss

ATL Adaptive Thresholding Loss

LDC Linguistic Data Consortium

TP True Positive

FP False Positive

FN False Negative

RNN Recurrent Neural Network

RNNs Recurrent Neural Networks

FFNN Feed-Forward Neural Network

1 INTRODUCTION

Relation extraction is an important process in the field of natural language processing (NLP), which is crucial for obtaining insightful information from unstructured textual input. Relation extraction's fundamental task is to locate and extract significant semantic connections or relationships between entities mentioned in a given text. These entities can include organizaitons, groups, places, and more. Relation extraction is important on several levels. In the beginning, it makes it easier to structure and arrange unstructured text, transforming it into a more arranged and semantically rich representation. A variety of downstream NLP applications, such as knowledge graph generation, information retrieval, question-answering, and summarization, are built on top of this structured data. It facilitates the building of knowledge graphs that model intricate links in difficult domains, assisting in improved comprehension and analysis, by automating the extraction of relationships.

Additionally, relation extraction is an effective technique for gaining insights from consumer comments, product evaluations, and market trends in the business and industrial areas as well. Businesses may use it to understand client feelings, find links between features and attributes of products, and make well-informed judgments about how to enhance products and implement marketing plans. Relation extraction is a tool used in the healthcare industry to help researchers uncover possible drug-disease links, find novel therapies, and advance medical knowledge by extracting key ideas from biological literature. Relation extraction fundamentally exposes the ability to extract important knowledge and insights from enormous volumes of unstructured text, making it a crucial part of the developing NLP environment. Its uses are broad, advancing decision-making and knowledge representation, and eventually enhancing our comprehension of the world through the use of textual data.

In order to fulfill the various requirements of NLP applications, there are two types of relation extraction methods: document-level and sentence-level. The complexity of the relationships in the text and the specific objectives of the study will determine which technique is optimal. While document-level extraction is crucial when dealing with sophisticated, long-distance, or complex linkages, sentence-level extraction is appropriate for localized and simple interactions. Both strategies are important techniques in the NLP toolbox that help ensure accurate information extraction from texts of various lengths and complexity.

The length, complexity, and applicability of relation extraction at the sentence and document levels are different. Sentence-level extraction is ideal for simpler interactions in brief contexts since it is granular and concentrates on links within individual sentences. In contrast, document-level extraction offers a wider perspective of complicated and overarching correlations by taking links spanning numerous phrases or the full documents into consideration. Sentence-level extraction is better suitable for delicate connections since it is less complicated and acts just within a single sentence. Conversely, document-level extraction is more complicated and involves co-reference resolution and discourse analysis across phrases, allowing for the extraction of relationships that are more complex and subtle. While document-level extraction is essential for building thorough knowledge graphs, locating long-distance linkages, and summarizing large documents, sentence-level extraction is frequently used for entity recognition, sentiment analysis, and extracting relationships in brief texts. These contrasts emphasize how crucial it is to select the proper degree of extraction based on the particular needs and goals of the NLP activity or application.

2

RESEARCH BACKGROUND

2.1 Sentence-Level Relation Extraction

Sentence-level relation extraction is a task in natural language processing that involves identifying and categorising the connections between diverse entities or elements in a sentence. These relationships could involve a range of interconnections, including links between subjects and verbs, semantic roles, and temporal dependencies. Sentence-level relation extraction is essential for various NLP applications, including information retrieval, knowledge graph development, question answering, and text summarisation and plays a significant role in identifying and extracting relationships between entities from text. One of the main difficulties encountered in the task of sentence-level connection extraction is the frequent presence of ambiguity inside sentences. Sentences often contain numerous entities, and these entities might engage in various interactions. The task of distinguishing amongst potential relationships becomes somewhat challenging when several entities are present. The proper extraction of connections relies on the critical task of determining the things that are related by a certain relation. Another problem occurs due to the intricate nature of contextual relationships. The interpretation of a connection might vary significantly based on the specific context in which it is considered. The appropriate extraction of a relation may need considering the surrounding context, since various statements may demand different interpretations of the same relation. The context in question has the potential to include information from preceding and succeeding phrases within a given document, so adding complexity to the work at hand. The identification and differentiation of complex relationships provide inherent challenges. Certain relationships may have apparent similarities, yet display substantial semantic variations. The ability to differentiate between these intricate relationships frequently requires a significant degree of accuracy and comprehension of sensitive aspects within the language. The effective extraction of relations relies on the precondition of accurate named entity recognition (NER). Mistakes in NER have the potential to affect the accuracy of connection extraction, resulting in erroneous outcomes. Moreover, the process of generating extensive labeled datasets for sentence-level connection extraction might be demanding in terms of resources, potentially leading to a scarcity of data, particularly in some areas or languages. The presence of an ample amount of data is crucial for the process of training models and achieving generalizability. Certain relationships exhibit a notable disparity in occurrence rates across textual data derived from real-world sources, resulting in imbalances within the datasets. The presence of this mismatch has the potential to introduce bias in models, favoring overrepresented relations and leading to suboptimal performance when it comes to minority class relations. In addition, the precise resolution of anaphora and co-reference inside sentences is crucial for the identification of relationships. The failure to address these dependencies has the potential to result in confusion. Real-world textual content frequently exhibits noise, including inaccuracies, casual vocabulary, colloquialisms, and grammatically incorrect constructions. The management of noisy data poses a persistent difficulty in the field of relation extraction. In summary, the resolution of these issues pertaining to sentence-level association extraction necessitates the utilization of sophisticated models, meticulous data preparation techniques, and domain-specific expertise.

2.2 Sentence-Level Relation Extraction Datasets

Datasets for sentence-level relation extraction play a crucial role in the advancement of the area of NLP and information extraction. The utilization of these datasets is of utmost importance in the training and assessment of machine learning models that are specifically developed to extract connections between entities referenced in textual data. This, in turn, facilitates the implementation of many applications, including the development of knowledge graphs, question-answering systems, and information retrieval. This part of article aims to examine the importance of datasets for extracting sentence-level relations, their distinctive features, and significant instances. Sentence-level relation extraction refers to the process of detecting and categorizing the relationships that exist between entities included inside a single sentence. These relationships encompass a range of sorts, such as binary linkages like "is-a" or "part-of," as well as more intricate associations like "works-for" or "authored." The precise extraction of relationships is of utmost importance in comprehending the semantic aspects of textual information and constructing organized representations of knowledge. The presence and accessibility of datasets that focus on extracting relations at the sentence level are crucial for the purpose of training and evaluating machine learning models, particularly those based on deep learning, that can efficiently carry out this task. The datasets mentioned are of great importance to both academic and industrial sectors, as they provide researchers with a valuable tool to create and evaluate models that possess the ability to comprehend the intricacies of language and its surrounding context. Table 2.1 shows the sentence-level relation extraction datasets details.

Datasets	**Relation Types**	**Dataset Access**
TACRED	41 relation types	$25 fee
Re-TACRED	40 relation types	Free
SemEval-2010 Task 8	9 relation types	Free
FewRel	100 relation types	CC BY-SA 4.0 license

Table 2.1: Sentence-level datasets details

2.2.1 TACRED

The TACRED dataset provides an important number of 106,264 instances specifically designed for relation extraction. These instances were collected from a combination of newswire and online text sources, originating from the corpus utilized in the annual TAC Knowledge Base Population (TAC KBP) challenges. The examples provided in TACRED contain a total of 41 relation types, which are utilized in the TAC KBP challenges. These relation types include per:schools attended and org:members, among others. In cases where no specified connection exists, the examples are designated as no relation. The previous instances have been generated by the combination of human annotations obtained from the TAC KBP tasks, as well as through the utilization of crowdsourcing [36].

There are four clear goals that are pursued in the TACRED dataset. Firstly, the dataset aims to be large-scale, meaning that it encompasses a substantial amount of data. Secondly, the dataset is designed to be representative of real-world scenarios, ensuring that it captures the complexities and nuances present in actual situations. Thirdly, TACRED includes negative examples, which are instances where a relation does not exist between entities, in order to provide a comprehensive and balanced training set. Lastly, the dataset is fully supervised, meaning that it is annotated with ground truth labels for each relation instance, enabling the development and evaluation of supervised learning models. Table 2.2 shows TACRED dataset train, dev, test details.

Split	**Examples count**
Train	68,124
Dev	22,631
Test	15,509
Total	106,264

Table 2.2: TACRED dataset details

The TACRED dataset is an essential resource for conducting research on connection extraction. It encompasses a wide range of relations, including the ’no relation’ class that denotes sentences lacking any explicit link. Through a thorough examination, it has been noticed that a significant proportion of the dataset, specifically around 79.5%, comprises phrases that have been categorized as ’no relation.’ Table 2.3 shows negative and positive ratio from dataset. The category labeled as ’no relation’ serves a crucial function in the training and assessment of relation extraction models, as it signifies the lack of any predetermined connections between items. In contrast, the remaining 20.5% of the dataset consists of diverse relations, each denoting distinct associations between things. The aforementioned relations cover a diverse array of semantic linkages, spanning from the concept of ’parent’ to that of ’founder.’ This dataset presents a complex yet valuable resource for the development and evaluation of relation extraction algorithms. A comprehensive comprehension of the distribution of the ’no relation’ label and other relations present in the TACRED dataset holds significant importance for researchers and practitioners in the field. This understanding plays a crucial role in informing the development of models, devising evaluation strategies, and shedding light on the imbalanced nature of the data. Consequently, it guides endeavors aimed at addressing this imbalance and enhancing the accuracy and performance of relation extraction systems.

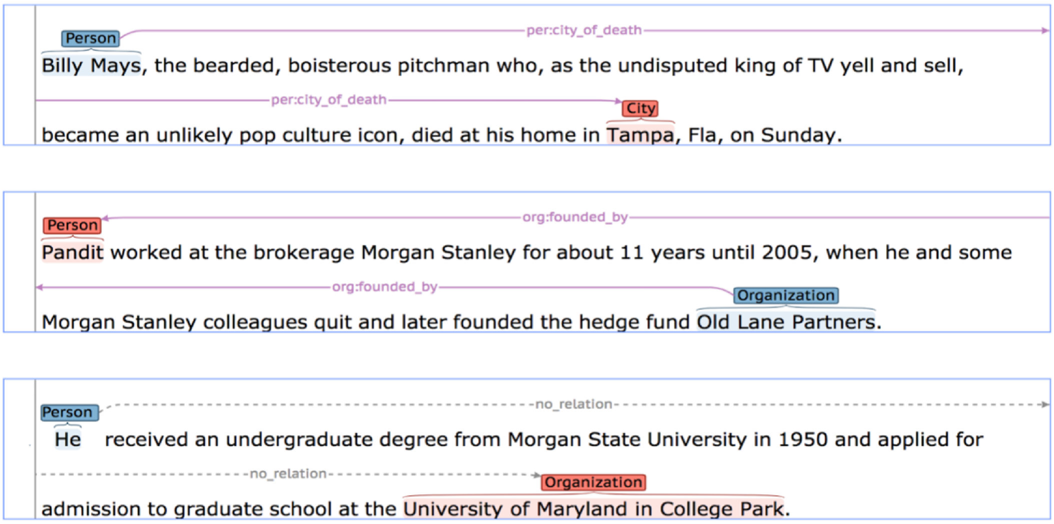
Label	Ratio (%)
Negatif	79.5
Positive	20.5

Table 2.3: Negative and positive percent in data

In the field of relation extraction, the TACRED dataset presents a significant division in its methodology by giving precedence to the identification of the presence or absence of a relation between two entities stated in a phrase as the first stage in the extraction procedure. The main aim in each phrase is to determine the presence of a relationship, indicated by the 'no relation' label, or to identify a certain type of relation. The use of a binary perspective in TACRED distinguishes it from other works by exploring the complexities of semantic linkages. Therefore, it encompasses the fundamental concept of relation extraction, where words convey information that includes entities that may or may not have predetermined connections. The central emphasis on the presence or absence of relationships provides a framework for conducting a comprehensive analysis of the interactions between entities and establishes the groundwork for subsequently categorizing particular types of relationships. The implementation of a multi-tiered approach guarantees that relation extraction models constructed using the TACRED dataset are capable of addressing both the identification of well-established relationships and the more intricate task of determining instances where no such relationship exists. This is a fundamental undertaking in the fields of natural language understanding and knowledge representation.

The TACRED dataset was constructed by extracting phrases containing mention pairs from the TAC KBP newswire and web forum corpus. Figure 2.1 shows the sentences example from dataset. Each example in TACRED is accompanied by annotations that include the spans of the subject and object mentions, the types of the mentions (drawn from the 23 fine-grained types used in the Stanford NER system [19]), and the relation between the entities (selected from the 41 TAC KBP canonical relation types). If no relation is found, a "no relation" label is assigned.

In order to minimize the potential bias of TACRED models towards generating false positive predictions on real-world text, the dataset has been carefully annotated to include negative instances. These negative examples consist of selected phrases in which no link was identified between the indicated pairings.



Billy Mays, the bearded, boisterous pitchman who, as the undisputed king of TV yell and sell, became an unlikely pop culture icon, died at his home in Tampa, Fla, on Sunday.

Pandit worked at the brokerage Morgan Stanley for about 11 years until 2005, when he and some Morgan Stanley colleagues quit and later founded the hedge fund Old Lane Partners.

He received an undergraduate degree from Morgan State University in 1950 and applied for admission to graduate school at the University of Maryland in College Park.

Figure 2.1: TACRED dataset samples

Consequently, a significant majority of the cases, specifically 79.5%, are categorized and tagged as "no relation." Among the instances in which a correlation was identified, the distribution of said correlations is as follows from Figure 2.2.

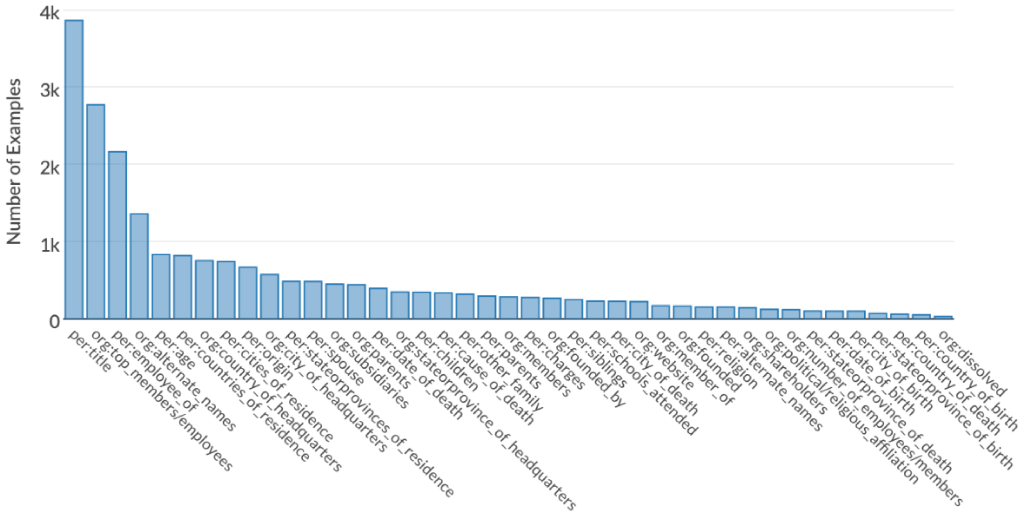


Figure 2.2: Relation Distribution

2.2.2 RE-TACRED

The Re-TACRED dataset is a notable enhancement of the TACRED dataset, specifically designed for the purpose of relation extraction. By using newly obtained crowd-sourced labels, the Re-TACRED framework effectively eliminates inadequately annotated phrases and resolves the issue of ambiguous connection definitions in the TACRED dataset. As a result, it successfully rectifies 23.9% of the erroneous labels included in TACRED. The dataset has a total of 91,000 sentences, which are distributed over 40 distinct associations. The suggested alternate rendering of the TACRED dataset, known as ReTACRED, stands out due to its extensive efforts to address the inherent constraints present in the original dataset. The complex procedure entails a comprehensive restructuring of the training, development, and test sets, along with a deliberate reassessment of certain connection types. The dataset was presented at the AAAI 2021 conference [24].

2.2.3 SemEval-2010 Task 8

SemEval, formerly referred to as semantic evaluation, is a sequence of global contests in the domain of NLP. The primary objective of these competitions is to promote and facilitate research and development in diverse NLP tasks, hence contributing to the advancement of the discipline. SemEval-2010 challenge 8 was a component of the SemEval-2010 competition, specifically designed to address the challenge of multi-way categorization of semantic links between pairs of nominals, which are nouns. The objective of this job was to tackle the difficulty of identifying the semantic connections between nouns in written language. This is a crucial aspect for a range of NLP applications, such as extracting information and constructing knowledge bases. The dataset utilized in the SemEval-2010 Task 8 refers to a multi-way classification task involving the identification of mutually exclusive semantic links between pairs of nominals [8].

SemEval-2010 Task 8 made significant contributions to the progress of scholarly investigations in the fields of relation categorization and information extraction. The dataset offered by this study served as a standardized reference point for researchers, enabling them to compare and evaluate various methodologies for the given goal within a shared framework. The dataset facilitated cooperation within the NLP research community and resulted in the advancement of techniques for extracting semantic relationships between nominals in textual data.

In summary, the participation in SemEval-2010 Task 8 played a crucial role in the progression of the natural language processing area. This task specifically focused on the categorization of semantic relations between pairs of nominals. The provision of a useful dataset and an assessment platform to researchers has eventually made a significant contribution towards the advancement of more precise and contextually-aware natural language processing models.

2.2.4 FewRel

The FewRel dataset [6], also known as the Few-Shot Relation Classification Dataset, comprises a collection of 100 distinct relations and around 70,000 instances sourced from Wikipedia. The FewRel dataset, also known as the Few-shot Relation Classification Dataset, is a commonly employed benchmark dataset in the field of NLP for the specific task of few-shot relation classification. The purpose of its introduction was to evaluate the capacity of NLP models in discerning and categorizing connections existing between entities inside textual data. The FewRel dataset has been specifically developed to assess the effectiveness of NLP models when confronted with limited training instances for relation categorization. This simulation presents a complex scenario when models are required to estimate from a constrained dataset. The FewRel dataset has a diverse range of relation categories, including but not limited to "author," "founder," "capital," "place of birth," and several others. The aforementioned relationships exhibit a wide range of variations and are representative of actual situations seen in the real world. The FewRel dataset is designed to operate in a few-shot learning context, wherein a limited number of labeled samples are available for each relation. This might occur in a limited number of situations, often ranging from one to two occurrences per relationship. Every entry in the dataset comprises of a phrase or paragraph that includes two mentions of entities and a corresponding connection. The objective is to categorize the relationship between the entities. The FewRel dataset is a highly significant resource that serves the purpose of testing and enhancing the skills of NLP models in the domain of few-shot relation classification. This particular task involves properly classifying relations between entities, despite the presence of severely restricted training data. The benchmark serves as a means to evaluate the capacity of NLP systems to generalize and adapt in low-resource circumstances.

2.3 Sentence-Level Relation Extraction Related Work

In this section, we will focus on state-of-the-art (SOTA) studies pertaining to sentence-level analysis. We will go into the methodologies employed in these papers and provide an overview of their respective findings and outcomes.

2.3.1 KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction

Chen et al. [2] are utilizing five datasets in their study, namely SemEval 2010 Task 8 (SemEval), DialogRE, TACRED, TACRED-Revisit, and Re-TACRED. The RoBERTA-large model is being employed for the purpose of fine-tuning. The initial phase involves the injection of knowledge into prompts that may be learned, followed by the proposal of a unique technique called Knowledge-aware Prompt-tuning with synergistic optimization (KnowPrompt) for the task of relation extraction (RE). The research demonstrates intriguing advancements in the field, particularly in terms of novel methodologies. The F1-scores achieved in several benchmark datasets are as follows: SemEval 2010 Task 8 (SemEval) with a score of 90.2, DialogRE with a score of 68.6, TACRED with a score of 72.4, TACRED-Revisit with a score of 82.4, and Re-TACRED with a score of 91.3. The researchers employ a technique including the incorporation of learnable virtual response words and virtual type words into the quick construction process in order to mitigate the labor-intensive nature of prompt engineering. Figure 2.3 shows the KnowPrompt model approach. In order to provide more clarity, rather of utilizing a conventional verbalizer that maps a single label word in the lexicon to a certain class, the authors suggest a novel approach that involves including learnable virtual response words. This is achieved by injecting semantic information, even if it is latent, in order to convey related labels. In addition, the researchers allocate virtual type words that may be learned to represent things in order to serve as weaker Type Markers. These virtual type words are initialized using previous information that is stored in connection labels. Significantly, they employ a novel approach by leveraging learnable virtual type words to adapt dynamically based on context, instead of relying on entity type annotation, which may be absent in datasets. The virtual words, which are created using previous knowledge and relation labels, have the ability to first identify various entity types. Through contextual optimization, these virtual words may effectively convey semantic information that closely aligns with the real entity type. In this way, they serve a role similar to that of a Typer Marker.

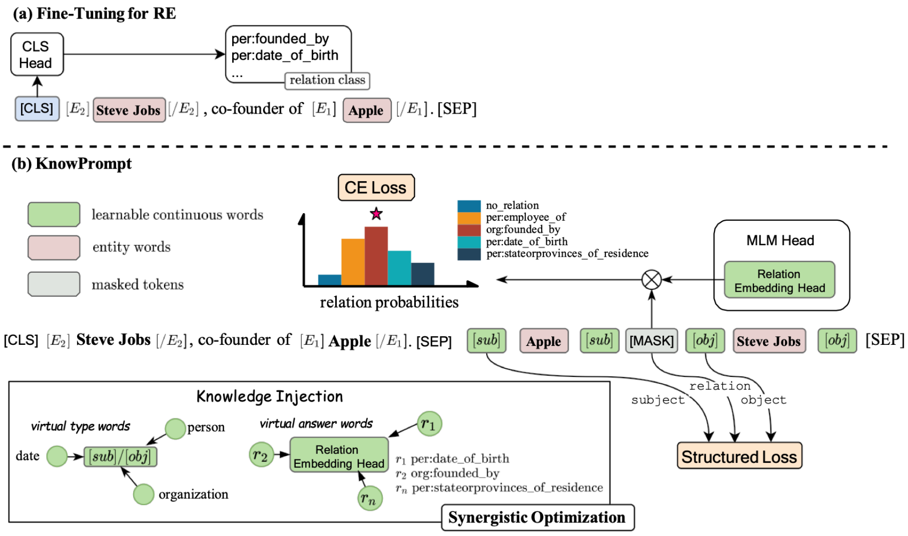


Figure 2.3: KnowPrompt model approach [2]

2.3.2 Relation Classification with Entity Type Restriction

Lyu et. al [16]'s research aim to utilize entity types as a means to limit possible relationships. Subsequently, the system acquires knowledge by constructing a distinct classifier for every combination of entity categories. The researchers employed SpanBert and GCN models in their methodologies. The research achieved F1-score of 75.2 is attained on the TACRED dataset. By imposing an entity type restriction, certain unsuitable relations are eliminated from the pool of potential relations for a given pair of entity types. For instance, when the entities under consideration are people, the model may take into account several connection categories such as "family," "colleagues," or "friends." The objective of limiting the entity types is to enhance the precision and significance of relation categorization, acknowledging that some associations may have a higher likelihood or significance based on the entities' types. For each combination of entity types, a distinct classifier is trained to identify a specific set of potential relations. Figure 2.4 shows the utilization of distinct classifiers for each entity pairs is seen, rather than implementing a generic classifier.

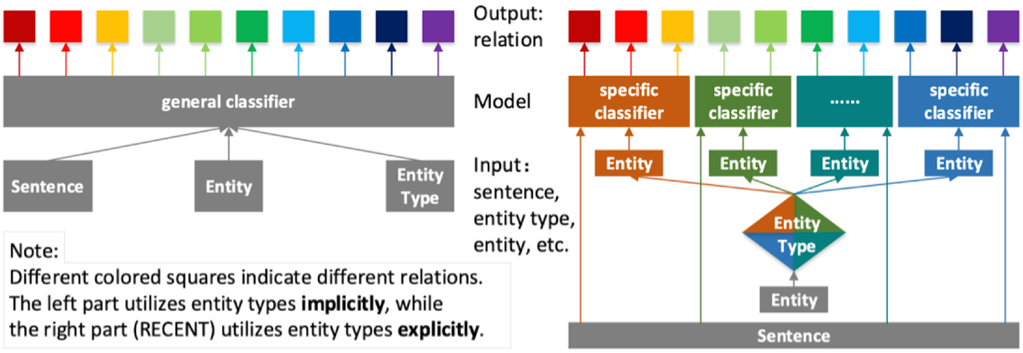


Figure 2.4: Relation Classification with Entity Type Restriction [16]

2.3.3 DEEPSTRUCT: Pretraining of Language Models for Structure Prediction

Wang et. al [29] provides utility in the context of relation classification, specifically in respect to the employment of TACRED and FewRel 1.0 datasets for relation classification tasks. Language Models (LMs) enhance their ability to comprehend structure. The model is trained using a set of task-agnostic corpora, which includes pre-existing large-scale alignments between text and triples. The following tools were employed: T-REx, TEKGEN, KELM, WebNLG, and Concept-Net. The dataset was utilized for challenges involving the prediction of entities and relations. OPIEC was employed for the triple prediction tasks. The authors suggest a method called structure pretraining, which involves pretraining LMs to comprehend textual structures. The researchers employed zero-shot and multi-task learning techniques in their methodology. The researchers conducted a comparative analysis and found that multi-tasking learning yielded more favorable outcomes. The TACRED dataset achieved a F1-score of 76.8, while the FewRel dataset achieved a F1-score of 100. Figure 2.5 shows DEEPSTRUCT pre-training structure. The goal of their method is to improve the structural understanding capabilities of LMs, i.e., understanding the structures of text. Instead of applying the traditional pretrain-finetune approach for individual tasks, the authors propose the adoption of structural pretraining, which seeks to instruct LMs to align with various task structures concurrently.

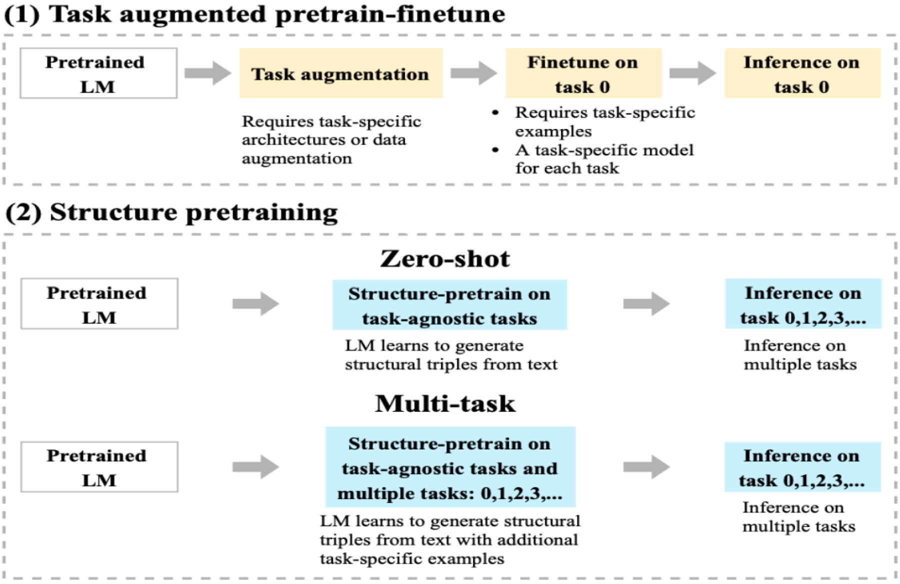

(1) Task augmented pretrain-finetune
Pretrained LM
Task augmentation
Requires task-specific architectures or data augmentation
Finetune on task 0
Requires task-specific examples
A task-specific model for each task
Inference on task 0
(2) Structure pretraining
Zero-shot
Pretrained LM
Structure-pretrain on task-agnostic tasks
LM learns to generate structural triples from text
Inference on task 0,1,2,3,...
Inference on multiple tasks
Multi-task
Pretrained LM
Structure-pretrain on task-agnostic tasks and multiple tasks: 0,1,2,3,...
LM learns to generate structural triples from text with additional task-specific examples
Inference on task 0,1,2,3,...
Inference on multiple tasks

Figure 2.5: DEEPSTRUCT pre-training structure
[29]

2.3.4 Joint Extraction of Entities and Relations via an Entity Correlated Attention Neural Model

Li et. al [12] introduces a two-stage tagging approach that distinguishes between potential head entities and many tail entities in certain relationships. Additionally, it suggests a joint extraction neural model that is based on the entity-first labeling strategy. The researchers used the CoNLL04 and ADE datasets, as well as a specialized Chinese dataset, for their research. The researchers employed a model that utilized a BiLSTM-based encoder module. According to their statement, the researchers reported superior outcomes when employing joint models as opposed to conventional models. The CoNLL04 dataset achieved a F1-score of 77.55, whereas the ADE dataset achieved a F1-score of 79.62. The model has three main components: an encoder module, a candidate head entity (CHE) recognition module, and a multiple tail entities (MTE) recognition module. Figure 2.6 shows their architecture. The encoder module employs a BiLSTM neural network to extract bidirectional sequence characteristics and provide a shared context representation, using the embedded vectors as its input. Subsequently, following the integration of the hidden state and global context characteristics derived from the encoder module, the CHE recognition module employs the BiLSTM-Conditional Random Field (CRF) approach to detect potential head entities. The MTE module receives the candidate head entities from the CHE module and the shared context representation from the encoder module as combined input. The utilization of an entity correlated attention unit is extended to compute the entity correlation inside a particular relation context. This, in conjunction with the BiLSTM-CRF model, facilitates the prediction of ultimate relation tags associated with the tail entities.

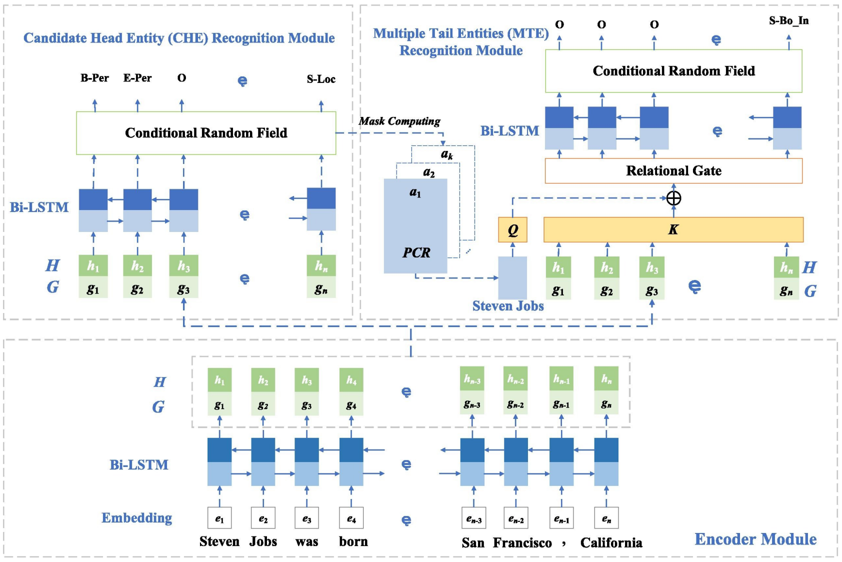


Figure 2.6: Joint extraction of entities and relations via an entity correlated attention neural model architecture [12]

2.3.5 Other Related Works

Zhang et. al [37] employ both extensive textual collections and knowledge graphs (KGs) in order to train an improved language representation model known as enhanced language representation with informative entities (ERNIE). This model is designed to effectively leverage lexical, syntactic, and knowledge-based information concurrently. ERNIE demonstrates notable advancements in a range of knowledge-driven tasks, while also exhibiting comparable performance to the state-of-the-art model BERT in other conventional NLP tasks. The datasets utilized in this study are TACRED, and FewRel. The TACRED dataset achieved a F1-score of 67.87, while the FewRel dataset achieved a F1-score of 88.32. The majority of existing supervised algorithms for relation classification employ a singular embedding to depict the relationship between a given pair of items. Cohen et. al [3] contention is that a more effective strategy is to consider the work of relation categorization as a Span-Prediction issue, akin to the method used in Question Answering. In their study they provide a system that utilizes span prediction for the task of relation classification and proceed to assess its performance in comparison to an existing embedding-based system. The results of the study suggest that the supervised span prediction aim produces much better outcomes in comparison to the conventional classification-based objective.The TACRED dataset achieved a F1-score of 74.8, while the SemEval task 8 dataset achieved a F1-score of 91.9. Span-based joint extraction models have demonstrated their effectiveness in the tasks of entity recognition and relation extraction. The models under consideration see text spans as potential entities and span tuples as potential relation tuples. The sharing of span semantic representations is observed in both entity identification and relation extraction tasks. However, current models have limitations in effectively capturing the semantic information of candidate entities and connections. In order to tackle these issues, Ji et. al [11] propose the implementation of a framework for joint extraction that operates on spans, using attention-based semantic representations.

2.4 Document-Level Relation Extraction

The identification and extraction of significant links between entities referenced in textual documents is a vital topic within the domain of NLP. This process, known as the extraction of document-related relationships, requires an objective evaluation of the textual content to ensure a clear, concise, and necessary presentation of information. The task at hand has great importance in a wide range of applications, including but not limited to improving the retrieval of information, developing organized knowledge graphs, strengthening question-answering systems, and optimizing recommendation algorithms. Despite this, the task of extracting relations from documents has several obstacles that must be addressed. These issues include the need for precise identification and disambiguation of entities, the ability to identify various relations, the filtering of noise in text, an advanced knowledge of contextual information, and the capacity to handle large volumes of documents. The field of NLP has numerous substantial obstacles when it comes to document-level relation extraction, which are crucial to tackle. One of the primary obstacles is in the intrinsic intricacy associated with comprehending and extracting connections that span the entirety of a document. In contrast to the extraction of relations at the sentence level, the extraction of relations at the document level necessitates models that are capable of capturing long-range dependencies and contextual subtleties. This can impose a significant computational effort. Furthermore, the process of disambiguating and accurately identifying entities throughout a document is a significant challenge, especially in cases when several entities possess identical names or when entities are referred to using different variations and aliases. In addition, it is common for documents to contain a significant amount of unstructured and extraneous information, necessitating the implementation of noise reduction and document summarizing as essential components of the extraction procedure. One notable obstacle that arises is the wide range of relationships found within documents, encompassing both clear and well-organized linkages as well as implicit and subtle interconnections. The presence of many sorts of relationships requires the use of adaptable models that can accommodate different language patterns and structures. Assessing the precision of document-level relation extraction models poses an additional difficulty, since the establishment of appropriate metrics for intricate and multifaceted links can be a formidable undertaking. It is imperative to confront these issues in order to progress the area of document-level relation extraction and facilitate the development of a wider array of applications, spanning from information retrieval to knowledge graph creation.

2.5 Document-Level Relation Extraction Datasets

Document-Level Relation Extraction is a key task within the field of NLP, which seeks to reveal and categorize the semantic connections between entities referenced in documents. This approach surpasses the conventional method of extracting information at the sentence level, as it offers a comprehensive comprehension of relationships that can extend over numerous phrases and parts within a document. In order to facilitate progress in the domain of research and development, a range of Document-Level Relation Extraction datasets have been carefully compiled, each presenting distinct problems and prospects. The task of relation extraction in document-level text has distinct obstacles when compared to relation extraction at the sentence level. It is common for documents to consist of many phrases, and it is possible for entities participating in relationships to be referenced in various sections of the document.

Document-level relation extraction plays a crucial role in several domains, such as information retrieval, knowledge graph development, and document processing. The datasets provided are essential for the purpose of training and assessing models that possess the ability to extract relationships within the context of a document. They play a crucial role in driving progress in NLP, allowing for the creation of models that possess the ability to understand and effectively use the intricate connections present within lengthy textual content. In contrast to datasets that focus on sentence-level, document-Level relation extraction in the context of data analysis, datasets often consist of documents that serve as the fundamental unit of analysis. These works encompass several entities, necessitating the identification and categorization of the relationships between them, often spanning numerous sentences and sections. The datasets frequently consist of intricate relationships wherein entities may be referenced in several portions of a text. The successful extraction of relations may need the consolidation of information from many sections within the document.

The task is to extract relationships between entities on a document-level. Datasets play a crucial role in facilitating the advancement of sophisticated NLP models capable of effectively handling the intricacies presented by lengthy textual data. These tools support the investigation of document comprehension and information extraction, providing the basis for innovative methodologies and models that may effectively grasp and utilize intricate connections inside texts. These datasets serve as evidence of the dynamic nature of NLP and its continuous efforts to enhance machine capabilities in extracting knowledge from vast amounts of textual data. Table 2.4 shows the document-level relation extraction datasets details.

Datasets	**Relation Types**	**Dataset Access**
DocRED	96 relation types	Public for commercial use
Re-DocRED	96 relation types	Public for commercial use
ACE 2004	24 relation types	$3,000.00 fee
ACE 2005	33 relation types	$4,000.00 fee

Table 2.4: Document-level datasets details

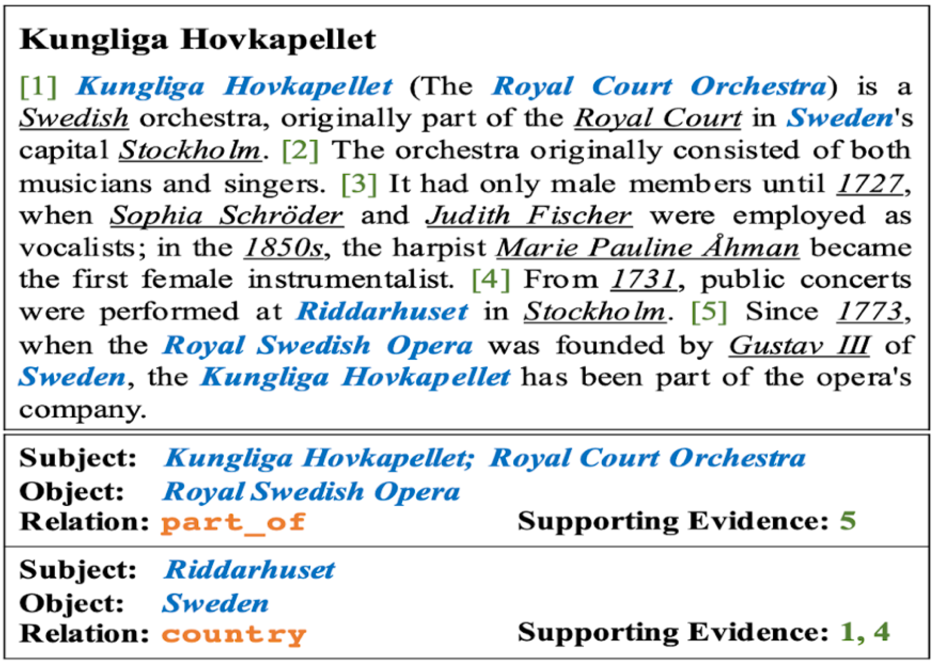
2.5.1 DocRED

The DocRED dataset, also known as the Document-Level Relation Extraction Dataset, is a collection of data specifically designed for relation extraction tasks. This dataset was created by utilizing information from reputable sources such as Wikipedia and Wikidata. Every document inside the collection has been manually annotated by humans, including named entity mentions, coreference information, intra-sentence and inter-sentence relationships, as well as supporting evidence. The DocRED system necessitates the examination of many sentences inside a document in order to extract entities and deduce their relationships by combining all available information from the document. In addition to the manually annotated data, the dataset includes a substantial amount of distantly supervised data on a huge scale. The DocRED dataset has a total of 132,375 entities and 56,354 relational facts that have been meticulously annotated throughout 5,053 Wikipedia entries. In addition to the data that has been annotated by humans, the collection also includes a substantial amount of distantly supervised data, spanning over 101,873 pages. Distant supervision is a method used for the process of annotating data for relation extraction by using an already established knowledge database.

The researchers gathered a dataset that was annotated by human annotators. The individuals involved completed four distinct stages. The collection of their human-annotated data occurs in four distinct stages. (1) The process of creating distantly supervised annotations for Wikipedia documents. (2) The task involves the annotation of all named entity mentions included in the papers, as well as the inclusion of coreference information. (3) Establishing connections between named entity references and corresponding objects in Wikidata. (4) The process of assigning labels to relationships and the corresponding evidence [34]. Figure 2.7 shows an example from DocRED dataset.

A random sample of 300 documents was taken from the development and test sets, which collectively contained 3,820 instances of relations. The researchers then conducted a manual analysis to determine the forms of reasoning necessary to extract these relations. Figure 2.8 presents statistical data pertaining to the primary forms of reasoning observed within the sample.

The DocRED dataset is a highly significant resource within the domain of NLP, since it has been particularly curated to be useful to relation extraction tasks within the context of document-level text. The training and evaluation of machine learning or deep learning models that seek to discover and categorize links between entities stated in documents play a vital role in allowing many applications, including knowledge base development, information retrieval, and document interpretation.



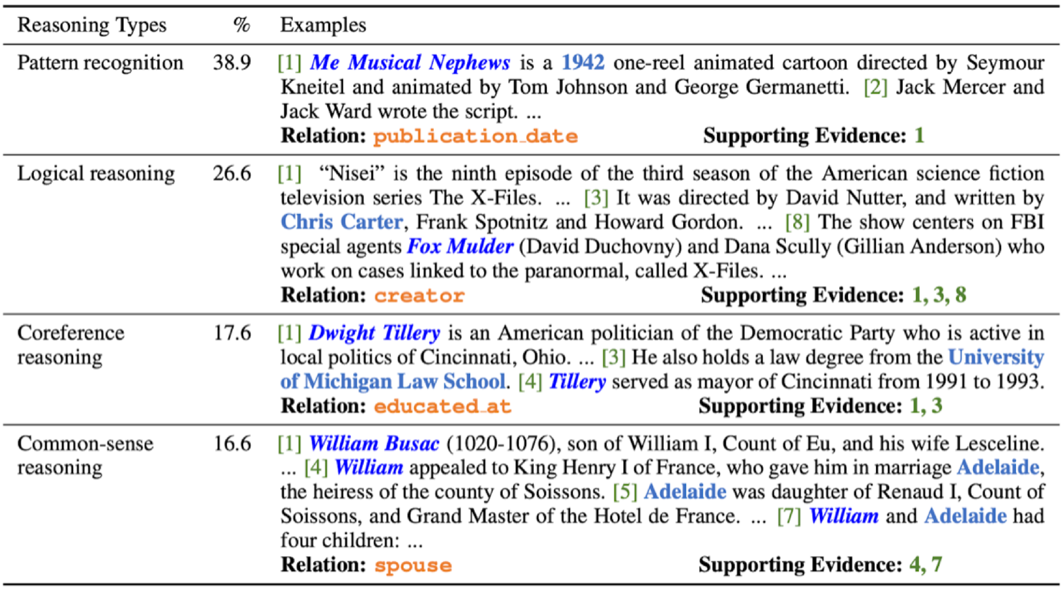
Kungliga Hovkapellet

[1] Kungliga Hovkapellet (The Royal Court Orchestra) is a Swedish orchestra, originally part of the Royal Court in Sweden's capital Stockholm. [2] The orchestra originally consisted of both musicians and singers. [3] It had only male members until 1727, when Sophia Schröder and Judith Fischer were employed as vocalists; in the 1850s, the harpist Marie Pauline Åhman became the first female instrumentalist. [4] From 1731, public concerts were performed at Riddarhuset in Stockholm. [5] Since 1773, when the Royal Swedish Opera was founded by Gustav III of Sweden, the Kungliga Hovkapellet has been part of the opera's company.

Subject: Kungliga Hovkapellet; Royal Court Orchestra
Object: Royal Swedish Opera
Relation: part_of **Supporting Evidence:** 5

Subject: Riddarhuset
Object: Sweden
Relation: country **Supporting Evidence:** 1, 4

Figure 2.7: An example from DocRED



Reasoning Types	%	Examples
Pattern recognition	38.9	[1] Me Musical Nephews is a 1942 one-reel animated cartoon directed by Seymour Kneitel and animated by Tom Johnson and George Germanetti. [2] Jack Mercer and Jack Ward wrote the script. ... **Relation:** publication_date **Supporting Evidence:** 1
Logical reasoning	26.6	[1] "Nisei" is the ninth episode of the third season of the American science fiction television series The X-Files. ... [3] It was directed by David Nutter, and written by Chris Carter, Frank Spotnitz and Howard Gordon. ... [8] The show centers on FBI special agents Fox Mulder (David Duchovny) and Dana Scully (Gillian Anderson) who work on cases linked to the paranormal, called X-Files. ... **Relation:** creator **Supporting Evidence:** 1, 3, 8
Coreference reasoning	17.6	[1] Dwight Tillery is an American politician of the Democratic Party who is active in local politics of Cincinnati, Ohio. ... [3] He also holds a law degree from the University of Michigan Law School. [4] Tillery served as mayor of Cincinnati from 1991 to 1993. **Relation:** educated_at **Supporting Evidence:** 1, 3
Common-sense reasoning	16.6	[1] William Busac (1020-1076), son of William I, Count of Eu, and his wife Lesceline. ... [4] William appealed to King Henry I of France, who gave him in marriage Adelaide, the heiress of the county of Soissons. [5] Adelaide was daughter of Renaud I, Count of Soissons, and Grand Master of the Hotel de France. ... [7] William and Adelaide had four children: ... **Relation:** spouse **Supporting Evidence:** 4, 7

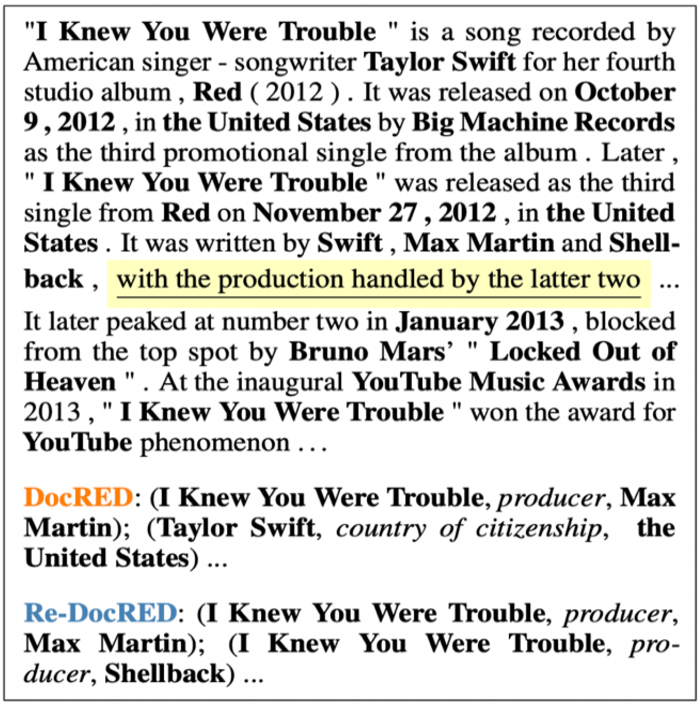
Figure 2.8: Types of reasoning requirement for DocRED

2.5.2 Re-DocRED

The DocRED benchmark is extensively utilized for the task of document-level connection extraction. Nevertheless, it is worth noting that the DocRED dataset exhibits a considerable proportion of false negative instances, which might be attributed to insufficient annotation. A total of 4,053 documents from the DocRED collection were subjected to revision and subsequent resolution of identified issues. The dataset was made publicly available under the name "Re-DocRED dataset." Figure 2.9 shows an example from Re-DocRED dataset.

The Re-DocRED dataset successfully addressed the previously mentioned challenges seen in the DocRED dataset [27]:

- The incompleteness problem was addressed by augmenting a substantial quantity of related triples.
- The logical contradictions present in DocRED were examined and discussed.
- The coreferential errors inside DocRED have been corrected.



"**I Knew You Were Trouble** " is a song recorded by American singer - songwriter **Taylor Swift** for her fourth studio album , **Red** (2012) . It was released on **October 9 , 2012** , in **the United States** by **Big Machine Records** as the third promotional single from the album . Later , " **I Knew You Were Trouble** " was released as the third single from **Red** on **November 27 , 2012** , in **the United States** . It was written by **Swift** , **Max Martin** and **Shellback** , with the production handled by the latter two ...
It later peaked at number two in **January 2013** , blocked from the top spot by **Bruno Mars**' " **Locked Out of Heaven** " . At the inaugural **YouTube Music Awards** in 2013 , " **I Knew You Were Trouble** " won the award for **YouTube** phenomenon . . .

DocRED: (**I Knew You Were Trouble**, *producer*, **Max Martin**); (**Taylor Swift**, *country of citizenship*, **the United States**) ...

Re-DocRED: (**I Knew You Were Trouble**, *producer*, **Max Martin**); (**I Knew You Were Trouble**, *producer*, **Shellback**) ...

Figure 2.9: One sample from Re-DocRED

The subsequent iteration of the system, Re-DocRED, surpasses its antecedent, DocRED, by the integration of several enhancements and modifications. The Re-DocRED dataset is a significant advancement in the field of document-level relation extraction, building upon the previous dataset, DocRED. The enhanced dataset presented here provides a crucial resource for the purposes of research and development, therefore establishing its position as a favored option for the progression of approaches related to document-level connection extraction.

2.5.3 ACE 2004

The ACE 2004 Multilingual Training Corpus encompasses the entirety of the English, Arabic, and Chinese training data utilized in the 2004 Automatic Content Extraction (ACE) technology evaluation. The corpus comprises annotated data of several sorts for entities and relations. It was developed by the Linguistic Data Consortium, with funding from the ACE Program, and with supplementary support from the DARPA TIDES Program (Translingual Information Detection, Extraction, and Summarization). The primary goal of the ACE program is to further the development of automatic content extraction technology, which facilitates the automated processing of human language in written form. The evaluation of sites in September 2004 encompassed the assessment of system performance across six distinct areas. These categories include Entity Detection and Recognition (EDR), Entity Mention Detection (EMD), EDR Co-reference, Relation Detection and Recognition (RDR), Relation Mention Detection (RMD), and RDR given reference entities. The evaluation of all tasks was conducted in three languages, namely English, Chinese, and Arabic [1]. The ACE annotators conducted tagging on several types of data, including broadcast transcripts, newswire, and newspaper data, in three different languages: English, Chinese, and Arabic. This process resulted in the creation of both training and test datasets, which were used for evaluating typical research tasks. The study encompassed three main ACE annotation tasks, which aligned with the three research objectives: Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC), and Event Detection and Characterization (EDC). One further annotation activity, known as Entity Linking (LNK), included the consolidation of all mentions of a certain entity and its associated attributes into a Composite Entity.

- **Entity Detection and Tracking (EDT)** The key annotation job of the study was EDT, which served as the fundamental basis for all subsequent tasks. Subsequent ACE tasks have delineated seven distinct categories of entities, namely Person, Organization, Location, Facility, Weapon, Vehicle, and Geo-Political Entity (GPEs). Each category was further subdivided into subcategories (such as Organization subcategories encompassing Government, Commercial, Educational, Non-profit, and Other). The annotators have assigned tags to all instances of each entity mentioned in the document, regardless of whether they are named, nominal, or pronominal. The annotator determined the maximum length of the string that represents the entity and assigned a label to the head of each mention. The capturing of nested mentions was also seen. Every entity was categorized based on its type and subtype, and subsequently labeled based on its class, which may be particular, generic, attributive, negatively quantified, or underspecified. In the LNK annotation work, annotators conducted a comprehensive examination of the text with the purpose of categorizing mentions of identical entities into cohesive groups. Additionally, they identified instances of metonymy, when the name of one thing is employed to refer to another entity (or entities) that are associated with it.

- **Relation Detection and Characterization (RDC)** The task of RDC encompasses the process of identifying and characterizing relationships that exist between items. The inclusion of this assignment occurred during the implementation of Phase 2 of the ACE project. The research and development committee focused on various types of relations, both physical and social/personal. These included located relations, near relations, and part-whole relations. Additionally, the committee examined a range of employment or membership relations, as well as relations between artifacts and agents, such as ownership. Affiliation-type relations, such as ethnicity, were also considered, along with relationships between individuals and geopolitical entities, such as citizenship. Lastly, discourse relations were explored as well. In each connection, the annotators identified two main arguments, specifically the two ACE items that are connected, together with the temporal properties of the relation. Delineations were made between relations that were substantiated by explicit textual evidence and those that relied on contextual inference by the reader.

- **Event Detection and Characterization (EDC)** In the field of EDC, annotators have successfully discovered and classified five distinct categories of events in which entities belonging to the EDT system engage. The sorts of events targeted in this study encompassed many categories, namely Interaction, Movement, Transfer, Creation, and Destruction. The textual mention or anchor for each occurrence was tagged by annotators, who also classified it based on its kind and subtype. The researchers further distinguished event arguments, including agent, object, source, and target, as well as qualities such as temporal and locative, in accordance with a template particular to each kind.

2.5.4 ACE 2005

The ACE 2005 Multilingual Training Corpus was created by the Linguistic Data Consortium (LDC). It comprises over 1,800 files that encompass a variety of genres in English, Arabic, and Chinese. These texts have been annotated to identify entities, relations, and events. This dataset encompasses the entirety of the training data available in the specified languages for the 2005 Automatic Content Extraction (ACE) technology evaluation. The genres encompassed in this study consist of newswire, broadcast news, broadcast conversation, weblog, discussion forums, and conversational telephone speech [28]. The ACE program placed its emphasis on several tasks pertaining to the extraction of information from textual data, encompassing NER, RE, and Event Extraction. The aforementioned tasks were specifically devised with the objective of extracting organized information from text sources that lack a predefined structure. The dataset encompassed many named entity categories, such as individuals, corporations, geographical areas, dates, times, numerical quantities, and more categories. The ACE dataset comprises documents that have undergone manual annotation by human annotators to identify named entities, relationships between entities, and events. The annotations provided explicit details on the spatial positions of entities, the categorization of entities, and the interconnections between them within the textual context. The ACE datasets, such as ACE 2005, have played a pivotal role in facilitating advancements in information extraction research and serving as a standard for evaluating the effectiveness of NLP systems. Scholars have utilized ACE data to create and evaluate different NLP methodologies for the purpose of detecting entities, relations, and events within textual data. This is particularly relevant when considering diverse languages and domains.

2.6 Document-Level Relation Extraction Related Work

In this section, we will focus on SOTA studies pertaining to document-level analysis. We will go into the methodologies employed in these papers and provide an overview of their respective findings and outcomes.

2.6.1 Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling

The task of document-level RE is more complicated in comparison to sentence-level RE. Document-level relation extraction introduces two challenges; the multi-entity problem and the multi-label problem. Zhou et. al [38] employed the computation of embeddings for individual items, which are subsequently combined into pairs. Following this, the classification will be performed based on the adaptive-thresholding loss technique. This technique allows for the learning of an adaptive threshold that is specific to each pair of entities. By doing so, it aims to minimize the judgment mistakes that arise from using a single global threshold. Additionally, the adaptation of embedding representation for entity pairs may be achieved by the utilization of localized context pooling. This technique involves capturing context that is closely associated with the entity pairs in order to enhance the quality of the entity representations. The datasets included in this study encompassed DocRED, CDR, and GDA.

2.6.2 Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation

Tan et. al [26] conducted a study on the topic of document-level connection extraction, specifically focusing on the utilization of adaptive focal loss and knowledge distillation techniques. The issue pertaining to document-level connection extraction has significant importance within the realms of information extraction and NLP research. The methodology presented for document-level connection extraction incorporates the utilization of knowledge distillation, axial attention, and adaptive focus loss.

Figure 2.10 shows the their architecture. The initial phase is extracting the contextual representation for each pair of entities using a language model that has undergone pre-training. The utilization of the feedforward neural network classifier is employed to obtain the logits and compute the associated losses. The utilization of the suggested adaptive focus loss aims to optimize the learning process for courses with low occurrence rates. Knowledge distillation is employed as a means to address the disparities that exist between human annotated data and distantly supervised data. The instructor model undergoes training using annotated data, and the resultant output is then employed as soft labels. The student model undergoes pre-training with both soft labels and remote labels. The pre-trained student model will be subjected to additional fine-tuning using the annotated data.

The performance of the model is assessed using the DocRED and HacRED benchmark. The model's performance falls short of human performance, suggesting the existence of potential areas for enhancement. In addition to evaluating the models' performance, it is remarkable that across all methods, HacRED consistently exhibits much superior absolute performance compared to its performance on DocRED. The concentration of HacRED on hard relations, as opposed to the more comprehensive approach of DocRED, presents a counter-intuitive aspect. This study focuses on the issues of class imbalance and logical reasoning in the context of connection extraction. Knowledge distillation is employed as a means to address the disparities that exist between human annotated data and distantly supervised data.

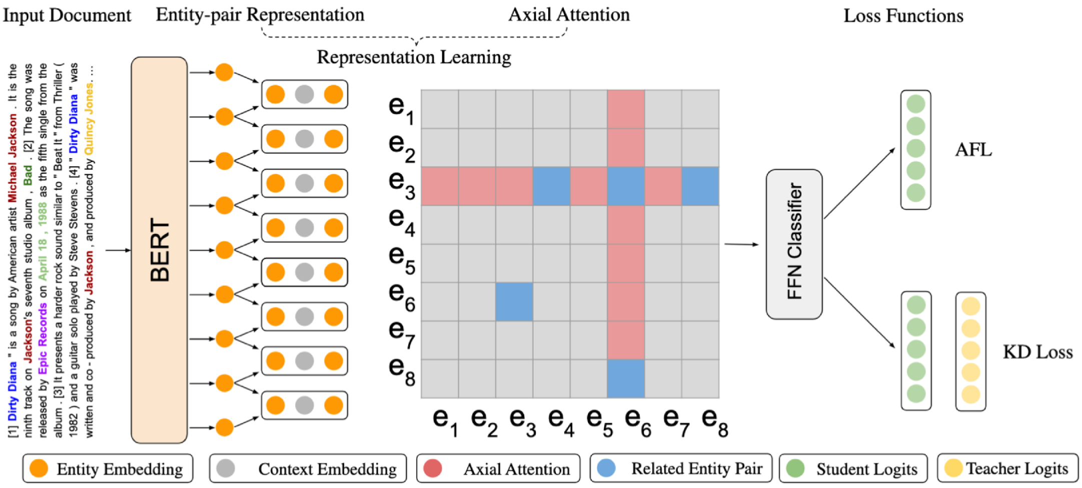

Input Document
Entity-pair Representation
Axial Attention
Representation Learning
Loss Functions
[1] Dirty Diana " is a song by American artist Michael Jackson . It is the ninth track on Jackson's seventh studio album , Bad . [2] The song was released by Epic Records on April 18 , 1988 as the fifth single from the album . [3] It presents a harder rock sound similar to " Beat It " from Thriller (1982) and a guitar solo played by Steve Stevens . [4] " Dirty Diana " was written and co - produced by Jackson , and produced by Quincy Jones. ...
BERT
e1 e2 e3 e4 e5 e6 e7 e8
FFN Classifier
AFL
KD Loss
Entity Embedding
Context Embedding
Axial Attention
Related Entity Pair
Student Logits
Teacher Logits

Figure 2.10: Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation Architecture [26]

2.6.3 Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling

Zhou et. al [38] focuses exclusively on the transformer architecture and introduces a unique adaptive thresholding loss for addressing the multi-label problem in the context of DocRE. Additionally, it integrates the contextual information with the aggregated attention weights assigned to each item. In their study, the authors assert that there is a lack of existing research specifically addressing the issue of class imbalance in the context of DocRE. This work only concentrates on threshold learning as a means to achieve a balance between positive and negative instances. However, it neglects to tackle the issue of class imbalance specifically within the positive examples. They suggest the utilization of localized context pooling (Figure 2.11) as a means to enhance entity representations by using pre-trained attention to extract relevant context for entity pairings. Context-enhanced Entity Representation facilitates the transfer of established dependencies from the pre-trained language model, hence avoiding the need to learn additional attention layers from the beginning.

The entities are mapped to hidden states using a linear layer followed by a non-linear activation function. The probability of the relation is then calculated

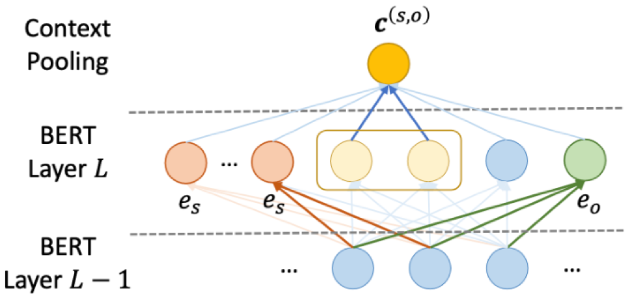


Figure 2.11: Illustration of localized context pooling [38]

using a bilinear function and a sigmoid activation function. The adaptive-thresholding loss in this study was formulated by modifying the normal categorical cross entropy loss. The study introduces the concept of Adaptive Focal Loss (AFL) as a potential improvement to Adaptive Thresholding Loss (ATL) (2.12) in the context of long-tail classes. This study has significant importance in the context of multi-label classification issues within the domain of DocRE [38].

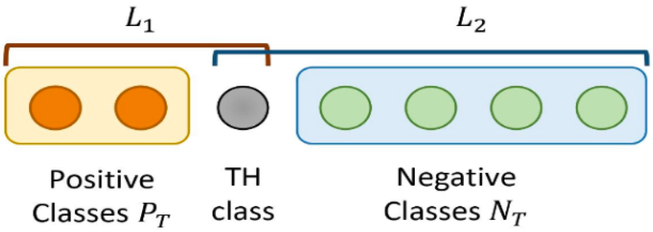


Figure 2.12: Illustration of adaptive-thresholding loss [38]

2.6.4 Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction

Xu et. al [32] has demonstrated that the utilization of distantly supervised data has the potential to enhance the efficacy of document-level connection extraction. They emphasized the need of incorporating contextual information for achieving accurate results. The suggested SSAN model presented in their study utilizes document text as its input and constructs contextual representations by including the entity structure throughout the encoding step. This study employs the "Naive Adaptation" approach, which involves two steps. Firstly, the model is pretrained using distantly supervised data and the relation extraction loss. Subsequently, the model is fine-tuned using human-annotated data, while maintaining the same aim. The SSAN model demonstrates superior performance compared to other competitor baselines, successfully accomplishing document-level relation extraction tasks.

2.6.5 Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation

The calculation of axial attention involves applying self-attention separately along the height and width axes. After each computation along these axes, a residual connection is made. The objective of Wang et al. [30] is to employ this approach in order to decrease the computing cost associated with semantic segmentation. The research asserts that axial attention demonstrates strong performance not just as an independent model for picture categorization, but also as a foundational component for panoptic segmentation, instance segmentation, and semantic segmentation tasks. According to this article, an axial-attention layer is responsible for the propagation of information along a certain axis. In order to effectively gather global information, they utilize two axial-attention layers in a sequential manner, with each layer focusing on the height-axis and width-axis accordingly. This approach helps to minimize the memory usage when dealing with large feature maps.

2.6.6 Document-level Relation Extraction as Semantic Segmentation

Zhang et. al [35] model approaches the job of connection extraction in a manner that is analogous to the methodology employed in semantic segmentation within the field of computer vision. In this study, a Convolutional Neural Network (CNN) architecture was employed to encode the interaction between entity pairs. However, it is important to note that the CNN structure utilized in this paper has limitations in capturing all the elements present within the two-hop reasoning paths. Additionally, the focus of the CNN structure was primarily on threshold learning to achieve a balance between positive and negative examples. However, it is worth mentioning that the issue of class-imbalance within positive examples was not specifically addressed in this research. Tan et. al [26] conduct a comparative analysis between their proposed approach and a previous study that utilizes Convolutional Neural Networks (CNNs) for encoding neighbor information in relation categorization. Tan et. al [26] posit that directing attention towards the axial elements yields more effectiveness and intuitiveness.

2.6.7 Learning from Context or Names? An Empirical Study on Neural Relation Extraction

Peng et. al [20] conduct an empirical investigation on the impact of two primary types of information in text: textual context and entity mentions. The researchers discover that (i) contextual information is the primary means by which predictions are supported in RE models, but these models also heavily depend on information derived from entity mentions, primarily in the form of type information. Additionally, (ii) it is observed that current datasets may inadvertently incorporate shallow heuristics through entity mentions, thereby contributing to the high performance observed in RE benchmarks. The authors put out a paradigm for contrastive pre-training in relation extraction (RE) that incorporates entity masking. This approach aims to enhance comprehension of textual context and type information, while mitigating the risk of memorizing entities or relying on superficial signals in mentions. The extraction of relations from documents is a crucial work in the field of natural language processing. This activity plays a significant role in acquiring organized knowledge and enhancing information retrieval in many applications.

2.6.8 Other Related Works

Liu et. al [15] put out a theoretical framework that is capable of encoding a document while simultaneously generating intricate structural connections. This method integrate a differentiable non-projective parsing algorithm into a neural model and employ attention mechanisms to effectively include the structural biases. The datasets included in this study encompassed a variety of sources, including Yelp reviews, IMDB ratings, Czech reviews, and Congressional floor discussions. The application of seq2seq approaches has led to advancements in addressing structured prediction challenges. Instead of constructing a linguistic representation of a series of words in a given target language, the researchers [14] develop a model that represents a collection of activities linked to each individual stage in the decoding process. The proposed approach involves the utilization of a Feed-Forward Neural Network (FFNN) for the purpose of scoring items, and an autoregressive log-linear model for the computation of probabilities. The dataset used for end-to-end relation extraction is ACE-05. Xiao et. al [31] suggest a method called Supervising and Augmenting Intermediate Steps (SAIS) for RE, which aims to train the model to effectively collect important contexts and object kinds. The SAIS technique, which is suggested, utilizes a range of meticulously constructed tasks. This method not only improves the quality of extracted relations through more effective supervision, but also enhances interpretability by properly retrieving the related supporting evidence. The SAIS system demonstrates exceptional performance in the field of RE across three benchmark datasets, namely DocRED, CDR, and GDA.

3

EXPERIMENTS AND ANALYSIS

3.1 Experiement

For our experiment, the TACRED dataset was utilized. Two different methods were carried out, the first using the LSTM model and the second using the BERT model. This section outlines the model specifics, presents a comparison analysis, undertakes an error analysis and evaluates our calculation metrics.

3.1.1 LSTM

LSTM is a specific architecture of a recurrent neural network (RNN) that aims to address the inherent constraints of conventional Recurrent Neural Networks (RNNs) when it comes to acquiring and retaining long-term dependencies within sequential input [10]. LSTM networks have a more intricate structural design in comparison to conventional RNNs. Memory cells are present inside these structures, enabling the storage of information over extended durations. Additionally, they possess a range of gates that regulate the transmission of information.

An obstacle of RNNs is their reliance on a "short-term memory" mechanism, which essentially involves storing and recalling past input. Upon reaching its memory capacity, the device proceeds to expunge the most chronologically preserved information and subsequently substitutes it with fresh data. LSTM model attempts to address this issue by selectively storing relevant information in its long-term memory. The storage of long-term memory occurs within a cellular structure known as the Cell State. Furthermore, it is worth noting the existence of the concealed state, a concept familiar in conventional neural networks, where it serves as a repository for short-term information derived from previous computational iterations. During each computational iteration, the current input is utilized alongside previous states of short-term memory and hidden state. The hidden state refers to the temporary memory of the model. The regulation of the cell state occurs via the input gate, forget gate, and output gate. Figure 3.1 shows LSTM architecture.

- The input gate is responsible for regulating the flow of information into the cell state.
- The forget gate is responsible for determining the preservation or elimination of information from the cell state.
- The output gate is responsible for controlling the flow of information that comes out from the cell state.

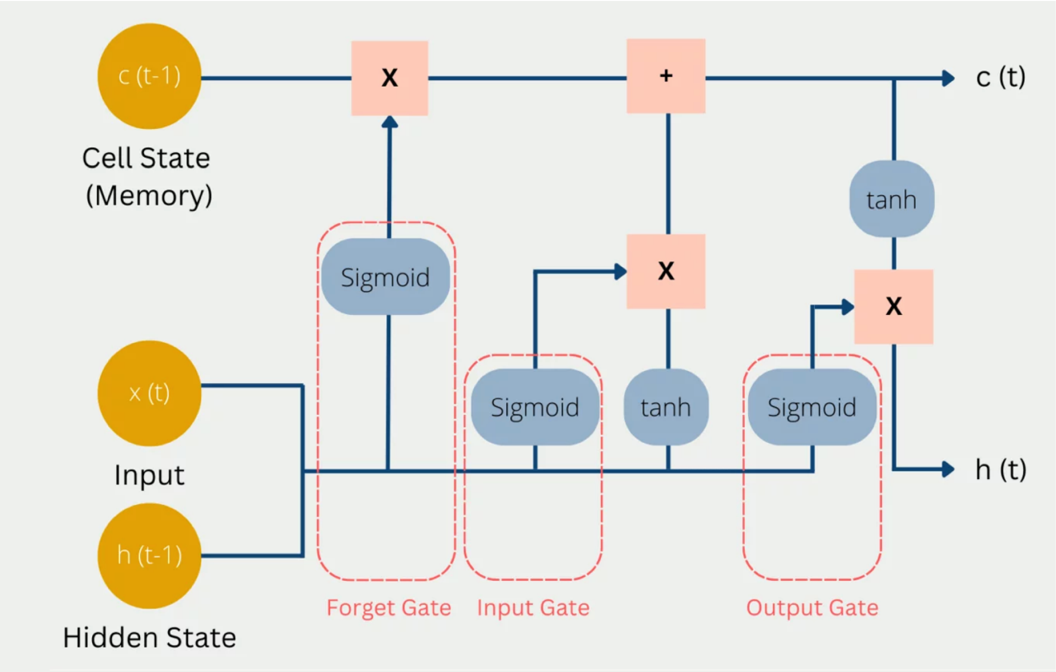


Figure 3.1: LSTM architecture

LSTM models are extensively employed in many NLP applications, such as machine translation, sentiment analysis, relation extraction. LSTM model have been used in several relation extraction applications such as clinical texts [18], cybersecurity [5], e-commerce [23]. The significance of LSTM models in relation extraction tasks lies in their capacity to adeptly understand and represent extensive relationships within sequential data. Relation extraction is the process of discovering and categorizing the connections between entities inside a given text. This entails determining if a sentence conveys a cause-and-effect relationship, an affiliation, or any other form of association. There are several justifications for the significance of LSTM in relation extraction models.

Contextual Understanding: LSTM models demonstrate exceptional proficiency in collecting contextual information and sequential relationships within textual data. The proper determination of the link between things in relation extraction heavily relies on a comprehensive grasp of the contextual factors surrounding these entities. LSTM networks has the ability to retain pertinent information throughout extended sequences, hence enabling them to take into account the contextual relationship between words and phrases that may be spatially distant within the input text.

Memory Cells for Long-Term Information: LSTM models are equipped with memory cells that possess the capability to retain knowledge for prolonged durations. This unique characteristic enables the model to effectively preserve and recall significant contextual information. This phenomenon is especially advantageous in the context of relation extraction, as the necessary information required to detect a link may be distributed across a given text.

Sequential Pattern Recognition: The process of relation extraction frequently necessitates the identification and comprehension of patterns and interdependencies within the sequential arrangement of words or sentences. LSTM models are particularly suitable for tasks that require the identification of patterns within sequential data, given to their inherent sequential character.

Effective Feature Extraction: LSTM models has the capability to autonomously learn hierarchical representations of incoming data. Within the domain of relation extraction, this implies that the model possesses the capability to autonomously extract relevant characteristics and representations from the input text, hence decreasing the necessity for manual feature engineering [13].

LSTM networks play a vital role in relation extraction models due to their capacity to capture extensive dependencies, comprehend contextual information, and efficiently handle sequential data of varying lengths. These capabilities are indispensable for accurately discerning connections between entities in textual data expressed in natural language.

3.1.2 BERT

BERT model is specifically developed to pre-train deep bidirectional representations from unlabeled text. It achieves this by simultaneously considering both the left and right context in all layers of the model. The model is built around the Transformer architecture and is specifically engineered to comprehend the contextual meaning of words inside a phrase by taking into account the neighboring words from both preceding and succeeding directions. BERT is trained on a substantial corpus of textual data and acquires the ability to anticipate missing of words within phrases, hence enhancing its comprehension of the complex relationships and subtleties of words [4]. The process of pre-training enhances the performance of BERT in a remarkable manner across a range of NLP activities, including question answering, sentiment analysis, and text classification, eliminating the requirement for training that is particular to each job.

BERT utilizes a Transformer, which is a neural network architecture that exploits an attention mechanism that acquires contextual connections among words inside a given text. The fundamental structure of a Transformer model comprises an encoder component responsible for processing the input text and a decoder component responsible for generating predictions for the given job. The input to the encoder in BERT consists of a series of tokens, which are initially transformed into vectors and subsequently subjected to neural network processing. The input representation of a particular token is formed by adding together the token embedding, segment embedding, and position embedding that correspond to that token. Figure 3.2 provides a graphic representation of this architecture.

- Token embeddings: The input word tokens are expanded to include a [CLS] token at the start of the first sentence, and each sentence ends with a [SEP] token.
- Segment embeddings: Each token has a marker attached to it that indicates either Sentence A or Sentence B. The encoder can now discriminate between sentences as a result of this.
- Positional embeddings: Each token has a positional embedding added to it to represent its place in the sentence.

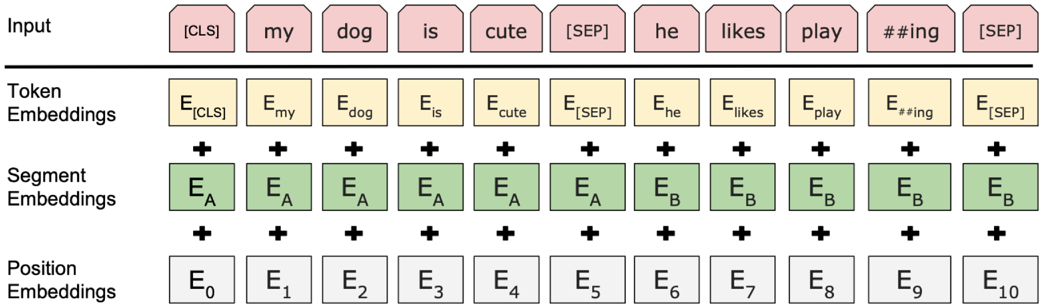


Figure 3.2: BERT input representation [4]

The model's pre-training does not involve the use of traditional left-to-right language models. Instead, it focuses on addressing two unsupervised tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). BERT utilizes an advanced methodology known as MLM, wherein words inside the phrase are randomly masked and subsequently predicted. Masking refers to the process in which a model examines the entirety of a phrase, encompassing both preceding and subsequent context, to predict the masked word. The model's main objective is in predicting the presence of the [MASK] token in the input, whereas its goal is for the model to accurately predict the correct tokens irrespective of the token present in the input. In order to address this problem, a subset of tokens amounting to 15% were chosen for the purpose of masking:

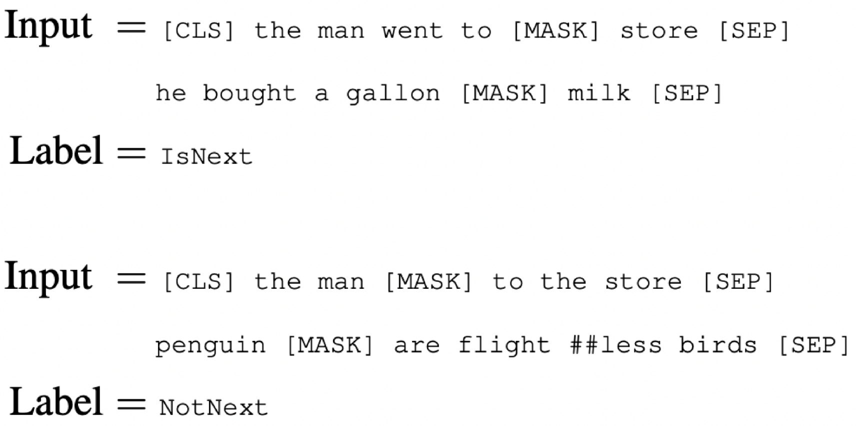
- 80% of the tokens are replaced with the token [MASK].
- Tokens are changed by random tokens 10% of the time.
- 10% of the time tokens are left unchanged.

During the training process, the BERT loss function exclusively takes into account the prediction of the masked tokens, disregarding the prediction of the non-masked tokens.

The BERT training technique incorporates next sentence prediction as a means to comprehend the link between two phrases. A pre-existing model with this level of comprehension is applicable in tasks such as question answering. During the training process, the model is provided with pairs of sentences as input and it acquires the ability to predict whether the second sentence is the subsequent sentence in the original text. The BERT model utilizes a distinct [SEP] token that marks sentence limits. During the training process, the model is provided with two input phrases simultaneously:

- Half of the time the second sentence comes after the first one.
- Half of the time, it is a sentence chosen at random from the whole corpus.

BERT is afterwards tasked with predicting if the second phrase exhibits randomness, operating on the assumption that the random statement will lack coherence with the first sentence. In order to determine the coherence between the first and second sentences, the entire input sequence is processed using the Transformer-based model. Figure 3.3 shows next sentence prediction task illustration.



Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]
Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]
Label = NotNext

Figure 3.3: The next sentence prediction task [4]

The utilization of BERT is of crucial significance in connection extraction activities owing to its capacity to effectively capture contextualize information and complicated connections present inside sentences. The process of relation extraction entails the identification and classification of connections between items inside a given text. This requires evaluating if two entities are linked by a certain sort of relationship, such as "born in" or "works for". BERT model have been used in several relation extraction applications such as medical [21, 33], finance [9], biomedical [25, 7] and geological reports [17]. The significance of BERT in connection extraction tasks is based on its ability to effectively capture contextual information and semantic relationships between entities, hence enhancing the accuracy and performance of relation extraction models. For relation extraction tasks, it's important to use the BERT following reasons for:

Contextual Understanding: BERT has exceptional proficiency in gathering and comprehending contextual information [4]. Many conventional relation extraction algorithms commonly depend on context windows of constant size, which could not effectively reflect the intricate and fluctuating characteristics of language. BERT, because to its bidirectional training and attention processes, takes into account the complete context of a phrase, hence enabling a more sophisticated comprehension of relationships.

Bidirectional Training: BERT is trained using a bidirectional approach, which takes into account both the left and right context. The understanding of the complete context surrounding things, including the preceding and succeeding words, is of the greatest significance in relation extraction. The implementation of a comprehensive strategy enhances the model's capacity to accurately perceive and analyze the underlying connections between things.

Transfer Learning: BERT gets pre-training using an enormous amount of varied textual data, enabling it to acquire knowledge of large language patterns and semantics. The process of pre-training enables the model to acquire a comprehensive comprehension of language [22]. By undergoing fine-tuning for certain connection extraction tasks, BERT effectively utilizes its pre-existing knowledge to achieve exceptional performance in the identification and classification of links between entities.

Handling Ambiguity: The interpretation of relations within text can sometimes be subject to ambiguity and dependence on the surrounding context. The contextual embeddings of BERT allow it to effectively address the issue of ambiguity by taking into account the complete context of the statement. This aspect has significant importance in relation extraction tasks, as the interpretation of a relation can be influenced by the context in which it is expressed.

The contextual comprehension, bidirectional training, transfer learning capabilities, and efficacy in managing ambiguity exhibited by BERT render it a potent instrument for relation extraction activities. Consequently, it enables the detection of connections between entities in natural language text with enhanced precision and complexity.

3.2 Evaluation Metrics

Metrics are of crucial significance when it comes to assessing the effectiveness of prediction models, algorithms, or classifiers within the fields of artificial intelligence and statistical analysis. Quantitative measures are utilized to evaluate the effectiveness of a model in addressing a certain activity or problem. The choice of evaluation metrics depends on the underlying characteristics of the problem, the goals of the study, and the specific aspects of the dataset. A variety of different metrics may be required for various activities, and in certain instances, a combination of metrics is necessary to thoroughly evaluate the performance of a model. The main objective of our study will be the evaluation of precision, recall, F1-score, and confusion matrix analysis.

3.2.1 Precision and Recall

Precision and recall are fundamental metrics utilized in the domains of information retrieval, machine learning, and statistics. These metrics are widely employed to assess the effectiveness of categorization. The aforementioned measures hold special significance in the context of unbalanced datasets, when one class exhibits a much higher frequency compared to the other. Precision, also known as positive predictive value, is a metric that quantifies the degree of accuracy exhibited by a model in its positive predictions. The calculation of precision is determined through the use of the following formula:

Precision = $\frac{TP}{TP+FP}$

- True positives (TP), refer to the count of accurately predicted positive cases.
- False positives (FP), refer to the situations that are inaccurately classified as positive when they are, in fact, negative.

A high level of precision is indicative of a model with a low incidence of false positives, hence demonstrating its proficiency in accurately recognizing positive cases.

Recall, which is often referred to as sensitivity or true positive rate, quantifies the capacity of a model to identify all positive instances. The calculation of recall is determined through the use of the following formula:

Recall = $\frac{TP}{TP+FN}$

- True positives (TP), refer to the count of accurately predicted positive cases.
- False negatives (FN), refer to the events that were really positive but were inaccurately classified as negative.

A high recall value signifies that the model possesses the ability to accurately detect a significant majority of positive cases while minimizing the number of false negatives. Precision and recall metrics offer a thorough evaluation of a model's performance, particularly in scenarios with unbalanced datasets or situations where one sort of mistake (either false positives or false negatives) has greater significance than the other. They assist in assessing the trade-offs involved in achieving accurate positive forecasts and capturing all true positive situations. Precision and recall have significant importance within the context of NLP as a result of many reasons.

- Imbalanced data is a common occurrence in NLP applications such as sentiment analysis, named entity identification, and information retrieval. In the context of sentiment analysis, it is possible that the number of neutral or negative examples exceeds the number of positive examples. The importance of precision and recall in such instances lies in their ability to offer a more comprehensive evaluation of the model's performance for the minority class.

- The evaluation of question answering systems depends heavily on recall and precision metrics, which play a significant role in determining the system's ability to accurately identify important sections or responses among an extensive amount of product. The achievement of high precision reduces the occurrence of false positives, and high recall guarantees the retrieval of all relevant responses.

- The evaluation of information extraction tasks involves the assessment of precision and recall, which measure how precise the system is in extracting targeted information, such as events or connections, from unstructured text.

- NER involves the identification of specified entities, such as names of individuals, organizations, or locations, within a given text. Precision and recall metrics are commonly used to evaluate the accuracy of NER systems in accurately identifying these entities. Achieving a high level of precision is crucial in order to minimize the occurrence of false positive entity recognition. On the other hand, a high level of recall is necessary to guarantee that no critical entities are missed or neglected.

- The field of NLP frequently encompasses applications related to information retrieval, which entail the process of searching and retrieving pertinent materials or providing responses to user inquiries. The evaluation of a retrieval system's performance in accurately identifying the most relevant documents and reducing the occurrence of false positives relies on the metrics of precision and recall.

The optimal trade-off between recall and precision in the field of NLP is contingent upon the particular application and its corresponding demands. Certain activities may place a greater emphasis on achieving high precision, which involves reducing the occurrence of false positives. Conversely, other tasks may prioritize good recall, which involves minimizing the occurrence of false negatives.

3.2.2 F1-Score

The F1-score, also referred to as the F1 measure or F1 statistic, is a commonly used metric in the domains of machine learning, information retrieval, and statistics. It serves the purpose of assessing the accuracy of binary classification models. Dealing with unbalanced datasets, characterized by a significant disparity in class frequencies, is especially advantageous. The F1-score is a mathematical measure that combines recall and precision in a harmonic mean, offering a well-balanced assessment of a model's effectiveness. The term has been defined as follows:

F1-score = $\frac{2*Precision*Recall}{Precision+Recall}$

The F1-score is a statistic that integrates precision and recall, offering a valuable tool for situations where achieving a trade-off between limiting false positives (precision) and decreasing false negatives (recall) is necessary. A high F1-score signifies the model's proficiency in achieving both high precision and high recall, hence indicating its effectiveness in accurately predicting positive cases while catching a significant proportion of the actual positive instances. The F1-score is a powerful measure for evaluating models in instances when there is a need to carefully control the trade-off between precision and recall.

The F1-score is frequently used by researchers as well as professionals in various domains, including text classification, medical diagnostics, anomaly detection, and fraud detection. These fields often encounter imbalanced datasets, where the occurrence of positive cases is significantly lower than negative cases. In such scenarios, accurately identifying positive cases and correctly classifying negative cases are of the highest priority due to the potential consequences associated with missing positive cases or misclassifying negative cases.

3.2.3 Confusion Matrix

The confusion matrix, additionally referred to as an error matrix, holds significant importance in the domains of machine learning and statistics since it serves as a key tool for assessing the effectiveness of classification algorithms. The organized approach offered by this method allows for the concise summarization of outcomes in both binary and multiclass classification tasks. The confusion matrix is commonly represented as a square matrix with dimensions N*N, where N represents the total number of classes involved in the classification task. In the context of binary classification, the matrix typically takes the form of a 2*2 matrix, encompassing the following elements:

- True Positives (TP) refers to the count of instances that have been accurately classified as belonging to the positive class.
- True Negatives (TN) refer to the count of occurrences that have been accurately classified as belonging to the negative class.
- False positives (FP) refer to the quantity of occurrences that are inaccurately classified as positive, although really belonging to the negative class. This is commonly known as a Type I mistake.
- False Negatives (FN) refer to the quantity of instances that are inaccurately classified as negative, although really belonging to the positive class. This type of error is commonly known as a Type II error.

In the context of multiclass classification, the confusion matrix is a square matrix of size N*N. Each row of the matrix corresponds to the examples that have been predicted to belong to a certain class, while each column represents the occurrences that really belong to a specific class. In this particular scenario, the matrix contains values that correspond to various classification results, including true positives, true negatives, false positives, and false negatives, for each individual class.

The confusion matrix holds significance in the field of NLP due to its many roles in assessing the efficacy of classification models in tasks involving textual data. The confusion matrix holds significant value in the field of NLP due to several reasons.

- **Imbalanced Data:** Imbalanced datasets, characterized by a substantial disparity in class distribution, are frequently encountered in the field of NLP. In the context of sentiment analysis, it is possible to encounter a greater proportion of text expressing neutrality as opposed to texts conveying positive or negative attitudes. The utilization of a confusion matrix facilitates the evaluation of a model's performance in relation to both the majority and minority classes, hence offering valuable insights into its efficacy in managing unbalanced data.
- **Multiclass Classification:** NLP tasks frequently entail the categorization of textual data into various labels or categories, encompassing subject classification, named entity identification, and part-of-speech tagging. In these instances, the utilization of a confusion matrix is crucial for assessing the efficacy of the model in relation to each class, enabling the identification of areas of proficiency and areas of difficulty.

- **Model Comparison:** In the field of NLP, it is common practice to evaluate and compare several models or algorithms in order to ascertain their relative performance for a certain job. The utilization of a confusion matrix offers a systematic approach for evaluating and contrasting models, hence facilitating the identification of the best appropriate model.

- **Error Analysis:** NLP models have the potential to exhibit a range of mistakes, including but not limited to the confusion of synonyms, the inability to accurately identify entities, and the misclassification of sentiment within a given context. The confusion matrix facilitates the analysis of these mistakes, enabling a comprehensive understanding of the prevalence of different types of misclassifications and highlighting areas that require improvement.

- **Fine-Tuning and Iterative Development:** The utilization of a confusion matrix facilitates the iterative refinement process of NLP models. This process assists in the identification and prioritization of certain concerns that require attention, hence enabling gradual enhancements to the model over a period of time.

In essence, the confusion matrix holds significant importance in the field of NLP as it serves as a crucial instrument for evaluating and enhancing the efficacy of text categorization models. It facilitates comprehension and resolution of particular obstacles and demands in NLP activities, hence enabling more efficient creation and implementation of models across diverse applications.

3.3 Analysis

In this study, we employed the LSTM and BERT models with the TACRED dataset for the purpose of relation extraction. In this part, we will discuss general comparisons of LSTM and BERT.

Model Architecture:

- LSTM is a specific variant of RNN architecture that has been specifically developed to effectively model and capture long-term dependencies present in sequential data. The system sequentially analyzes input sequences, while simultaneously utilizing a memory cell capable of storing and retrieving information across significant spatial intervals.

- BERT is a transformer-based model that effectively captures contextual information by taking into account the complete input sequence in both forward and backward directions. Attention processes are employed in order to assign weights to various words within the input sequence, hence enabling the model to effectively capture intricate connections.

Contextual Information:

- LSTM models are designed to incorporate contextual information by processing data sequentially. However, they may encounter challenges in properly capturing relationships that span across long distances. Hidden states are utilized as a means of preserving information from preceding phases.
- BERT has exceptional proficiency in gathering contextual information across the full input sequence. This capability empowers BERT to comprehend the intricate connections between words in a more comprehensive manner. Through the training process, it takes into account both the left and right contexts.

Pretraining:

- LSTM models are commonly trained either from the beginning or started using pre-trained word embeddings. However, they do not exhibit any advantages when subjected to extensive pretraining on large language corpora.
- The BERT model undergoes unsupervised pretraining on a large corpus of text data prior to being fine-tuned for specific tasks. The process of pretraining BERT on a wide variety of language problems enables it to effectively capture intricate contextual representations.

Training Efficiency:

- LSTM models may necessitate more computational time, particularly when dealing with extensive datasets, due to their sequential processing nature.
- The process of training BERT from the beginning can require significant computer resources, but fine-tuning BERT for specific tasks is a more efficient approach that takes advantage of the pretrained model's prior training on extensive language datasets.

Performance on Relation Extraction:

- LSTM models may have challenges in effectively capturing complex relationships as a result of their inherent limits in contextual comprehension.
- BERT demonstrates exceptional performance in tasks related to connection extraction, particularly in cases where the relationships are contextually sensitive and need a comprehensive grasp of the full phrase.

Handling Out-of-Vocabulary Words:

- LSTM model necessitates a pre-established vocabulary, and the presence of terms outside of this vocabulary might provide difficulties.
- BERT has the ability to handle out-of-vocabulary terms to a certain degree by utilizing subword tokenization, which involves breaking words down into smaller subword components.

Interpretability:

- LSTM models have a certain degree of interpretability through their hidden states; yet, comprehending the decision-making process of these models can be a formidable task.
- The challenge of interpretability comes from the intricate attention processes employed by BERT, making it difficult to determine the specific contributions of different segments within the input sequence to the resulting output.

In conclusion, it can be seen that although LSTM models have been extensively employed for sequential tasks, the advanced capability of BERT to comprehend and incorporate comprehensive contextual information has resulted in notable enhancements across diverse natural language processing tasks, such as relation extraction. The selection between LSTM and BERT is contingent upon several aspects, including the intricacy of linkages within the data, the computing resources at hand, and the magnitude of the dataset. BERT, due to its extensive pretraining on vast corpora and its ability to comprehend bidirectional context, frequently exhibits superior performance compared to LSTMs in tasks that demand complex contextual comprehension.

3.3.1 Comparative Analysis of LSTM and BERT Models

Relation extraction, a fundamental task in natural language processing, aims to identify and classify relationships between entities in text. In our approach, we initially developed a traditional LSTM model, followed by a BERT model based on transformers. LSTM model, which falls under the category of recurrent neural networks, operates by sequentially processing input sequences and utilizes hidden states to record relationships over time. In contrast, BERT, a model based on transformers, incorporates bidirectional context throughout the whole sequence, enabling it to comprehend more intricate relationships by taking into account the interplay between words in both forward and backward directions. LSTM model is commonly trained either from the beginning or started with pre-existing word embeddings. BERT, on the other hand, engages in pretraining using huge quantities of unlabeled textual input, which allows it to acquire comprehensive contextual representations. The process of pretraining BERT on a variety of linguistic tasks serves as a solid foundation for subsequent tasks such as relation extraction. The scores obtained from the model output are shown in the Table 3.1. The results of our comparison research demonstrate that BERT, with its transformer architecture and bidirectional contextual awareness, presents notable benefits in relation to standard LSTM for the purpose of relation extraction. When deciding between the two models, it is important to take into account several criteria like the difficulty of the problem, the computing resources available, and the size of the dataset. However, BERT's exceptional performance and versatility make it a very attractive option for relation extraction tasks in the field of NLP. The BERT model showed better results compared to the LSTM model. We conducted an examination of both models using a confusion matrix and then examined the sentences examples to identify more closely related relationship types.

Model	**Precision**	**Recall**	**F1-score**
LSTM	62.13%	48.66%	54.57%
BERT	72.44%	69.41%	70.89%

Table 3.1: LSTM and BERT model results

The confusion matrix of the LSTM model allows the review of the model's errors. Figure 3.4 shows the analysis of "no relation" type details from a confusion matrix, which were subsequently examined in comparison to other relation types.

A common error in LSTM models occurs when the real label is classified as "no relation," while the predicted label is classified as "org:top_members/employees.", figure 3.6 shows this error correlation and figure 3.5 shows example from dataset.

LSTM models occurs error when the real label is classified as "no relation" while the predicted label is classified as "per:title", figure 3.8 shows this error correlation and figure 3.7 shows example from dataset.

LSTM models occurs error when the real label is classified as "per:cities_of_residence" while the predicted label is classified as "no relation", figure 3.10 shows this error correlation and figure 3.9 shows example from dataset.

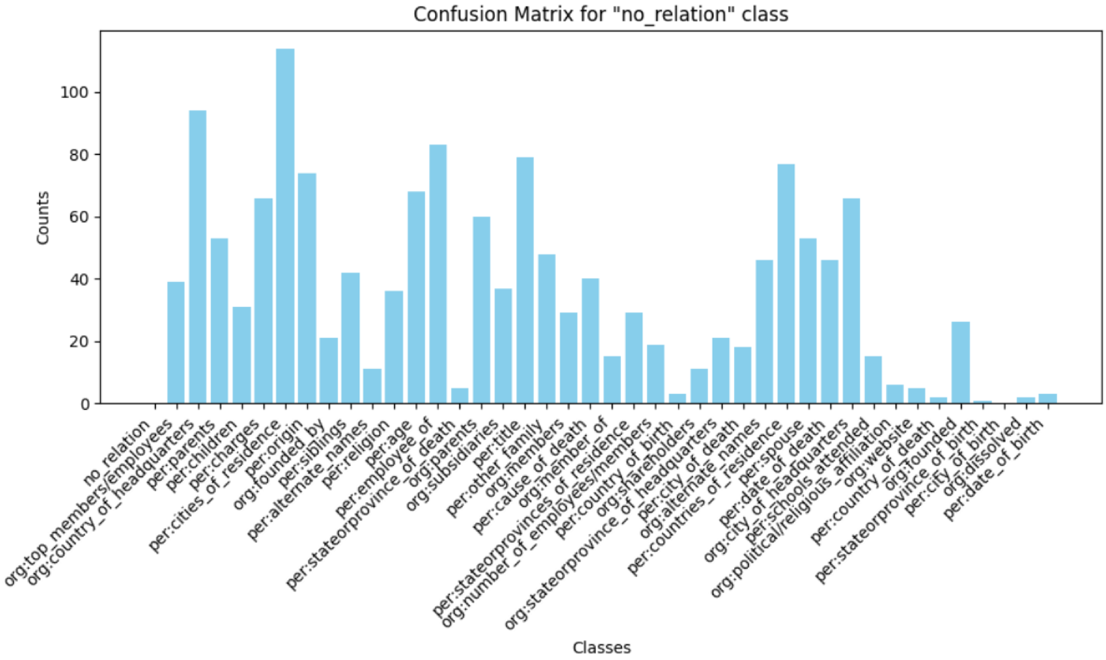
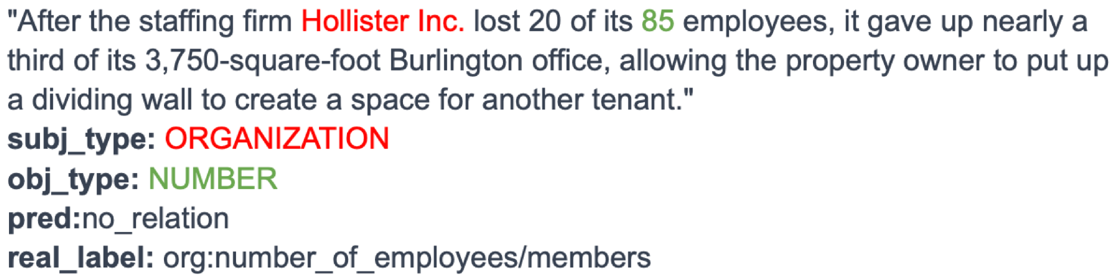


Figure 3.4: "no relation" type error analysis



"After the staffing firm Hollister Inc. lost 20 of its 85 employees, it gave up nearly a third of its 3,750-square-foot Burlington office, allowing the property owner to put up a dividing wall to create a space for another tenant."
subj_type: ORGANIZATION
obj_type: NUMBER
pred:no_relation
real_label: org:number_of_employees/members

Figure 3.5: Sentence example for "no relation" "org:top_members/employees" error

The similar analytical approach was conducted for the BERT model as well. The confusion matrix of the BERT model allows the review of the model's errors. Figure 3.11 shows the analysis of "no relation" type details from a confusion matrix, which were subsequently examined in comparison to other relation types.

BERT models occurs error when the real label is classified as "per:title" while the predicted label is classified as "org:founded_by", figure 3.13 shows this error correlation and figure 3.12 shows example from dataset.

BERT models occurs error when the real label is classified as "per:employee_of" while the predicted label is classified as "org:alternate_names", figure 3.15 shows

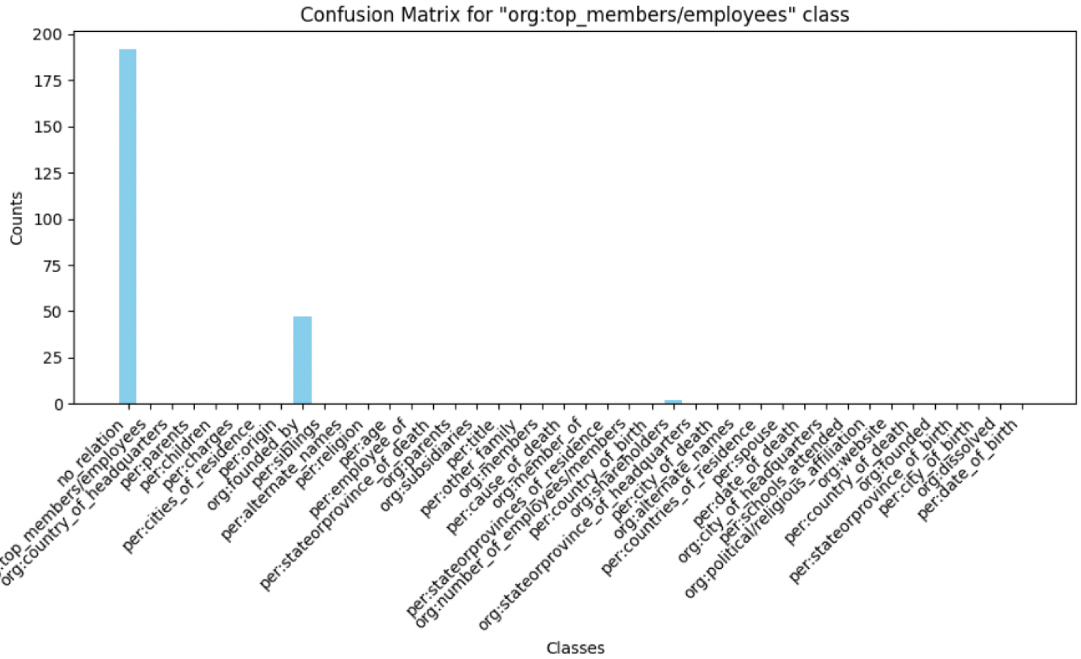
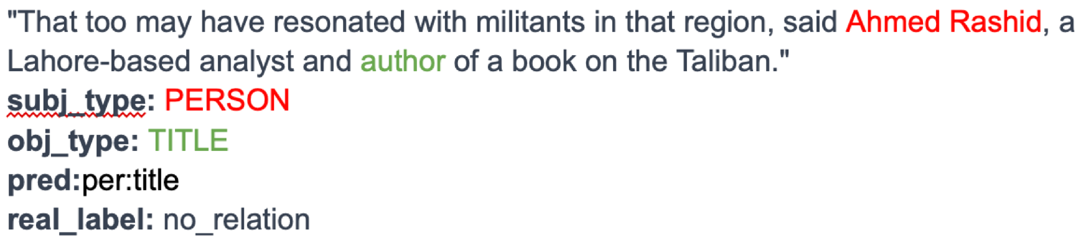


Figure 3.6: Error from the true label is "no relation" but the prediction label is "org:top_members/employees"



"That too may have resonated with militants in that region, said Ahmed Rashid, a Lahore-based analyst and author of a book on the Taliban."
subj_type: PERSON
obj_type: TITLE
pred:per:title
real_label: no_relation

Figure 3.7: Sentence example for "no relation" - "per:title" error

this error correlation and figure 3.14 shows example from dataset.

BERT models occurs error when the real label is classified as "org:alternate_names" while the predicted label is classified as "per:children", figure 3.17 shows this error correlation and figure 3.16 shows example from dataset.

The BERT model demonstrates better results compared to the LSTM model in related prediction tasks. However, it is important to note that both models exhibit errors in their predictions. One of the major issues seen in this study is the presence of errors in the dataset, where the same sentences are labeled differently in various versions. Figure 3.18 shows an example for same sentence with different labels from TACRED dataset. This inconsistency in labeling poses a challenge for both models, leading to confusion.

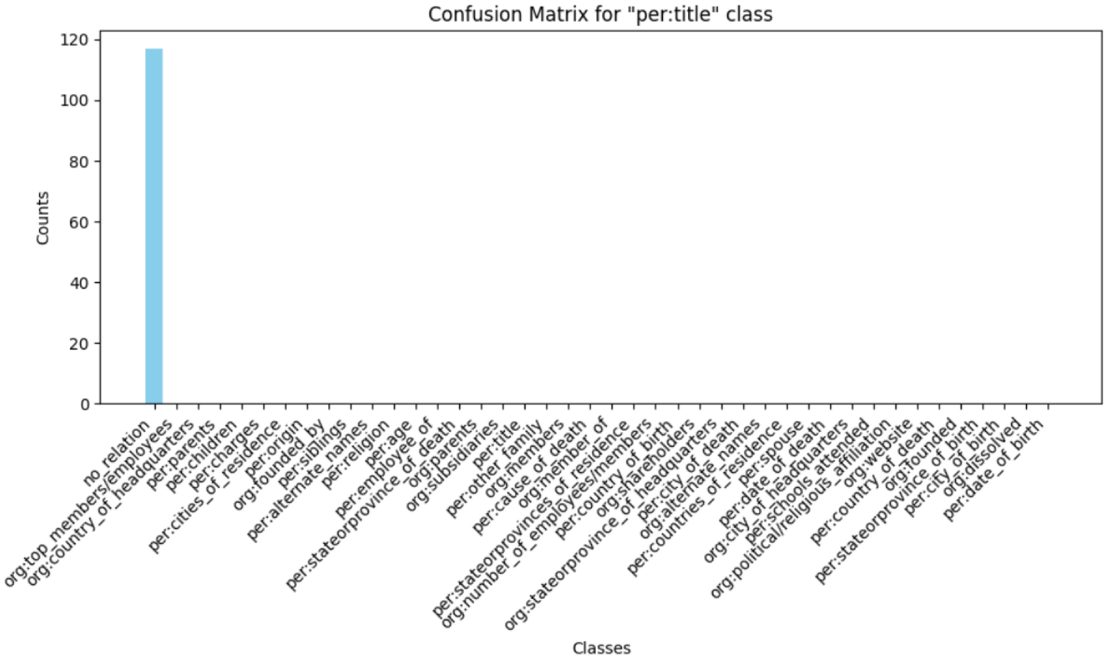
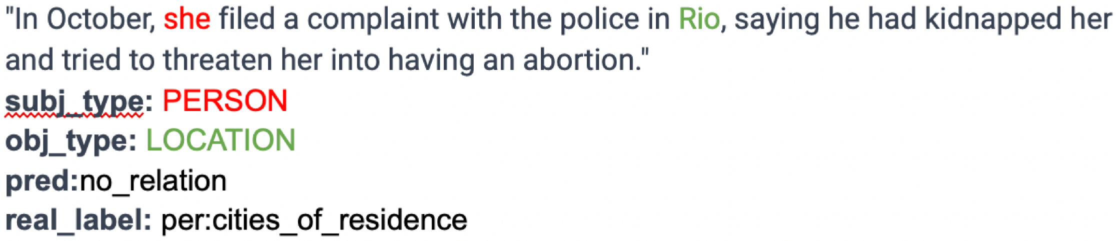


Figure 3.8: Error from the true label is "no relation" but the prediction label is "per:title"



"In October, she filed a complaint with the police in Rio, saying he had kidnapped her and tried to threaten her into having an abortion."

subj_type: PERSON

obj_type: LOCATION

pred:no_relation

real_label: per:cities_of_residence

Figure 3.9: Sentence example for "per:cities_of_residence" - "no relation" error

Another issue occurs from the insufficient quantity of data available for certain relation types, which poses a challenge for training the model. This lack of data further complicates the performance of the models. In the meantime, the TACRED dataset exhibits imbalanced data, with no connection type accounting for 79.5% of the dataset. This lack of coverage for a certain relation type leads to confusion when dealing with other relation types.

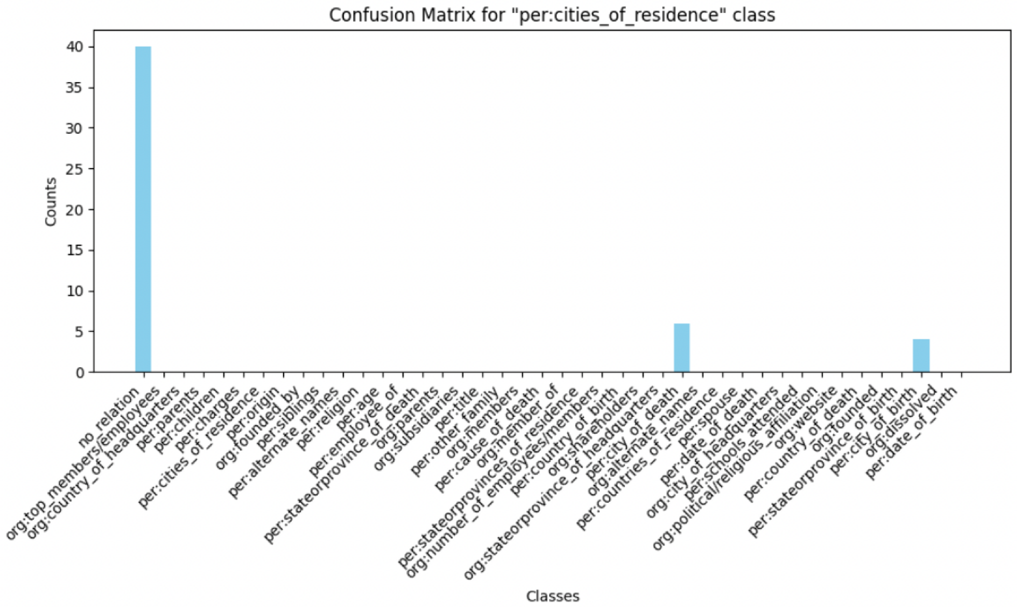


Figure 3.10: Error from the true label is "per:cities_of_residence" but the prediction label is "no relation"

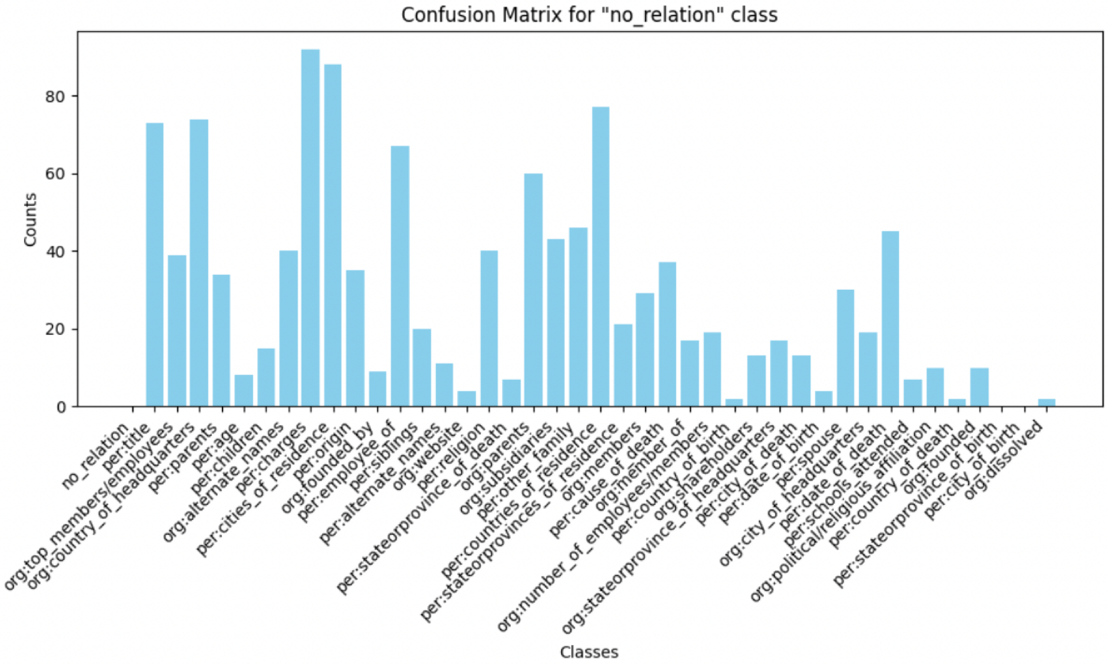
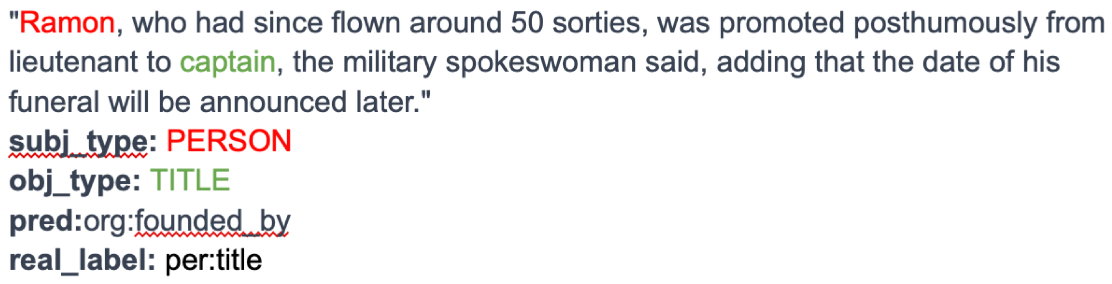


Figure 3.11: "no relation" type error analysis



"Ramon, who had since flown around 50 sorties, was promoted posthumously from lieutenant to captain, the military spokeswoman said, adding that the date of his funeral will be announced later."

subj_type: PERSON

obj_type: TITLE

pred:org:founded_by

real_label: per:title

Figure 3.12: Sentence example for "per:title" - "org:founded_by" error

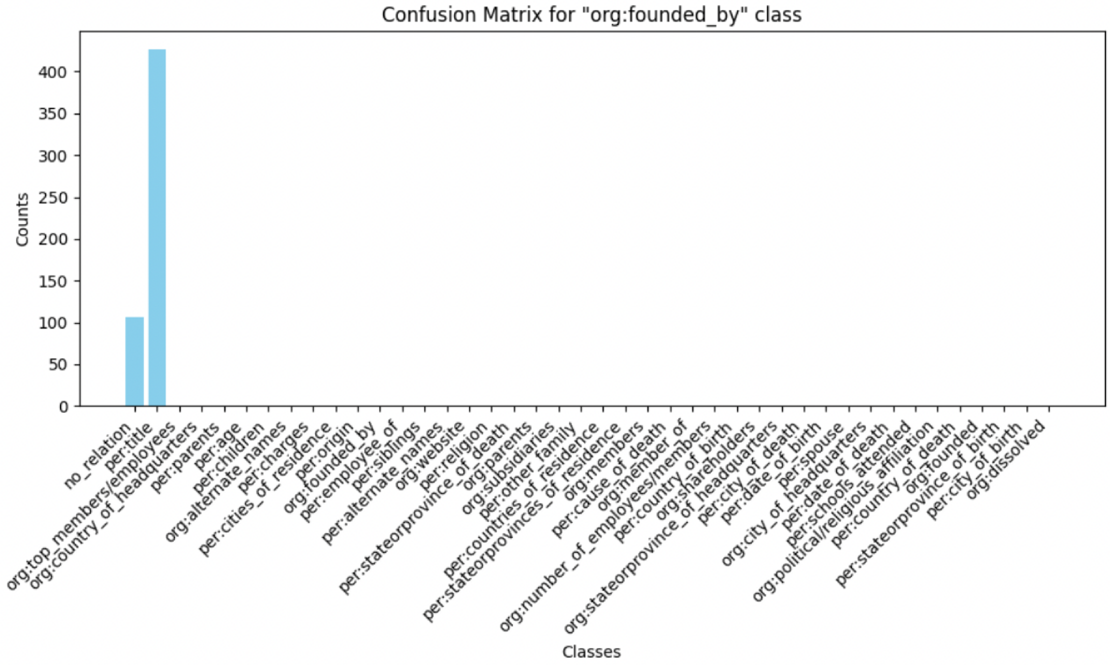
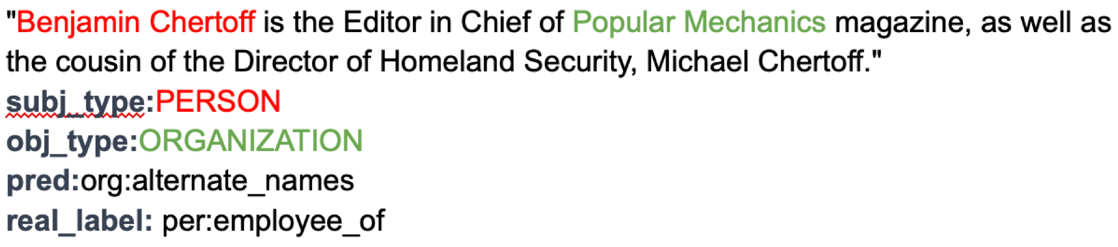


Figure 3.13: Error from the true label is "per:title" but the prediction label is "org:founded_by"



"Benjamin Chertoff is the Editor in Chief of Popular Mechanics magazine, as well as the cousin of the Director of Homeland Security, Michael Chertoff."

subj_type:PERSON

obj_type:ORGANIZATION

pred:org:alternate_names

real_label: per:employee_of

Figure 3.14: Sentence example for "per:employee_of" - "org:alternate_names" error

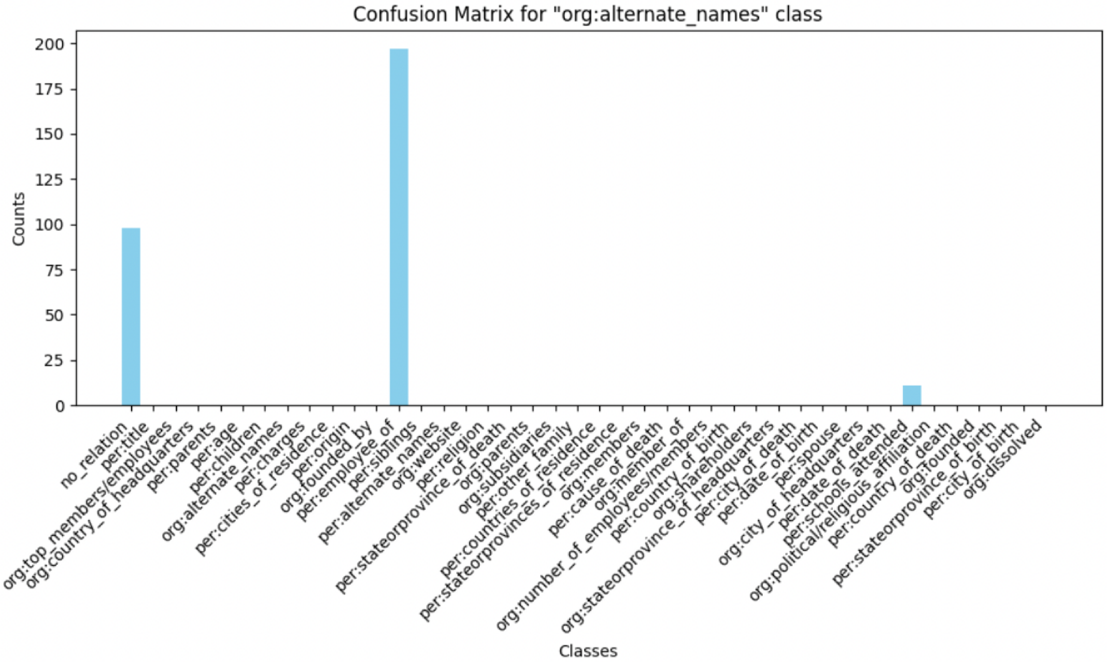


Figure 3.15: Error from the true label is "per:employee_of" but the prediction label is "org:alternate_names"

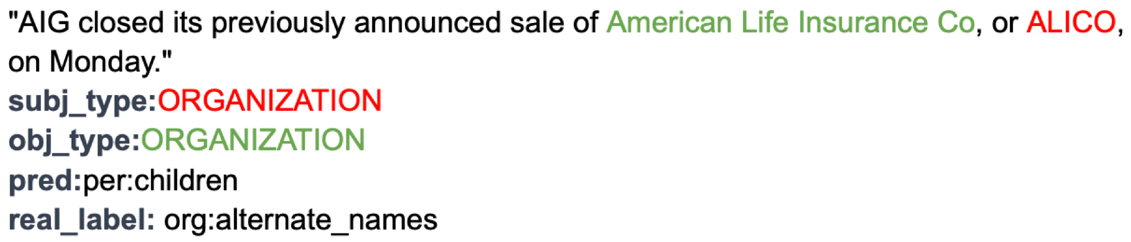

"AIG closed its previously announced sale of American Life Insurance Co, or ALICO, on Monday."
subj_type:ORGANIZATION
obj_type:ORGANIZATION
pred:per:children
real_label: org:alternate_names

Figure 3.16: Sentence example for "org:alternate_names" - "per:children" error

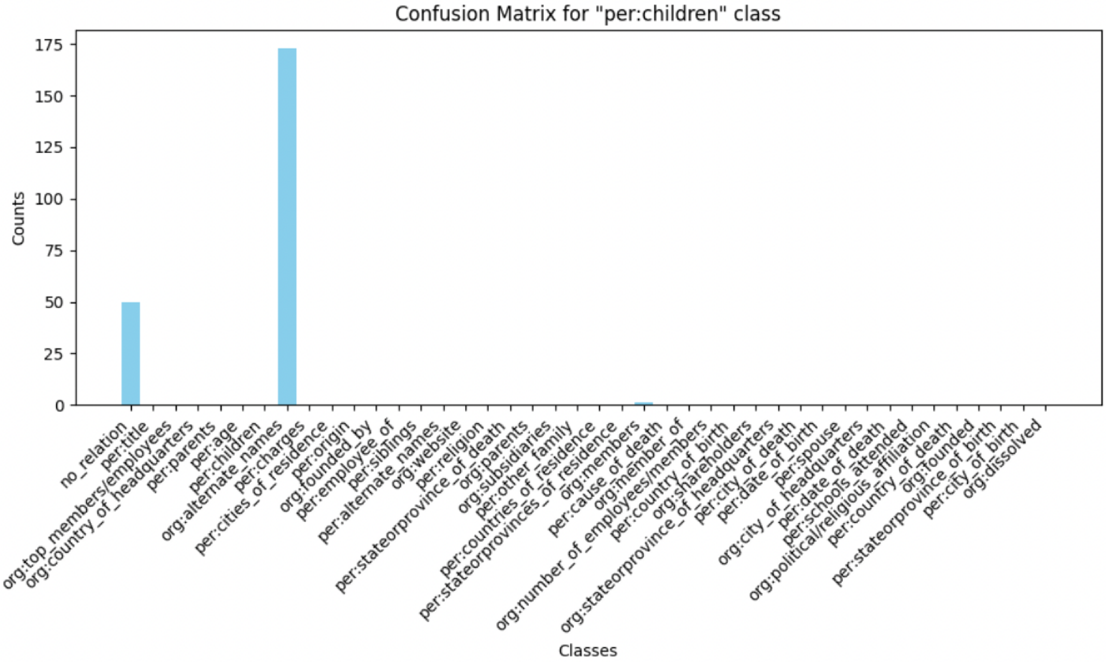
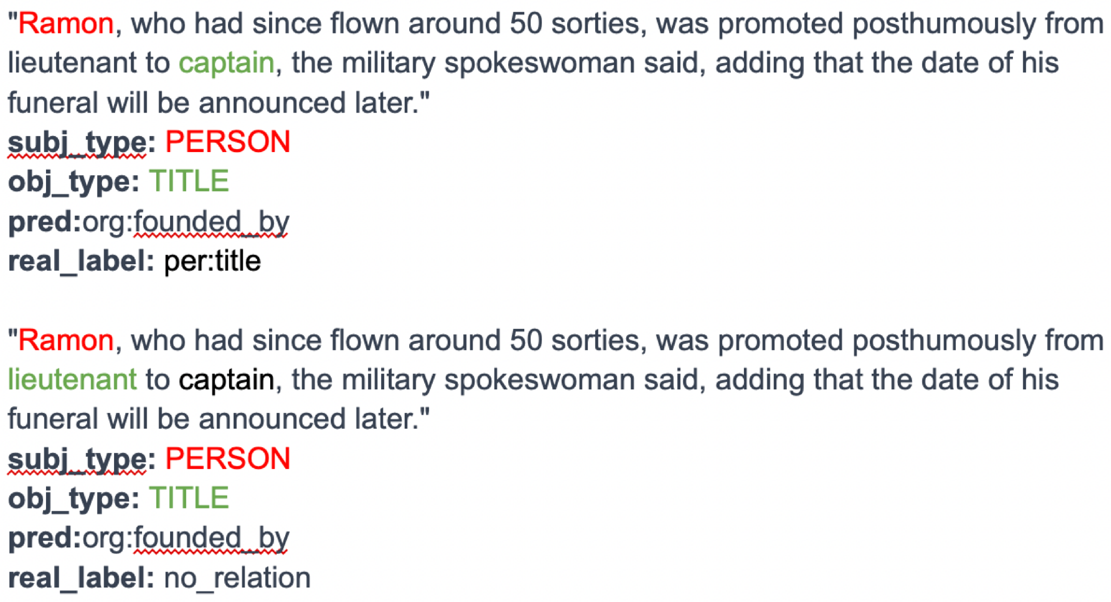


Figure 3.17: Error from the true label is "org:alternate_names" but the prediction label is "per:children"



"Ramon, who had since flown around 50 sorties, was promoted posthumously from lieutenant to captain, the military spokeswoman said, adding that the date of his funeral will be announced later."
subj_type: PERSON
obj_type: TITLE
pred:org:founded_by
real_label: per:title

"Ramon, who had since flown around 50 sorties, was promoted posthumously from lieutenant to captain, the military spokeswoman said, adding that the date of his funeral will be announced later."
subj_type: PERSON
obj_type: TITLE
pred:org:founded_by
real_label: no_relation

Figure 3.18: TACRED dataset same sentence with different labels

4

CONCLUSIONS AND FUTURE WORKS

Relation extraction takes a significant role in the domain of NLP, since it is essential for extracting valuable insights from unstructured textual data. The primary objective of relation extraction is to identify and extract meaningful semantic associations or links between entities referenced in a given text. This study delved into the intricate domain of sentence-level and document-level relation extraction within the realm of NLP applications. The research part, encompassed various approaches, extending to relation extraction, with the overarching aim of constructing a robust model tailored for this specific purpose.

The evaluation also revolved around the effectiveness of current approaches for extracting relations at the sentence level, with a particular emphasis on utilizing the TACRED dataset as a fundamental source. Significantly, the incorporation of the BERT model played a crucial impact, demonstrating its remarkable powers in improving the identification of connections between things included inside sentences. Simultaneously, we evaluated the efficacy of the LSTM model using the identical TACRED dataset, enabling a comparative examination to uncover the inherent advantages and disadvantages of each methodology. The insights derived from this project offer a nuanced understanding of how BERT and LSTM models perform in the specific context of sentence-level relation extraction. By meticulously examining their precision in extracting relationships, we have contributed valuable information to the ongoing discourse surrounding the relative advantages and disadvantages of these two prominent models.

In addition, in order to contextualize what we discovered within the wider scope of scholarly research and progress in this particular domain, our study included an extensive examination of existing literature and research articles pertaining to methodologies for extracting connections at both the sentence and document levels. The inclusion of these additional sources not only enhanced the comprehensiveness of our investigation, but also functioned as reference points for comparative analysis, offering significant methodological perspectives. By leveraging cutting-edge technologies and powerful NLP models, our outcomes can be significantly enhanced.

In conclusion, the domain of NLP is now seeing a notable surge in technical progress, and various research regularly highlight the capacity for ongoing enhancements. The dynamic environment, characterized by the emergence of novel models and technologies, contributes a substantial aspect to this advancement. The utilization of various technologies and techniques contributes to the enhancement of our understanding of language processing and also presents new opportunities for innovative applications and progress. With the emergence of advanced models and technologies, there is a growing emphasis on collaborative efforts to enhance NLP approaches in diverse study domains. These initiatives are expected to have a significant impact on obtaining improved outcomes and the development of more complex and efficient language processing systems.

References

[1] Shudong Huang Alexis Mitchell Stephanie Strassel and Ramez Zakhary. "ACE 2004 Multilingual Training Corpus". In: Linguistic Data Consortium (LDC). 2004.

[2] Xiang Chen et al. "KnowPrompt: Knowledge-Aware Prompt-Tuning with Synergistic Optimization for Relation Extraction". In: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 2778–2788. ISBN: 9781450390965. DOI: 10.1145/3485447.3511998. URL: https://doi.org/10.1145/3485447.3511998.

[3] Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. "Relation Classification as Two-way Span-Prediction". In: 2021. arXiv: 2010.04829 [cs.CL].

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *North American Chapter of the Association for Computational Linguistics*. 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[5] Houssem Gasmi, Jannik Laval, and Abdelaziz Bouras. "Information extraction of cybersecurity concepts: An LSTM approach". In: vol. 9. 19. MDPI, 2019, p. 3945.

[6] Xu Han et al. "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*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4803–4809. DOI: 10.18653/v1/D18-1514. URL: https://aclanthology.org/D18-1514.

[7] Shashank Hebbar and Ying Xie. "CovidBERT-biomedical relation extraction for Covid-19". In: *The International FLAIRS Conference Proceedings*. Vol. 34. 2021.

[8] Iris Hendrickx et al. "SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 33–38. URL: https://www.aclweb.org/anthology/S10-1006.

[9] Lars Hillebrand et al. "Kpi-bert: A joint named entity recognition and relation extraction model for financial reports". In: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE. 2022, pp. 606–612.

[10] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: vol. 9. Dec. 1997, pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[11] Bin Ji et al. "Span-based Joint Entity and Relation Extraction with Attention-based Span-specific and Contextual Semantic Representations". In: Jan. 2020, pp. 88–99. DOI: 10.18653/v1/2020.coling-main.8.

[12] Ren Li et al. "Joint extraction of entities and relations via an entity correlated attention neural model". In: vol. 581. 2021, pp. 179–193. DOI: https://doi.org/10.1016/j.ins.2021.09.028. URL: https://www.sciencedirect.com/science/article/pii/S0020025521009592.

[13] SiLiang Li, Bin Xu, and Tong Lee Chung. "Definition extraction with lstm recurrent neural networks". In: *International Symposium on Natural Language Processing Based on Naturally Annotated Big Data*. Springer. 2016, pp. 177–189.

[14] Tianyu Liu et al. "Autoregressive Structured Prediction with Language Models". In: 2022. arXiv: 2210.14698 [cs.CL].

[15] Yang Liu and Mirella Lapata. "Learning Structured Text Representations". In: vol. 6. 2017, pp. 63–75. URL: https://api.semanticscholar.org/CorpusID:39871772.

[16] Shengfei Lyu and Huanhuan Chen. "Relation Classification with Entity Type Restriction". In: Jan. 2021, pp. 390–395. DOI: 10.18653/v1/2021.findings-acl.34.

[17] Kai Ma et al. “Ontology-Based BERT Model for Automated Information Extraction from Geological Hazard Reports”. In: vol. 34. 5. Springer, 2023, pp. 1390–1405.

[18] Arjun Magge, Matthew Scotch, and Graciela Gonzalez-Hernandez. “Clinical NER and relation extraction using bi-char-LSTMs and random forest classifiers”. In: *International workshop on medication and adverse drug event detection*. PMLR. 2018, pp. 25–30.

[19] Christopher Manning et al. “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 55–60. DOI: 10.3115/v1/P14-5010. URL: https://aclanthology.org/P14-5010.

[20] Hao Peng et al. “Learning from Context or Names? An Empirical Study on Neural Relation Extraction”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 3661–3672. DOI: 10.18653/v1/2020.emnlp-main.298. URL: https://aclanthology.org/2020.emnlp-main.298.

[21] Arpita Roy and Shimei Pan. “Incorporating medical knowledge in BERT for clinical relation extraction”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 5357–5366. DOI: 10.18653/v1/2021.emnlp-main.435. URL: https://aclanthology.org/2021.emnlp-main.435.

[22] Injy Sarhan and Marco Spruit. “Can we survive without labelled data in NLP? Transfer learning for open information extraction”. In: vol. 10. 17. MDPI, 2020, p. 5758.

[23] Syed Afeef Ahmed Shah, Muhammad Ali Masood, and Amanullah Yasin. “Dark Web: E-Commerce Information Extraction Based on Name Entity Recognition Using Bidirectional-LSTM”. In: vol. 10. IEEE, 2022, pp. 99633–99645.

[24] George Stoica, Emmanouil Antonios Platanios, and Barnab'as P'oczos. "Re-TACRED: Addressing Shortcomings of the TACRED Dataset". In: *AAAI Conference on Artificial Intelligence*. 2021. URL: https://api.semanticscholar.org/CorpusID:233296843.

[25] Peng Su and K Vijay-Shanker. "Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction". In: vol. 23. 1. Springer, 2022, p. 120.

[26] Qingyu Tan et al. "Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022. DOI: 10.18653/v1/2022.findings-acl.132. URL: https://aclanthology.org/2022.findings-acl.132.

[27] Qingyu Tan et al. "Revisiting DocRED Addressing the False Negative Problem in Relation Extraction". In: *Proceedings of EMNLP*. 2022. URL: https://arxiv.org/abs/2205.12696.

[28] Christopher Walker et al. "ACE 2005 multilingual training corpus. Linguistic Data Consortium." In: 2006. URL: https://catalog.ldc.upenn.edu/LDC2006T06.

[29] Chenguang Wang et al. "DeepStruct: Pretraining of Language Models for Structure Prediction". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 803–823. DOI: 10.18653/v1/2022.findings-acl.67. URL: https://aclanthology.org/2022.findings-acl.67.

[30] Huiyu Wang et al. "Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation". In: 2020. arXiv: 2003.07853 [cs.CV].

[31] Yuxin Xiao et al. "SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 2395–2409. DOI:

10.18653/v1/2022.naacl-main.171. URL: https://aclanthology.org/2022.naacl-main.171.

[32] Benfeng Xu et al. "Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction". In: 2021. arXiv: 2102.10249 [cs.CL].

[33] Kui Xue et al. "Fine-tuning BERT for joint entity and relation extraction in Chinese medical text". In: *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2019, pp. 892–897.

[34] Yuan Yao et al. "DocRED: A Large-Scale Document-Level Relation Extraction Dataset". In: *Proceedings of ACL 2019*. 2019.

[35] Ningyu Zhang et al. "Document-level Relation Extraction as Semantic Segmentation". In: 2021. arXiv: 2106.03618 [cs.CL].

[36] Yuhao Zhang et al. "Position-aware Attention and Supervised Data Improve Slot Filling". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. 2017, pp. 35–45. URL: https://nlp.stanford.edu/pubs/zhang2017tacred.pdf.

[37] Zhengyan Zhang et al. "ERNIE: Enhanced Language Representation with Informative Entities". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Llus Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1441–1451. DOI: 10.18653/v1/P19-1139. URL: https://aclanthology.org/P19-1139.

[38] Wenxuan Zhou et al. "Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling". In: 2020. arXiv: 2010.11304 [cs.CL].