DATASETS AND TRANSFORMER MODELS FOR CROSS-LINGUAL RELATION CLASSIFICATION

by

Abdullatif Köksal

B.S., Computer Engineering, Boğaziçi University, 2018

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2021

ACKNOWLEDGEMENTS

First and foremost, I would like to express my profound gratitude to my advisor, Assoc. Prof. Arzucan Özgür for her continuous guidance and encouragement during my studies. I have started my NLP journey in sophomore year with her support. Since then, she has always motivated and guided me, and I greatly appreciate her efforts, understanding, and dedicated involvement.

I am deeply thankful to all professors at Boğaziçi University, especially to Assoc. Prof. Elif Özkırımlı, Prof. Tunga Güngor, Prof. Taylan Cemgil, and Prof. Lale Akarun for their assistance and support. I also would like to thank Assist. Prof. Reyyan Yeniterzi for being on my thesis committee. I would like to express my gratitude to my colleagues in TABILAB and AILAB for their wonderful collaboration and friendship.

Further, I am also greatful to Ahmed Yusuf Asiltürk, Barbaros Eriş, Gizem Özkanal, Ramazan Pala, and Taha Küçükkatırcı for their contributions to the translation process and El Turco company for their quality check service in this work.

I would like to show my gratitude and appreciation to my family and my wife, Esmanur Yılmaz Köksal, for their endless support through my academic journey.

This work is supported by TÜBITAK-BIDEB 2210-A Scholarship Program (TÜBİTAK BİDEB 2210-A Genel Yurt İçi Yüksek Lisans Burs Programı), and TÜBİTAK is gratefully acknowledged. This work is also supported by Boğaziçi University Research Fund under the Grant Number 16903.

ABSTRACT

DATASETS AND TRANSFORMER MODELS FOR CROSS-LINGUAL RELATION CLASSIFICATION

Relation classification is one of the key topics in information extraction, which can be used to construct knowledge bases or to provide useful information for question answering. Current approaches for relation classification are mainly focused on the English language and require lots of training data with human annotations. Creating and annotating a large amount of training data for low-resource languages is impractical and expensive. To overcome this issue, we propose two cross-lingual relation classification models: a baseline model based on Multilingual BERT (mBERT) and a new multilingual pretraining setup called Matching the Multilingual Blanks (MTMB), which significantly improves the baseline with distant supervision. For evaluation, we introduce a new public benchmark dataset for cross-lingual relation classification in English, French, German, Spanish, and Turkish, called RELX. We also provide the RELX-Distant dataset, which includes hundreds of thousands of sentences with relations from Wikipedia and Wikidata collected by distant supervision for these languages. We observe that MTMB significantly outperforms the mBERT baseline in presented languages by 2.14% absolute improvement of F1-score on average. We further investigate MTMB's effectiveness in low-resource settings, and when 10% of the training data is used, 10.58% absolute improvement of F1-score on average over mBERT is observed.

ÖZET

ÇAPRAZ DİLLİ İLİŞKİ SINIFLANDIRMASI İÇİN DÖNÜŞTÜRÜCÜ MODELLERİ VE VERİ KÜMELERİ

İlişki sınıflandırması, bilgi tabanları oluşturmak ve soru cevaplama sistemleri için faydalı bilgiler sağlamak için kullanılabilen bilgi çıkarımındaki önemli konulardan biridir. İlişki sınıflandırmasındaki mevcut yaklaşımlar, temel olarak İngilizce dilinde gerçekleşmektir ve çok sayıda işaretli eğitim verisi gerektirir. Az kaynaklı diller için bu miktarda işaretli eğitim verisi oluşturmak pratik değildir ve yüksek maliyetlidir. Bu sorunun üstesinden gelmek için iki farklı çapraz dilli ilişki sınıflandırma modeli öneriyoruz: Çok Dilli BERT'e (mBERT) dayalı temel bir model ve temel modeli önemli ölçüde iyileştiren Çok Dilli Boşlukları Eşleştirme (MTMB) adını verdiğimiz, uzak denetim kullanılarak özgün bir ön eğitim aşamasına sahip olan çok dilli bir dönüştürücü modeli. Çapraz dilli ilişki sınıflandırması için RELX adını verdiğimiz, İngilizce, Fransızca, Almanca, İspanyolca ve Türkçe dillerinden verilere sahip olan yeni bir değerlendirme veri seti sunuyoruz. Ayrıca, bu diller için Wikipedia ve Wikidata'dan uzak denetim yöntemiyle toplanan yüz binlerce cümle içeren RELX-Distant ilişki sınıflandırma veri kümesini de sağlıyoruz. Sonuç olarak çapraz dilli ilişki sınıflandırmasında MTMB'nin mBERT temel modeline göre sunulan dillerde önemli ölçüde daha iyi performans gösterdiğini ve ortalama olarak F1 puanında %2,14 iyileşme sağladığını gözlemliyoruz. Eğitim verisinin %10'unun kullanıldığı az kaynaklı ortamda da MTMB'nin etkinliğinin daha iyi olduğunu ve mBERT'e göre ortalama F1 puanını %10,58 iyileştirdiğini gözlemliyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

D_s	Dataset in the source language
D_t	Dataset in the target language
S_i^s	i_{th} sentence in the source dataset
S_i^t	i_{th} sentence in the target dataset
w_i	i_{th} word in a sentence
$E1_i^s$	First entity in i_{th} sentence in the source dataset
$E1_i^t$	First entity in i_{th} sentence in the target dataset
$E2_i^s$	Second entity in i_{th} sentence in the source dataset
$E2_i^t$	Second entity in i_{th} sentence in the target dataset
R	Fixed relation set
r_i	i_{th} relation in R
D_t	Dataset in the target language
<e1>	Start entity marker for the first entity
</e1>	End entity marker for the first entity
<e2>	Start entity marker for the second entity
</e2>	End entity marker for the second entity

LIST OF ACRONYMS/ABBREVIATIONS

BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BPE	Byte-Pair Encoding
CNN	Convolutional Neural Network
EMNLP	Empirical Methods in Natural Language Processing
GLUE	General Language Understanding Benchmark
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
KB	Knowledge Base
LSTM	Long Short Term Memory
mBERT	Multilingual BERT
MTB	Matching the Blanks
MTMB	Matching the Multilingual Blanks
NLP	Natural Language Processing
RE	Relation Extraction
RNN	Recurrent Neural Network
SOTA	State of the Art
SOV	Subject - Object - Verb
SVO	Subject - Verb - Object
SemEval	Semantic Evaluation
TPU	Tensor Processing Unit
WALS	World Atlas of Language Structures
XLM	Cross-lingual Language Model

1. INTRODUCTION

Extracting useful information from unstructured text is one of the most essential topics in Natural Language Processing (NLP). Relation classification can help to achieve this objective by enabling the automatic construction of knowledge bases and by providing useful information for question answering models [1]. Given an entity pair $(e1, e2)$ and a sentence S that contains these entities, the goal of relation classification is to predict the relation $r \in R$ between $e1$ and $e2$ from a set of predefined relations, which may include 'no relation' as well. For example, with the help of relation classification, we can create semantic triples such as *(Rocky Mountain High School, founded, 1973)* from a sentence like *"Rocky Mountain High School opened at its current location in 1973 and was expanded in 1994."*, where 'Rocky Mountain High School' and '1973' are the given entities and 'founded' is the relation between them based on this sample sentence.

Traditionally, relation classification methods rely on hand-crafted features [2]. Lately, pretrained word embeddings [3] with RNN-LSTM architecture [4,5] or transformers-based models [6] have gained more attention in this domain. Recent works on relation classification have usually focused on English, even though non-English content on the web is around 40% [7] and the number of multilingual text-corpora is increasing [8]. These supervised approaches for relation classification are not easily adaptable to other languages, since they require large annotated training datasets, which are both costly and time-consuming to create.

The challenge of creating manually labeled training datasets for different languages can be alleviated through cross-lingual NLP approaches. In cross-lingual relation classification, the objective is to predict the relations in a sentence in a target language, while the model is trained with a dataset in a source language, which may be different from the target language. For example, a cross-lingual relation classification model should be able to extract semantic triples such as *(CD Laredo, founded,*

1927) from a Spanish sentence like *"CD Laredo fue fundado en 1927 con el nombre de Sociedad Deportiva Charlestón.*[1] *"* for the given entities 'CD Laredo' and '1927', even when the annotated training data is solely in English.

Thanks to multilingual pretrained transformer models like Multilingual BERT (mBERT) [9] and XLM [10], cross-lingual models have been studied in depth for several NLP tasks such as question answering [11–13], natural language inference [10, 13, 14], and named entity recognition [13].

In this thesis, we first propose a baseline cross-lingual model for relation classification based on the pretrained mBERT model [9]. Then, we introduce an approach called Matching the Multilingual Blanks to improve the relation classification ability of mBERT in different languages with the help of a considerable number of relation pairs collected by distant supervision. Prior works on cross-lingual relation classification use additional resources in the target language such as aligned corpora [15], machine translation systems [16], or bilingual dictionaries [17]. Our mBERT baseline model does not require any additional resources in the target language. The Matching the Multilingual Blanks model improves mBERT by utilizing the already available Wikipedia and Wikidata resources with distant supervision.

We present two new datasets for cross-lingual relation classification, namely RELX and RELX-Distant. RELX has been developed by selecting a subset of the commonly-used KBP-37 English relation classification dataset [4] and generating human translations and annotations in the French, German, Spanish, and Turkish languages. The resulting dataset contains 502 parallel test sentences in five different languages with 37 relation classes. To our knowledge, RELX is the first parallel relation classification dataset, which we believe will serve as a useful benchmark for evaluating cross-lingual relation classification methods.

[1] *English Translation:* CD Laredo was founded in 1927 with the name "Sociedad Deportiva Charlestón".

RELX-Distant is a multilingual relation classification dataset collected from Wikipedia and Wikidata through distant supervision for the aforementioned five languages. We gather from 50 thousand up to 800 thousand sentences, whose entities have been labeled by the editors of Wikipedia. The relations among these entities are extracted from Wikidata.

Our contributions can be summarized as follows:

(i) We introduce the RELX dataset, a novel cross-lingual relation classification benchmark with 502 parallel sentences in English, French, German, Spanish, and Turkish.
(ii) To support distantly supervised models, we introduce the RELX-Distant dataset, which has hundreds of thousands of sentences with relations collected from Wikipedia and Wikidata for the mentioned five languages.
(iii) We first present a baseline mBERT model for cross-lingual relation classification and then, propose a novel multilingual distant supervision approach to improve the model.

The work presented in this thesis is published in EMNLP - Findings 2020, with the title of "The RELX Dataset and Matching the Multilingual Blanks for Cross-Lingual Relation Classification" [18]. This thesis is organized as follows. The related work including cross-lingual NLP, English relation classification, and cross-lingual relation classification is discussed in Chapter 2. Background of relation classification and transformers architecture are discussed in Chapter 3. The datasets, KBP-37, RELX, RELX-Distant, are introduced in Chapter 4. The task definition, our baseline and novel methods are described in Chapter 5. The experimental results for BERT, mBERT, and MTMB transformer models are presented in Chapter 6. Finally, we conclude and discuss future work in Chapter 7.

2. RELATED WORK

Natural Language Processing (NLP) is a domain with an interest in understanding textual and spoken data, aiming at human language understanding. In the early era of NLP, many studies solely focus on linguistic features and extracting rules to understand texts, called symbolic NLP. Until the 1990s, symbolic NLP takes the lead, and several topics are studied such as chatbots [19], word-sense disambiguation [20], and generative grammars [21]. Even though they build strong baselines for several NLP tasks, they are not able to generalize well and need complex hand-crafted rules which require expertise and time.

Statistical NLP approaches have gained popularity to solve generalizability and human-time costs around the 1990s. For example, the interest towards statistical machine translation [22] increases as glossary creation and word-level alignment are achieved thanks to the statistical techniques. Furthermore, many fields such as text classification [23], named entity recognition [24], part-of-speech tagging [25] use statistical models like Hidden Markov models, support vector machines, and maximum entropy by encoding the textual input with the help of machine learning techniques. Neural models follow similar patterns, and recently, architectures and encoding of textual inputs are studied within deep neural models.

Recently, statistical and neural models have attracted attention in NLP, which are generally "data-hungry" approaches. To generalize well on different tasks, lots of annotated training data are required in these models. For example, Soares and coworkers [6] show that in relation classification, it can be achieved only a 43.4% F1 score when 681 samples are used for the training data in the TACRED dataset [26] while 68,120 samples in the training data can achieve a 70.6% F1 score. Collecting such dataset is possible for well-studied and industry-backed tasks in English NLP however it is usually not possible for low-resource languages. Instead of forming a large amount of annotated data for each NLP task in every language, recent works focus on adapting

multilingual or cross-lingual studies which aim to generalize to other languages while the language of the training data is different.

2.1. Cross-lingual NLP

Cross-lingual NLP aims to generalization of a given task with training data of a source language to another target language without giving any annotated samples to the model in the target language. In other words, it is zero-shot learning for the target language with the help of a source language that is different from the target language, and it is usually a high-resource language.

The introduction of dense word embedding models such Word2Vec [3] and Glove [27] enable to learn better word embeddings for languages with a corpus without any annotation. Creating word embeddings for different languages easily steps up efforts in the cross-lingual NLP. By training an additional alignment function between the embedding spaces of the source language and the target language, models trained with the source language generalize to the target language without any annotated data except alignment data. They are used in different NLP tasks such as cross-lingual natural language inference [28] and cross-lingual named entity recognition [29]. The summary for general cross-lingual NLP approaches with word embedding alignment can be seen in Figure 2.1. This figure illustrates cross-lingual sentiment analysis with the word embedding alignment function. First, it learns the alignment function between English and Turkish word embeddings with a small parallel corpus. Second, the model is trained with English sentiment analysis dataset. Finally, the alignment function is combined with a fine-tuned model to predict sentiments for the Turkish language.

Another approach to cross-lingual NLP is exploiting existing machine translation systems. Machine translation systems can perform high-quality performance thanks to the attention mechanism [30]. Even, recent attention-based machine translation models achieve near-human performance in BLEU scores [31]. Even though these models still have drawbacks like missing common-sense reasoning or weakness to spelling errors,

they provide an alternate solution to cross-lingual NLP. Some studies in cross-lingual NLP use existing machine translation systems to translate inputs in the target language to the source language during the inference time [16]. On the other hand, many studies focus on translating existing datasets in the source language to the target language by machine translation systems and train a model for the target language from scratch [32]. Both of these approaches are summarized in Figure 2.2. This figure illustrates cross-lingual sentiment analysis with a machine translation system. On the left, the approach of translating an existing dataset in the source language to the target language is illustrated. On the right, the approach of translating the target language to the source language during the inference time is illustrated. On the left, there is a trained model in the target language, while there is a trained model in the source language on the right.

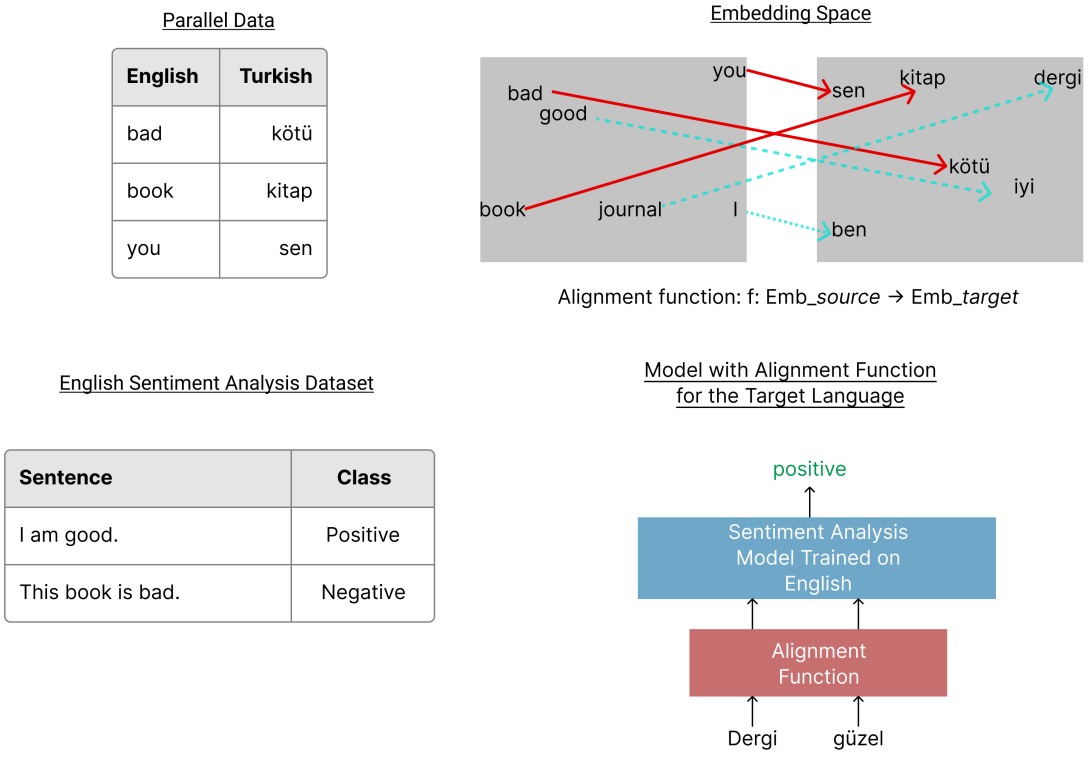


Figure 2.1. Illustration of cross-lingual task with word embedding alignment.

Machine translation or word embedding alignment models are widely used in cross-lingual NLP before the introduction of transformer models. Starting with the multilingual BERT model [9], different multilingual transformer architectures with different pretraining techniques are introduced and gained popularity in cross-lingual NLP. While some of these models (XLM [10]) leverage parallel sentences between language pairs in their training objective, others (mBERT [9], XLM-Roberta [13]) do not take benefit of any parallel data and train on only multilingual corpora without alignment. Still, these multilingual transformer models achieve better scores than word embedding alignment or machine-translation-based systems in cross-lingual NLP. They attract attention in different cross-lingual tasks such as natural language inference [10, 13, 14], question answering [11–13], and named entity recognition [13].

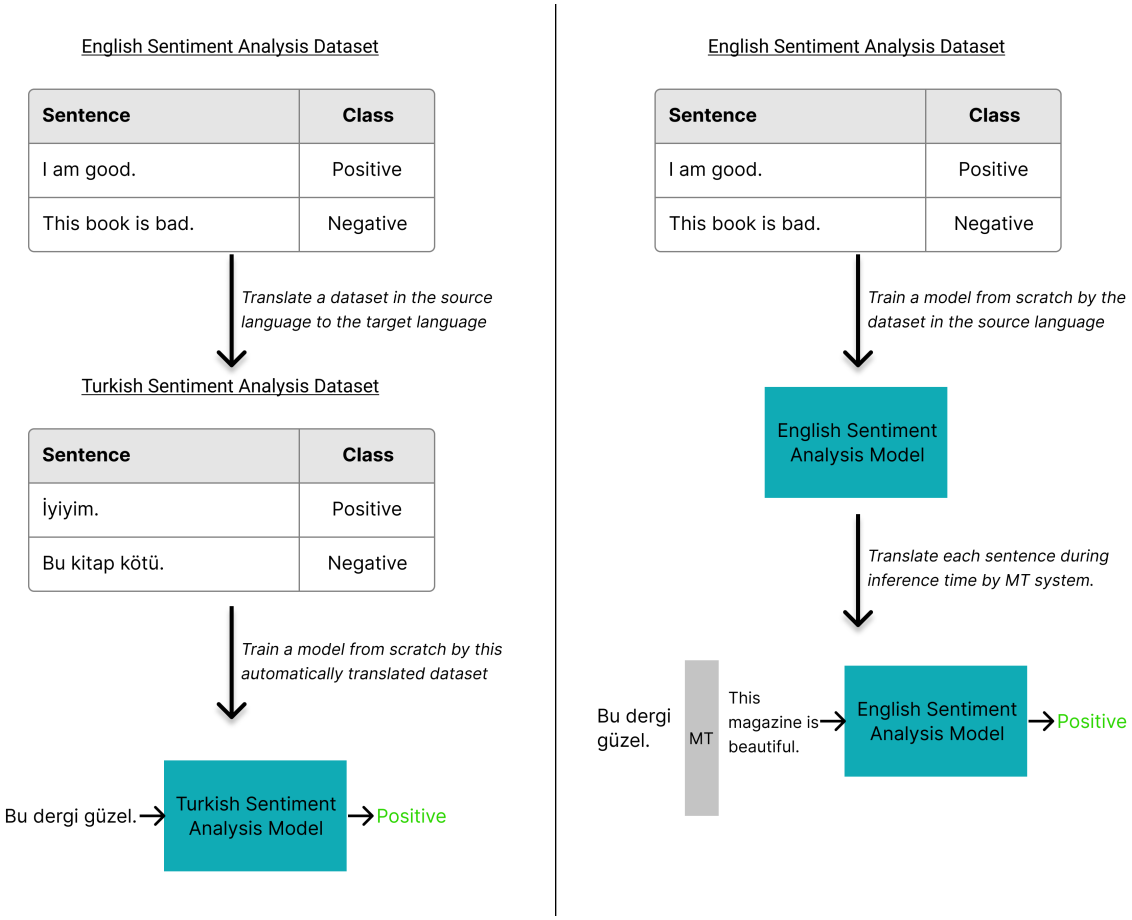


Figure 2.2. Illustration of cross-lingual task with machine translation.

2.2. English and Cross-lingual Relation Classification

Relation classification is the task of extracting relations from a fixed set of classes between given entities in a given text. It differs from *relation extraction* by entity extraction. Relation extraction aims to extract all relations out of a fixed set of classes from a given text without given entities. Therefore, relation extraction contains a named entity recognition task in it. However, relation extraction and relation classification terms are interchangeably used in the domain. On the other hand, open relation extraction [33] aims to extract all types of relations from a given text without any fixed set of classes. It contains named entity recognition, related pair detection, and phrase generation which denotes the relation between a pair. All of these tasks focus on different aspects of relation extraction/classification and help to create a structured information extraction by enabling the automatic construction of knowledge bases and by providing useful information for question answering models [1].

Relation classification and extraction systems help to increase the amount of structured data stored in knowledge bases. Several approaches in knowledge base population task [34] take advantage of relation extraction models. However, these pipelines still require a good amount of annotated data and are generally performed in the English language. Even though knowledge bases generally contain information that is suitable for multiple languages, they do not include necessary information for entities from different regions. For example, the President of the United States (Joe Biden) has more than 489 relations (property-value pairs) in Wikidata including *name, occupation, member of sports team, eye color, medical condition, Goodreads author id.* However, the President of Ghana (Nana Akufo-Addo) has 92 relations (property-value pairs) in Wikidata. By looking at this, we can deduce that some regional information is not included in knowledge bases, and the size of the available information is directly related to the availability of relation extraction systems in written languages of the region. We show the number of relations in Wikidata of capitals and presidents of different regions in Table 2.1.

Table 2.1. Number of relations of different regional entities in Wikidata showing entities residing in regions where high-resource languages spoken have more number of relations than low-resource languages in Wikidata.

Entity	**Number of Relations**	**Sample Relations**
Joe Biden (President of the US)	489	name, occupation, member of sports team, eye color, medical condition, Goodreads author id
Angela Merkel (Chancellor of Germany)	439	residence, personal pronoun, occupation
Ilham Aliyev (President of Azerbaijan)	149	birth name, spouse, signature
Nana Akufo-Addo (President of Ghana)	92	field of work, occupation, religion
Washington, D.C. (Capital of US)	336	head of government, list of monuments, seal image
Berlin (Capital of Germany)	476	located in or next to body of water, twinned administrative body, IPA transcription
Baku (Capital of Azerbaijan)	166	official name, population, coat of arms image
Accra (Capital of Ghana)	137	official language, coordinate location, pronunciation audio

Traditionally, monolingual relation classification models depend on hand-crafted features [2]. With the presentation of word embedding architectures [3, 27], many relation classification methods benefit from pretrained word embeddings with the RNN [4, 5] or CNN [35, 36] architectures. In recent years, with the high performance of transformer architectures for several NLP tasks [9, 10, 37], Soares and co-workers [6] applied BERT with different representations of the sentence with relation and showed the strength of transformer models on several English datasets. Although it is estimated that the non-English content on the web exceeds 40% [7] and the number of multilingual text-corpora is also increasing [8], recent research on relation classification has generally focused on the English language. These supervised methods for relation classification cannot be easily adapted to other languages because they require a large amount of annotated training datasets, which is expensive and time-consuming to create.

Cross-lingual word embeddings have been widely applied in zero-shot cross-lingual NLP with word embedding alignments for different tasks such as named entity recognition [29] and natural language inference [28], as discussed before. Cross-lingual relation classification has benefited from a similar approach [17]. However, recently, multilingual deep transformers [9,10,13] have attracted lots of attention in several cross-lingual tasks such as NLI [14], NER [13], and question answering [11, 12].

Many cross-lingual relation classification works depend on aligned corpora, supervised machine translation systems, or bilingual dictionaries. In [15, 38], projection of English labeled dataset to Korean with aligned corpora is made to train Korean relation classification models. Faruqui and co-workers [16] make use of a pretrained machine translation system to translate the sentence in a target language to a source language so that a relation classification model trained with the source language can be used. Zou and co-workers [39] take benefit of Generative Adversarial Networks to transfer the feature representations from the source language to the target language with the help of machine translation systems. Recently, Ni and co-workers [17] utilize word embedding mappings of two languages, trained with bilingual dictionaries to introduce a cross-lingual relation classification model.

To the best of our knowledge, we introduce the first transformer model for the task of cross-lingual relation classification. In addition, we present a multilingual distant supervision objective for pretraining to improve the baseline transformer model, mBERT. Soares and co-workers [6] use a similar approach for English relation classification, called Matching the Blanks. For the pretraining, they gather English sentence pairs based on the shared entities, annotated by an already existing, supervised entity linking system. On the other hand, we propose a multilingual approach that utilizes Wikipedia and Wikidata, which are already available for many languages and have been successfully used for tasks such as multilingual question answering [40] and named entity recognition [41].

3. BACKGROUND

3.1. Relation Classification

Relation classification is one of the well-studied and key topics in Natural Language Processing to convert unstructured text into structured text. It is first formulated in Message Understanding Conference in 1998 [42] as a subcategory of information extraction. Afterward, many approaches such as rule-based extraction [43], kernel-based machine learning models [44], word-embedding models with RNN [45], and transformer-based large language models [46] are proposed to work on this problem.

Table 3.1. Different inputs in relation classification for the same sentence *"Emma Watson is an English actress, model, and activist who is known by her role as Hermione Granger in the Harry Potter film series."*.

Entity Pairs	**Relation**
(Emma Watson, English)	languages-spoken
(Emma Watson, actress)	occupation
(Hermione Granger, Emma Watson)	performer
(Hermione Granger, Harry Potter)	present in work
(Harry Potter, English)	original language of film

Relation classification is a type of text classification problem. For a given text and entity pair in this text, the aim is to find a relational class out of a fixed set of classes. Contrary to general text classification problems, one text might produce different inputs with different entity pairs. Consider the sentence *"Emma Watson is an English actress, model, and activist who is known by her role as Hermione Granger in the Harry Potter film series."*. In Table 3.1, we show 5 possible relations that can be extracted from a given sentence with different entity pairs. All of these relations are valid and represented in Wikidata. Therefore, relation classification has a unique challenge to represent entities in a sentence properly to extract relations.

Relation classification systems start to gain importance with the rise of text content on the web. To illustrate this change, we check the number of words written in English Wikipedia [47]. As it can be seen from Figure 3.1, the number of words in English Wikipedia has increased from around 5 million words to 3 billion and 900 million words in the last 20 years. Therefore, when we consider the amount of text released on the web each year, proper relation classification algorithms are necessary to create a structured database from unstructured free text.

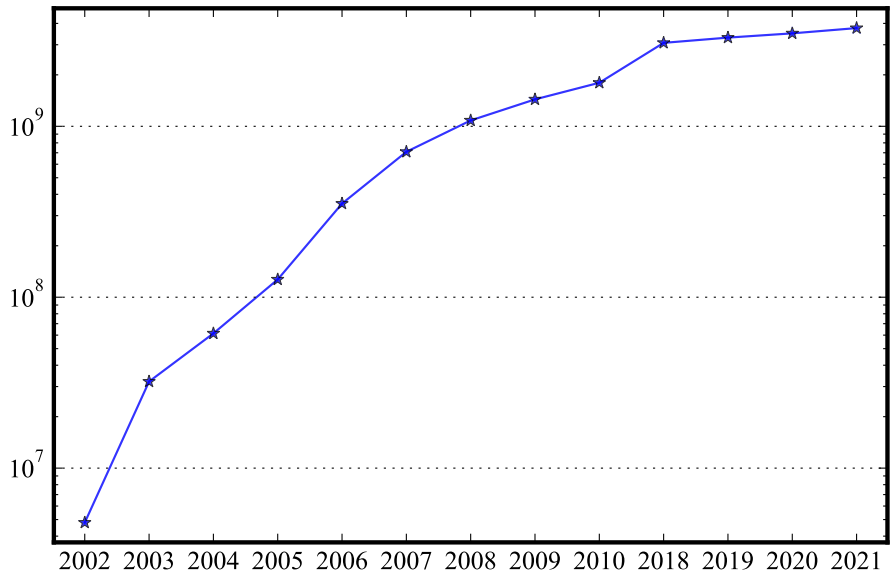


Figure 3.1. The change in the total number of words in the English Wikipedia over the years.

3.2. Transformers

In recent years, deep learning models have gained popularity in various NLP tasks such as question answering, dependency parsing, and intent classification. After the introduction of pretrained word embeddings in 2013, many studies have integrated these word embeddings such as Word2Vec, Glove, FastText to their models. RNN architecture is mainly used in these tasks with word embeddings however RNN architecture

is not scalable due to its sequential nature. This feature disallows to pretrain a large language model as in Computer Vision models like VGGNets [48] and ResNets [49].

RNN models are widely used in machine translation systems with encoder and decoder steps. However, forgetting problems in decoding steps in RNN lead to solutions like the attention mechanism [30]. In recent years, Vaswani and co-workers [50] introduced an architecture with only an attention mechanism that is more scalable and better learner than RNN with attention models. This architecture, transformers, enabled researchers to pretrain large language models by benefiting parallelism in GPUs.

The earlier transformer models like GPT-2 [51] focused on pretraining by the classical left-to-right language modeling approach. They don't benefit from the context of the right part of texts because of the memorization issues during training. Later, Devlin and co-workers [9] introduced a masking mechanism, which helps to avoid memorization and including the right context during pretraining. The model is called BERT, Bidirectional Encoder Representations from Transformers, in which the bidirectional term comes from considering right context as well as left context.

3.2.1. BERT

Bidirectional Encoder Representations from Transformers (BERT) [9] is a pretrained language model based on transformers architecture. It separates from its predecessors by successfully considering left and right context at the same time during training. Furthermore, BERT implements WordPiece [31] tokenization, which overcomes out-of-vocabulary issues which occur in traditional NLP approaches.

WordPiece [31] is a sub-word tokenization mechanism that helps to build a vocabulary for pre-trained transformer models. It is based on Byte-Pair Encoding (BPE) [52], and helps out-of-vocabulary problems and improve performance, especially in morphologically rich languages as Turkish or Finnish. It creates a fixed number of tokens

(around 30,000 for BERT-base) based on word frequency in a given corpus. In the end, frequent words like “and”, “the”, or “you” would have dedicated tokens for themselves however rare words like “linguistically” or “simplification” are tokenized as “linguistic–##ally” and “sim–##pl–##ification”, consecutively. As it can be seen from the examples, tokens starting with ## are continuation tokens and different from tokens for the beginning of the word.

BERT has two special tokens called [CLS] and [SEP]. [CLS] token is used to represent the whole input, and its feature vector is generally fed to another softmax layer for classification tasks. [SEP] token is included to separate multiple inputs for some NLP tasks. For example, the next sentence prediction task includes two sentences as input, and the aim is to determine whether the second sentence comes after the first sentence. Instead of giving these two sentences directly to the model, they are separated with [SEP] token to emphasize the separation in the input.

For the pretraining procedure, BERT focuses on two tasks: masked language model and next sentence prediction. In traditional NLP, standard conditional language models aim to find the next word based on the left context. At the same time, it is intuitively believed considering both contexts (bi-directionality) improves its performance (and it is also shown for BERT [9]). However, it is not possible to use them in standard conditional language models, as every word indirectly sees itself. To overcome this issue, BERT randomly masks each word in the input (i.e. replace words with special [MASK] token) and predicts them based on other non-masked tokens. This helps to consider left and right context during language model pretraining and produces contextualized word embedding which extract different feature vectors for the same word with different contexts. The architecture and the input-output types of masked language model task are given in Figure 3.2.

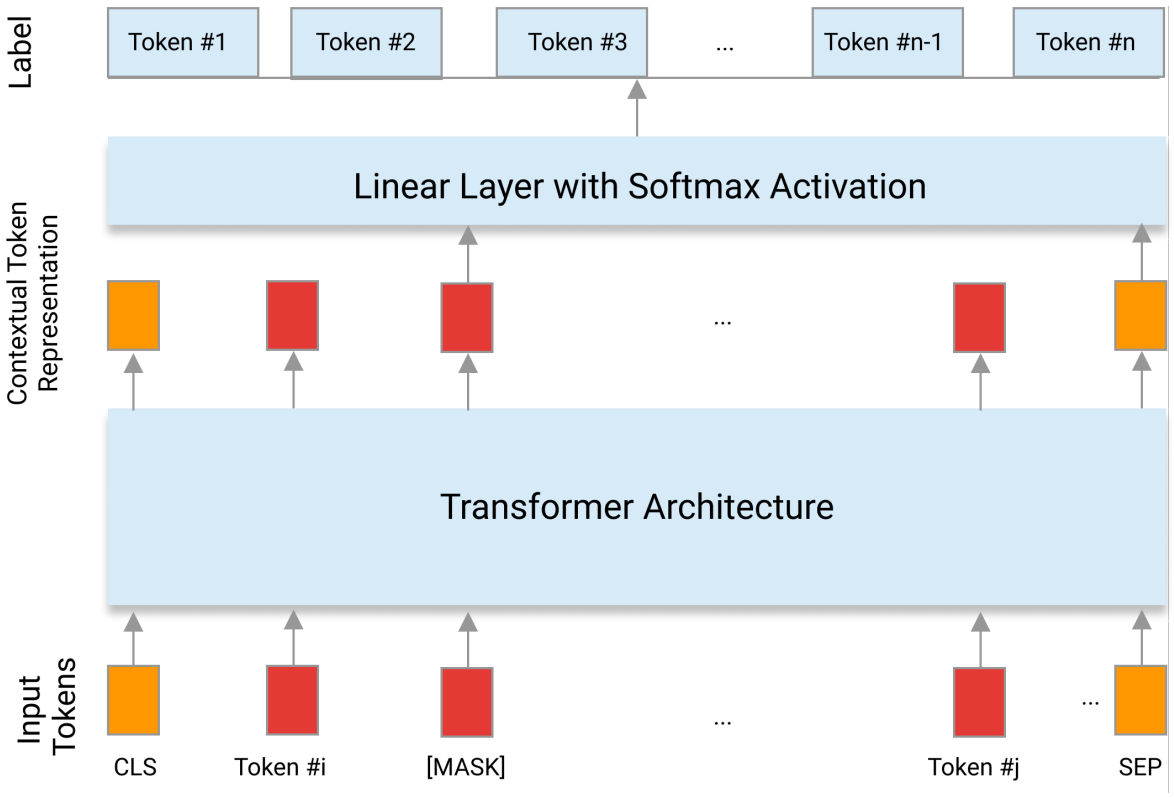


Figure 3.2. Masked language model pretraining in BERT. For each masked token in the input, the model predicts the corresponding token.

As many NLP tasks such as question answering and natural language inference rely on relations between sentences, only masked language models pretraining would not be sufficient. To handle these types of tasks, BERT is pretrained with another task called next sentence prediction. During the pretraining, two sentences, separated with [SEP] token, are given to the model. The feature vector of [CLS] token is fed into the classifier layer, and the model has decided whether the second sentence is the next sentence of the first sentence or not. The architecture and the input-output types of the next sentence prediction task are given in Figure 3.3.

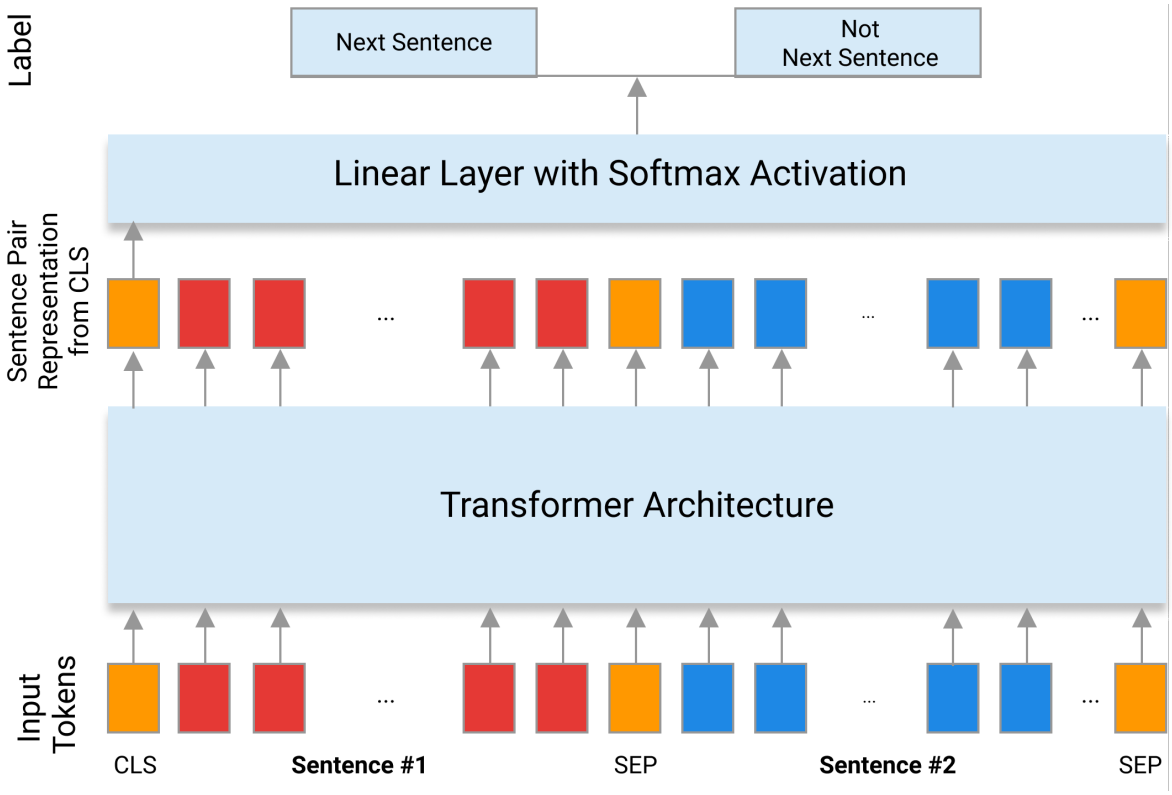


Figure 3.3. Next sentence prediction pretraining in BERT. [CLS] token is used to represent input with two sentences separated by [SEP] token. This [CLS] token is fed to binary classification.

BERT is first pretrained for the English language in [9]. They have used BookCorpus [53] with 800 million words, and English Wikipedia with 2,500 million words to create a dataset for both next sentence prediction and masked language model tasks. The final BERT model with a large setup (340 million parameters) outperformed GPT-2 model by a 7.7% improvement on the GLUE benchmark [54].

3.2.2. Multilingual BERT

Devlin and co-workers [9] first introduced the pretrained BERT model for the English language, and show its effectiveness in several English NLP tasks such as natural language inference, question answering, and named entity recognition. Even though

state-of-the-art performance is achieved for several tasks in English, applications to the other languages require a pretraining BERT model from scratch with a fairly good unannotated corpus. Pretraining a BERT model requires hardware such as GPUs or TPUs which is expensive and time-consuming. Furthermore, many languages don't have an unannotated corpus with sufficient data and good quality. At the time of writing, only 18 languages have more than 1,000,000 articles, and 253 languages have less than 10,000 articles on Wikipedia [55].

Following the introduction of the English BERT model, Devlin and co-workers [9] released a multilingual BERT (mBERT) model. Instead of BookCorpus and English Wikipedia, it is pretrained on Wikipedia dump of 104 languages. No additional steps are performed for multilinguality apart from changing vocabulary size from around 30,000 to around 120,000 and fixing normalization issues. Still, mBERT outperforms previous works in several cross-lingual tasks, such as named entity recognition.

Pires and co-workers [56] perform several experiments to show the effectiveness of mBERT in different cross-lingual settings and try to clarify its underlying learning mechanism. They state that vocabulary overlap, language similarity, code-switching, and shared space help to transfer knowledge between languages. They show that when the number of overlapped vocabulary or common World Atlas of Language Structures (WALS) [57] features increases between two languages, the cross-lingual performance improves. Furthermore, code-switching (having more than one language in the same statement) helps contextual mapping between languages. Finally, their dataset (Wikipedia) includes lots of common tokens such as URLs, numbers, and symbols in different languages. They show that these common tokens create a shared space between languages. Thanks to these features, the ability of information transfer in a cross-lingual setup in mBERT is powerful between languages, especially within similar ones.

4. DATASETS

In this chapter, we first analyze one of the used datasets in cross-lingual relation classification, ACE05 [58]. Afterward, we compare widely used English relation classification datasets, KBP-37 [4], SemEval [59], TACRED [26], because we propose a cross-lingual evaluation benchmark called RELX with a translation of the most suitable English relation classification dataset. Then, we give a detailed explanation about our proposed datasets: *RELX*, human-annotated cross-lingual relation classification benchmark, and *RELX-Distant*, weakly-supervised large-scale multilingual relation classification dataset.

The cross-lingual relation classification task requires training data for the source language, which is generally in English, and evaluation data for the target language. In recent works [60], ACE05 [58] data are used to train and evaluate cross-lingual relation classification. In Table 4.1, we briefly summarize the features of ACE05.

Table 4.1. Statistics of the ACE05 dataset.

	Statistics
Number of Relation Mentions	
English	8,738
Chinese	9,317
Arabic	4,731
Number of Unique Classes	6

As it can be seen from Table 4.1, ACE05 contains training and test data just for 3 languages: English, Arabic, and Chinese. Furthermore, the number of relations in this dataset is 6 which might create obstacles between academic studies and real-life settings as knowledge graphs generally contain more than thousands of relations [61, 62]. In addition to these, the ACE05 dataset is not publicly available, and it requires $4,000

to get this dataset from LDC for a non-member user. Furthermore, in [60], a relation classification dataset for 6 languages with 53 relation types has been proposed, but the dataset is not publicly released.

Due to discussed issues for ACE05 [58], we propose a new publicly available cross-lingual relation dataset. We aim to include a high number of languages and relation classes in the proposed dataset and make it publicly available. However, creating a proper dataset might be a very expensive task and includes several challenges such as deciding on relation labels, semi-automatically collecting a variety of sentences containing these relations in different languages, extracting entities, and labeling relations with the help of multiple annotators. To address these problems, we decide to start with an already existing and widely used relation classification dataset in the English language and translate the sentences in the test set to different languages with human annotators to create an evaluation benchmark. By doing this, we would simplify this process and support works that are already done in the selected English relation classification dataset.

Table 4.2. Datasets for English Relation Classification: KBP-37, SemEval, TACRED.

Properties	**KBP-37**	**SemEval**	**TACRED**
# of Documents in Training	15,917	8,000	68,120
# of Documents in Validation	1,724	-	22,631
# of Documents in Test	3,405	2,717	15,509
# of Unique Classes	37	19	42
Average Length	30.3	17.2	36.4
Percentage of Negatives	9.7%	17.4%	79.5%
An Example Label	Founded By	Cause-Effect	Spouse
Domain	General	Semantic	General
Publicly Available	Yes	Yes	Yes
Cost	$0	$0	$25
SOTA Test F1 [6]	69.3	89.5	71.5

We analyze the widely used relation classification datasets in the English language to select a proper dataset for the translation process. As we discussed in previous chapters, recent relation classification methods rely on neural networks with supervised settings. Even though we work on cross-lingual relation classification, meaning that no training data is required for target languages, we need a large amount of training data for the source language, English. We select widely used three English relation classification datasets for the training: SemEval 2010 dataset [59], TACRED [26], and KBP-37 [4]. In Table 4.2, we summarize these datasets.

As one can see in Table 4.2, these datasets have different characteristic features. SemEval 2010 dataset contains semantic relations between pairs of nominals with 10 distinct classes including others. However, it has a different challenge that requires common-sense knowledge and understanding of the nominals while widely used knowledge graphs generally contain named entities. Furthermore, its average length in terms of words is much less than TACRED and KBP-37 which makes it an easier task as we can see from the result of the SOTA [6], at the time of the writing, is much higher than the others. TACRED and KBP-37 datasets are in a very similar domain and contain a similar number of unique classes and sentences with positive labels (sentences with a relation except no_relation class) in the training data. Further, the state-of-the-art approach [6] at the time of the dataset selection process has similar scores for both datasets. We select KBP-37 as our baseline because it is on par with our motivation and freely available contrary to TACRED. In the next section, we analyze the KBP-37 dataset in depth.

4.1. KBP-37

KBP-37 [4] is one of the widely used and known datasets in the relation classification domain. It has 18 unique relation classes with direction and *no_relation* class, which result in 37 classes in total. Many relations in the relation classification tasks are not transitive, and they require a direction. For an example sentence, “As a member of the <e1> Eagles </e1>, <e2> Frey </e2> has won six Grammys and five American

Music Awards.", the relation between entities is *employee_of*. However, the direction should also be found to discover whether Eagles is the employee of Frey or Frey is the employee of Eagles. In order to ensure that, all relations except no_relation class have 2 directions (e.g. *employee_of(e1,e2)*, *employee_of(e2,e1)*) during classification.

The detailed statistics of KBP-37 can be seen in Table 4.2. As the number of sentences in the training data is 15,917, it might be suitable for supervised training with deep neural networks.

The distribution of the classes can be seen in Table 4.3. The most common three classes in KBP-37 are *per:employee_of*, *per:countries_of_residence*, and *no_relation* consecutively. Even though the distribution of labels in the training set is not uniform, they have similar statistical features across sets; training, development, and test. Furthermore, the average number of characters and words are very similar across all sets.

Table 4.3. Samples from KBP-37 for each unique labels.

Relation	**Sample**	**Frequency**	**# Directions**
per: employee of	<*e1*> *Rubin* </*e1*> 69 is a former secretary of the US <*e2*> *Treasury* </*e2*> .	3472	1862, 1610
per: countries of residence	<*e1*> *Daniel Ortega* </*e1*> was sworn in as president of <*e2*> *Nicaragua* </*e2*> on Wednesday .	1660	955, 705
no relation	The <*e1*> *300* </*e1*> <*e2*> *area* </*e2*> is operated by the Battelle Memorial Institute .	1545	- (No dir.)

Table 4.3. Samples from KBP-37 for each unique labels. (cont.)

Relation	**Sample**	**Frequency**	**# Directions**
org: city of headquarters	<e1> *Lufthansa* </e1> 's corporate headquarters are located in <e2> *Cologne* </e2> Germany .	1267	980, 287
org: country of headquarters	<e1> *Hinduja Foundries* </e1> is <e2> *India* </e2> largest casting maker .	1006	618, 388
org: subsidiaries	<e2> *OBC* </e2> was sold to the <e1> *Go-Ahead Group* </e1> in 1994 .	832	402, 430
per: stateorprovinces of residence	While at the University of <e2> *Florida* </e2> she became engaged to <e1> *Bob Graham* </e1> .	720	491, 229
org: members	The <e2> *Assam Regiment* </e2> is an infantry regiment of the Indian <e1> *Army* </e1> .	703	217, 486
per: cities of residence	<e1> *Mel* </e1> resides in <e2> *Orlando* </e2> .	663	425, 238
per: title	<e1> *Collins* </e1> ' biological father Robert Latta was a New York stage <e2> *actor* </e2> .	641	221, 420
org: top members / employees	<e1> *GlaxoSmithKline* </e1> promotes <e2> *Andrew Witty* </e2> to chief executive .	576	265, 311
org: stateorprovince of headquarters	The <e1> *National Airlines* </e1> division has its offices in Orlando <e2> *Florida* </e2> .	517	399, 118

Table 4.3. Samples from KBP-37 for each unique labels. (cont.)

Relation	**Sample**	**Frequency**	**# Directions**
org: alternate names	It was renamed first Airfield B *<e2> 152 </e2>* and later *<e1> RAF Fassberg </e1>* .	511	278, 233
org: founded	He began by forming the rock band *<e1> Deus </e1>* in Antwerp in *<e2> 1989 </e2>* .	393	252, 141
org: founded by	*<e1> Rap Snacks </e1>* were created by *<e2> James </e2>* Fly Lindsay in 1994 .	355	187, 168
per: country of birth	*<e1> Natalia </e1>* Estrada was born on September 3 1972 in Asturias *<e2> Spain </e2>* .	355	250, 105
per: origin	Whale Music is a novel by *<e2> Canadian </e2>* writer *<e1> Paul Quarrington </e1>* .	266	127, 139
per: spouse	*<e1> Korda </e1>* 's wife *<e2> Maria Corda </e2>* starred in several Corvin productions .	258	139, 119
per: alternate names	Previous line ups included *<e1> Alex Kapranos </e1>* (*<e2> Franz Ferdinand </e2>*) .	177	93, 84

Finally, sample sentences for each class with direction can be observed from Table 4.3. *(<e1> ,</e1>)* and *(<e2> ,</e2>)* tag pairs are used to indicate the target entities. For illustration purposes, all the samples are collected from the same direction (e1 $\rightarrow$ e2) or changed to this direction by changing tags. In conclusion, KBP-37 is selected as our training and validation set, and a subset of KBP-37's test set is proposed to create a cross-lingual relation classification benchmark by translation.

4.2. Proposed Datasets

In this section, we define the proposed datasets to create an evaluation benchmark and pretrain a transformer-based large language model for cross-lingual relation classification. First, we propose a manually annotated (through translation) RELX benchmark and describe the annotation process. Second, a distantly supervised multilingual relation classification dataset is presented.

4.2.1. RELX

As we stated at the beginning of Chapter 4, we aim to translate a subset of already existing and widely used English relation classification dataset to create a cross-lingual benchmark to four languages, French, German, Spanish, and Turkish. We follow four steps to create a cross-lingual benchmark, called RELX. First, we find a proper subset of KBP-37's test set by preserving specific statistics. Then, we use a machine translation system to automatically translate this subset with the tags. After that, human annotators edit and translate sentences with machine translation. Finally, a professional translation service called El Turco evaluates the quality of RELX.

KBP-37 [4] contains 3,405 sentences in the test set. Due to cost constraints, we select 502 sentences out of this. Instead of selecting randomly, we preserve statistical features of KBP-37. We consider average character length, average word length, and distribution of class labels. 10,000 different subsets are selected randomly by conforming to the class distribution of KBP-37. Out of these 10,000 subsets, the most similar one to KBP-37 is selected in terms of the sum of the normalized average character length and normalized average word length. Average character/word length normalization is performed by dividing the average character/word length in the original KBP-37 test dataset. Due to the variety in the languages, the average number of characters and words in the sentences can differ for different languages, but the RELX-English and KBP-37 test sets have similar distributions. The average number of characters and words in the KBP-37 test set are 180.2 and 30.2, and the average number of characters

and words in RELX-English are 171.2 and 28.9. It still has differences because we force the subset to have the same distribution of labels as KBP-37 test set.

Table 4.4. Different translation techniques to capture entity alignments in Google API.

Technique	**English Sentence (Source)**	**Google API Translation to Turkish Sentence (Target)**
Baseline	<e2> CNN </e2> and Court TV units of <e1> Turner Broadcasting System </e1> are owned by Time Warner Inc .	<e2> CNN </e2> ve <e1> Turner Broadcasting System </e1> 'in Court TV birimlerinin mülkiyeti Time Warner Inc'dir.
Removing	E2 and Court TV units of E1 are owned by Time Warner Inc .	E1'in E2 ve Court TV üniteleri Time Warner Inc.'e aittir.
Simple Tags	E2 CNN E2 and Court TV units of E1 Turner Broadcasting System E1 are owned by Time Warner Inc.	E1 Turner Broadcasting System E1'in E2 CNN E2 ve Court TV üniteleri Time Warner Inc'e aittir.
W/o Tags	CNN and Court TV units of Turner Broadcasting System are owned by Time Warner Inc .	Turner Broadcasting System'in CNN ve Court TV üniteleri Time Warner Inc.'e aittir.

After selecting a proper subset of KBP-37, which is RELX-English, we use Google API to translate these sentences into French, German, Spanish, and Turkish to make annotators' jobs easier. Current commercial models do not release word alignments between source and target languages. Therefore, we translate sentences with tags (<e1> ,</e1>, <e2> ,</e2>) to track entities in the translated sentences. However, current commercial translation models perform poorly when special tags are introduced because the syntax of the sentence would not be processed correctly. Overall, their commercial models have worse performance for the sentences with these tags. We also replace entities and corresponding tags between e1 tags with symbols like E1, but

the model encounters difficulties to capture the property of entities when we replace them. Furthermore, when we replace entities with symbols, the translation of the entities would not be available. Finally, we use E1 and E2 tags for (<e1> ,</e1>), and (<e2> ,</e2>), consecutively. Through few samples, we observe that this has the best performance in Google API machine translation. In Table 4.4, we show an illustration of these three translation techniques to capture entity alignments with English (source) and Turkish (target) languages.

In the third step, our annotators have edited manual translations. In total, we have seven human annotators who are proficient in both source and target languages that are assigned to them. An Excel spreadsheet is shared with annotators, as shown in Figure 4.1.

The first column, Verified, is to check the progress of annotators. They mark these cells with 'X' whenever they finish editing the target sentence. If they have any questions, they leave these cells blank, and we would analyze the corresponding sentences in depth. The second column contains an original dataset with simple tags. The annotators are not allowed to edit English sentences even if there are mistakes (typo, long URLs, etc.) in the original sentence. The third column contains the translated sentence in the target language. They edit these columns to make the sentence grammatically and semantically correct. They also change the position of E1 and E2 tags to make sure that they point to the same entities as the original sentence. The last two columns contain JavaScript macros that show entities between E1 and E2 tags in both sentences. Thanks to this, annotators can easily check whether tags are placed correctly.

Verified	İngilizce Cümle	Türkçe Cümle	E1	E2
X	The E1 Sacramento Heatwave E1 is an American Basketball Association (ABA) team based in Sacramento E2 California E2.	E1 Sacramento Heatwave E1, Sacramento E2 California E2 merkezli bir Amerikan Basketbol Birliği (ABB) takımıdır.	Sacramento Heatwave = Sacramento Heatwave	California = California
X	E1 Sydenham Institute of Management Studies E1 Research and Entrepreneurship Education (E2 SIMSREE E2) is one of the Premiere Management Institute of the country imparting Management Studies under University of Mumbai named after the then governor of Bombay Lord Sydenham of Combe in 1913.	E1 Sydenham Yönetim Araştırmaları Enstitüsü E1 Araştırma ve Girişimcilik Eğitimi (E2 SYAEAGE E2), 1913 yılında Combe, Bombay Lord Sydenham'ın valisi tarafından adı verilen Mumbai Üniversitesi'ne bağlı Yönetim Araştırmaları veren ülkenin Premiere Yönetim Enstitülerinden biridir.	Sydenham Institute of Management Studies = Sydenham Yönetim Araştırmaları Enstitüsü	SIMSREE = SYAEAGE
X	The 2001 E1 Indiana Hoosiers E1 football team represented E2 Indiana University E2 Bloomington during the 2001 NCAA Division I-A football season.	2001 E1 Indiana Hoosiers E1 futbol takımı, 2001 NCAA Division I-A futbol sezonunda E2 Indiana Üniversitesi E2 Bloomington'ı temsil etti.	Indiana Hoosiers = Indiana Hoosiers	Indiana University = Indiana Üniversitesi
X	Some staff at Osaka University are represented by the General Union a member of the E1 National Union of General Workers E1 (NUGW) which is itself a member of the National Trade Union Council (E2 Zenrokyo E2).	Osaka Üniversitesi'ndeki bazı personeller, Ulusal Ticaret Sendika Konseyi (E2 Zenrokyo E2) üyesi olan E1 Ulusal Genel İşçiler Sendikası E1 'nın (NUGW) bir üyesi olan İşçi Sendikası tarafından temsil ediliyorlar.	National Union of General Workers = Ulusal Genel İşçiler Sendikası	Zenrokyo = Zenrokyo
X	When the E1 schools E1 were consolidated in E2 1974 E2 Haubstadt student athletes had to participate in the Pocket Athletic Conference.	E1 Okullar E1 E2 1974 E2 'te birleştirilince Haubstadt'daki öğrenci atletler Cep Atletik Konferansı'na katılmak zorunda kaldı.	schools = Okullar	1974 = 1974
X	By E1 1995 E1 the merger of TSB Group Plc. and Lloyds Bank led Hill Samuel to become a subsidiary of E2 Lloyds TSB E2.	E1 1995 E1 tarihinde TSB Grup Plc. ve Lloyds Bank'ın birleşmesi, Hill Samuel'in E2 Lloyds TSB E2'nin bir yan kuruluşu olmasını sağladı.	1995 = 1995	Lloyds TSB = Lloyds TSB
-	Jonny Walker (born in Preston Lancashire) is an English rugby league footballer playing for the E1 Wigan Warriors E1 in the European Super League competition Blackpool Panthers and E2 Batley Bulldogs E2 as a.	Jonny Walker (Preston Lancashire doğumlu), Avrupa Süper Lig yarışması Blackpool Panthers ve E2 Batley Bulldogs E2'de E1 Wigan Warriors E1 için oynayan bir İngilitere Rugby Ligi futbolcusu.	Wigan Warriors = Wigan Warriors	Batley Bulldogs = Batley Bulldogs
X	In Eckington was also the Eckington Secondary Modern School on School Street ; when the grammar school became E1 Derbyshire E1 's first comprehensive school - the Westfield School - in 1957 this became E2 Eckington Junior School E2.	Eckington'da, ayrıca School caddesindeki Eckington Modern Ortaokulu vardı, 1957'de dilbilgisi okulu E1 Derbyshire E1'ın ilk meslek okulu - Westfield Okulu - olduğunda, bu E2 Eckington İlkokulu E2 oldu.	Derbyshire = Derbyshire	Eckington Junior School = Eckington İlkokulu
X	E1 NBC Universal E1 and Microsoft the parents of E2 msnbc.com E2 are holding high-level talks about changing its name an unusual and potentially risky endeavor for the third most popular news website in the United States.	E2 msnbc.com E2'un ana kuruluşları olan E1 NBC Universal E1 ve Microsoft, msnbc.com'un adını, ABD'deki üçüncü en popüler haber sitesi için alışılmadık ve potansiyel olarak riskli bir çabayla, değiştirme konusunda üst düzey görüşmeler yapıyor.	NBC Universal = NBC Universal	msnbc.com = msnbc.com

Figure 4.1. An Excel spreadsheet for annotators to edit manual translations.

The guideline is shared with all annotators. The guideline includes steps that we describe in the previous paragraph. Due to linguistic differences between source and target languages, annotators sometimes add or omit some words in the target sentence and may produce sentences with lower pragmatic effects to tag entities. For example, Turkish has a null subject but when the source language contains a sentence similar to "E1 He E1 knows", annotators translate it as "E1 O E1 biliyor" to preserve tags instead of translating it as "Biliyor" with a null subject.

Even though our human annotators are proficient in both languages, they are not professional translators. However, we work with a professional translation company to evaluate the quality of translated sentences. We randomly select 50 (10% of) English sentences from RELX-English with the human translations in the target languages, French, German, Spanish, Turkish. A professional translation company, El Turco, performs language quality assessments with the help of professional translators. Except for article and synonym mistakes, there were less than three sentences with errors in each language and no critical errors were found in any of the translations.

Table 4.5. Comparative statistics of KBP-37 and RELX in different languages.

Dataset	**Total Sentences**	**Average Chars**	**Average Words**
KBP-37			
Train	15917	181.21	30.28
Dev	1724	181.77	30.55
Test	3405	180.20	30.23
RELX			
English	502	171.18	28.88
French	502	186.63	30.99
German	502	188.27	27.73
Spanish	502	188.37	31.85
Turkish	502	170.76	23.60

Table 4.6. Sample parallel sentences from RELX in different languages.

English	*<e1> Hoyte </e1>* was born in *<e2> Guyana </e2>* 's capital Georgetown.
French	*<e1> Hoyte </e1>* est né à Georgetown, la capitale d' *<e2> Guyane </e2>*
German	*<e1> Hoyte </e1>* wurde in der Hauptstadt Georgetown von *<e2> Guyana </e2>* geboren.
Spanish	*<e1> Hoyte </e1>* nació en la capital de *<e2> Guyana </e2>* , Georgetown.
Turkish	*<e1> Hoyte </e1>* , *<e2> Guyana </e2>* 'nın başkenti Georgetown'da doğdu.
Category	*per:country_of_birth(e1,e2)*

Finally, we propose a human-annotated cross-lingual relation benchmark called RELX. The statistics of RELX, compared to KBP-37, can be seen in Table 4.5. In RELX, Turkish translations have a lower number of words on average in the sentences due to the agglutinative nature of Turkish. The characters and words represent the average length of sentences in the corresponding dataset. In Table 4.6, we show an example of a parallel sentence from RELX with the marked entities for a sample relation.

4.2.2. RELX-Distant

As we discuss in the previous sections, supervised neural models are "data-hungry". Furthermore, all pretraining objectives in transformers architecture usually require a large amount of weakly annotated data. As we propose a new pretraining objective called MTMB, we provide a weakly labeled multilingual relation classification dataset from Wikipedia and Wikidata via a distant supervision approach called RELX-Distant.

Distant supervision is one of the techniques in supervised learning that aims to train a supervised model with a weakly labeled dataset. These kinds of datasets are easier and cheaper to create than fully supervised datasets and may help to have better models with a large number of examples.

The idea to implement distant supervision techniques to the relation classification is introduced in [63]. In recent years, knowledge bases such as Freebase [64], Wikidata [61], DBpedia [65] have become publicly available and include numerous entities and relations. Construction [66] and completion [67] of a knowledge base require several NLP algorithms such as text classification and entity recognition to create a structured database from unstructured text. However, the process is thought backward in distant supervision. If two entities have a relation in a KB, the assumption in the distant supervision is that a sentence containing these two entities has information about a relation. Even though this assumption might not hold several times [68], it helps to improve the accuracy. Thus, entity recognition and linking methods have been used to map entities in the unstructured text to the structured KBs to find possible relations.

With the recent improvements of transformers-based large language models [9,37], distant supervised data become the center of pretraining, not fine-tuning. In [6], they collect a distantly supervised English relation classification dataset by using an off-the-shelf English entity linking system. First, they extract paragraphs from English Wikipedia. Then, they use the off-the-shelf tool to link entities with unique identifiers, such as Freebase or Wikidata ID. After that, they further pretrain $\text{BERT}_{\text{large}}$ model with an additional objective called Matching the Blanks, and show improvements over $\text{BERT}_{\text{large}}$ model for English relation classification.

Soares and co-workers [6] show that distantly supervised data are beneficial to learn better embeddings for the relation classification task. However, they use a commercial English tool to link entities and show improvements just for the English language. Following the same line of thought, we overcome these problems by using hyperlinks on Wikipedia and map them to Wikidata. By doing this, we ensure that

our system does not rely on any NLP algorithm and might be extended to any language that Wikipedia is available. An example from Turkish Wikipedia and Wikidata is shown in Figure 4.2.

Figure 4.2. An example from Turkish Wikipedia and Wikidata. Thanks to hyperlinks in Turkish Wikipedia, we can easily map entities to Wikidata ID's, and check relations from Wikidata to collect distantly supervised data in multiple languages.

Finally, we collect a large number of multilingual sentences with relations from Wikipedia and Wikidata by a distant supervision scheme [63] and create the RELX-Distant weakly-labeled dataset for relation classification in English, French, German, Spanish, and Turkish.

The following steps are used to create RELX-Distant:

(i) The Wikipedia dumps for the corresponding languages are downloaded and converted into raw documents with Wikipedia hyperlinks in entities.
(ii) The raw documents are split into sentences with spaCy [69], and all hyperlinks, which refer to entities, are converted to their corresponding Wikidata IDs.
(iii) Sentences that include entity pairs with Wikidata relations [61] are collected.

Table 4.7. Number of sentences with a relation in each language in RELX-Distant.

Language	**Number of Sentences**
English	815,689
French	652,842
German	652,062
Spanish	397,875
Turkish	57,114

The statistics about the created RELX-Distant dataset are provided in Table 4.7. After merging similar relations such as *capital* and *capital of*, RELX-Distant contains the following 24 relations, each of which includes at least 1000 sentences in English Wikipedia.

author, *capital*, *characters*, *continent*, *country of citizenship*, *country of origin*, *developer*, *ethnic group*, *father*, *instance of*, *language*, *located in country*, *member of*, *mother*, *owned by*, *parent organization*, *parent taxon*, *part of*, *partner*, *performer*, *place of birth*, *religion*, *sibling*, *spouse*

5. METHODOLOGY

In this chapter, we propose our methodology and present and discuss the results for cross-lingual relation classification. The first proposed model is based on multilingual BERT [9], and we finetune mBERT with the KBP-37 [4] dataset and make predictions on both RELX and KBP-37 datasets. The second proposed model pretrains a public checkpoint of mBERT, released by [9], with two objectives: Masked Language Model (MLM) which is already used in [9] and Matching the Multilingual Blanks which is a unique pretraining objective. We finetune this pretrained model (MTMB) with the KBP-37 dataset and check results on KBP-37 and RELX datasets.

5.1. Task Definition

In the cross-lingual relation classification task, a source language dataset D_s with n_s sentences containing related entity pairs is given.

$$\begin{aligned} D_s &= \{(S_i^s, E1_i^s, E2_i^s, r_i)\}_{i=1}^{i=n_s} \text{ where} \\ S_i^s &= [w_1, w_2, ..., w_n] \\ E1_i^s &= (w_k, w_{k+1}, ..., w_l) \\ E2_i^s &= (w_p, w_{p+1}, ..., w_q) \\ r_i &\in R \end{aligned} \tag{5.1}$$

$E1_i^s$ and $E2_i^s$ correspond to entities and w_i correspond to tokens in the sentence S_i^s. r_i is the directional relation between $E1_i^s$ and $E2_i^s$ in S_i^s, selected from a predefined relation set R.

Given a test set in the target language $D_t = \{(S_j^t, E1_j^t, E2_j^t)\}_{j=1}^{j=n_t}$, the objective of cross-lingual relation classification is capturing the probability of relation $P(r_j|S_j^t, E1_j^t, E2_j^t)$ where $r_j \in R$ for a given sentence and entity pair in the target language while the supervision of D_s in the source language.

5.2. Multilingual BERT and BERT

BERT is a strong baseline for many English NLP tasks, and mBERT is for many cross-lingual NLP tasks, as we discussed in Chapter 2. Even at the time of publication, they achieve state-of-the-art results at the General Language Understanding Benchmark [54] which contains several NLP tasks such as natural language inference, sentiment analysis, and textual similarity detection.

We follow the same line of thought and propose pretrained mBERT and $\text{BERT}_{\text{base}}$ models as our baseline models. $\text{BERT}_{\text{base}}$ model is only used to compare results in English (monolingual) relation classification, while mBERT is used to compare in both English and cross-lingual relation classification. Results of English relation classification in the $\text{BERT}_{\text{large}}$ model are reported in [6], and we compare our results with them.

Table 5.1. Comparing architecture of pretrained mBERT, $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{large}}$.

Features	**mBERT**	**$\text{BERT}_{\text{base}}$**	**$\text{BERT}_{\text{large}}$**
Number of Hidden Layers	12	12	24
Number of Self Attention Heads	12	12	16
Hidden Size	768	768	1024
Vocabulary Size	119,547	30,522	30,522
Number of Parameters	110M	110M	340M
Number of Languages	104	1 (English)	1 (English)
Training Corpus	Wikipedia	BookCorpus and English Wikipedia	BookCorpus and English Wikipedia

The comparisons between transformer models can be seen in Table 5.1. While $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ are pretrained on only English with additional BookCorpus, mBERT is pretrained on 104 languages with only Wikipedia corpus. Furthermore, $\text{BERT}_{\text{large}}$ is around 3 times larger than both $\text{BERT}_{\text{base}}$ and mBERT in terms of the number of parameters, which is affected by the number of hidden layers and the number of self-attention heads. Finally, the vocabulary size is four times bigger than $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ in order to capture various tokens and encoding in different languages.

For the input representation, entity markers are added to emphasize the location of the entities, following [6]. For example, for a given sentence *"Holy Cross High School is a Catholic secondary school founded in Waterbury Connecticut in 1968 by the Congregation of Holy Cross"*, where 'Holy Cross High School' and 'Congregation of Holy Cross' are the given entities and 'founded_by' relation, the input is marked as *"<e1> Holy Cross High School </e1> is a Catholic secondary school founded in Waterbury Connecticut in 1968 by the <e2> Congregation of Holy Cross </e2> "*.

The pipeline for the fine-tuning as follows:

(i) Entities in each sentence in the datasets is marked with "<e1> ", "</e1> ", "<e2> ", and "</e2> " markers. These entity start and end markers are special tokens given to the model and learned from scratch. We use placeholder tokens for these markers in the implementation of HuggingFace's tokenization library [70]. Each marker is a specific token and has a length of one (which means they are not divided by the tokenizer).

(ii) [CLS] and [SEP] tokens are included at the beginning and end of the input, consecutively.

(iii) The input is fed into transformers architecture, and then the fixed-length sentence representation of [CLS] token (in our case, relation representation) is fed into a softmax classifier that has the number of outputs equal to the number of unique relations. KBP-37 and RELX contain 18 directional and no_relation classes, which result in 37 unique relations.

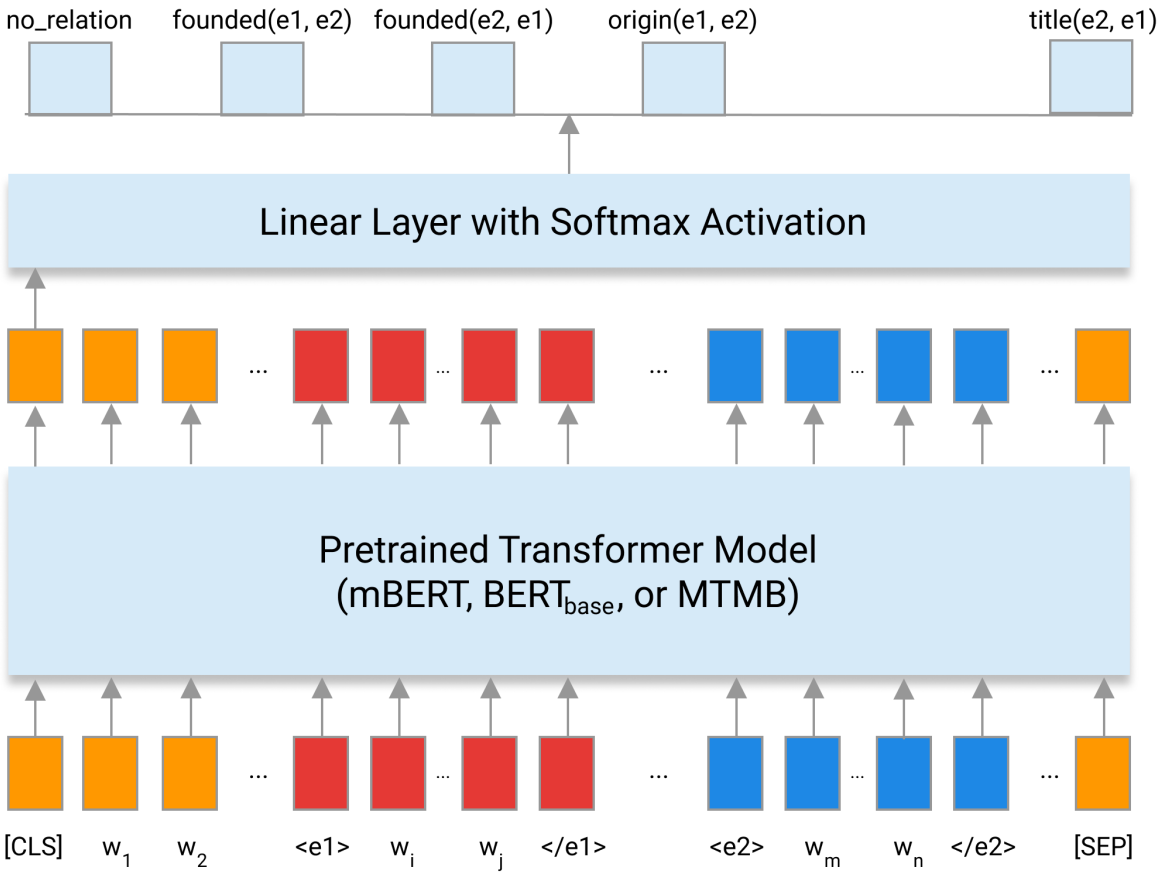


Figure 5.1. The architecture of fine-tuning with transformers.

The architecture of fine-tuning is illustrated in Figure 5.1. All pretrained transformer models including mBERT, $\text{BERT}_{\text{base}}$, and Matching the Multilingual Blanks (MTMB) are finetuned using the same flow. The hyperparameters are selected by the F1-score of the KBP-37 validation set. 1e-3, 1e-4, 3e-4, 1e-5, 3e-5, and 1e-6 learning rates; 0, 0.01, and 0.1 weight decay values with the AdamW [71], Adam [72], and SGD optimizers evaluated with PyTorch [73] and HuggingFace's Transformers [70] frameworks. The best hyperparameter value is determined as the $3e-5$ learning rate and 0.1 weight decay with the AdamW optimizer.

5.3. Matching the Multilingual Blanks

The usual process for fine-tuning a transformer model for a classification task is as follows. First, the most suitable pretrained transformer model is found for the task. Then, it is finetuned by adding a classifier layer on top of the pretrained transformer model, and all parameters are finetuned with a specific dataset. However, Soares and co-workers [6] propose a specific pretraining objective for relation classification called Matching the Blanks (MTB), and it achieves state-of-the-art performances in four diverse relation classification datasets.

MTB and our approach, MTMB, take benefit of the distant supervision scheme [63]. Distant supervision in relation classification assumes that if a sentence contains two entities, which has a relation based on a knowledge base, the sentence contains a relation between these entities. For example, "Boğaziçi University" and "Istanbul" have a "location" relation in Wikidata. Distant supervision scheme assumes that all sentences including these entities have a "location" relation. However, this assumption may not hold all the time, for example, the sentence "Boğaziçi University football team won a game played in Atatürk Olympic Stadium in İstanbul" contains both entities but the information about "location" relation between them does not exist in the sentence. Still, the distant supervision scheme can help to create weakly-labeled datasets to improve the performance of the models.

MTB collects distantly supervised pairs from English Wikipedia. They use an off-the-shelf entity linking system to map entities on Wikipedia to a knowledge base. After collecting a large number of sentences containing a relation, they perform a pretraining objective called Matching the Blanks, aiming for a fake task to find sentence pairs with the same relation. They state that the relation classification datasets finetuned over MTB have a better performance than $\text{BERT}_{\text{large}}$.

Sentences

S_{en}	In the 3rd century, E2 wrote his "E1" and other exegetical and theological works while living in Caesarea.
S_{es}	Este es un palimpsesto de una copia de la obra de E2 llamada la E1.
S_{tr}	İreneyus ve E2 gibi kilise babalarının metinlerinde aktarılanlara göre esasen E3li olan Marcellina, Anicetus döneminde Roma'ya göç etmiş ve çok sayıda takipçi toplamıştır.

Entities

E1	*Q839739* (Hexapla, Hexapla, Hexapla)
E2	*Q170472* (Origen, Orígenes, Origenes)
E3	*Q87* (Alexandria, Alejandría, İskenderiye)

Relations

(E1, E2)	*P50* (Author)
(E2, E3)	*P19* (Place of Birth)

Pairs

Positive	(S_{en}, S_{es})
Negative	(S_{en}, S_{tr})

Figure 5.2. Sample positive and negative pairs constructed from RELX-Distant. Entities and relations are linked with their Wikidata ID's (shown in italic) and words in parentheses in entities represent English, Spanish, and Turkish correspondence.

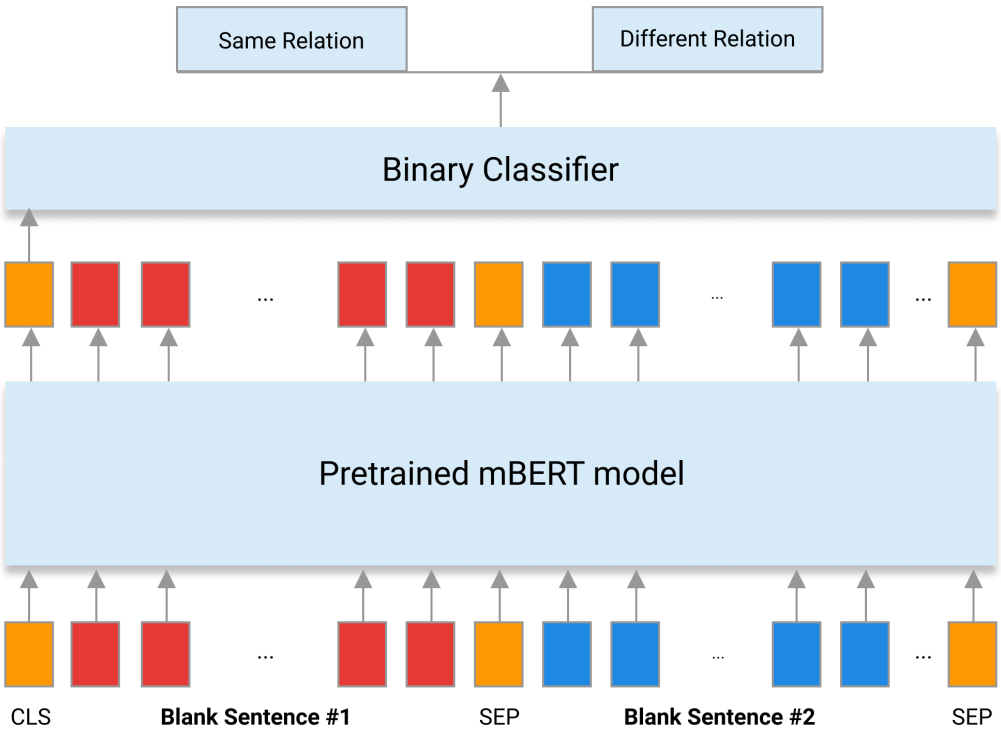
We aim to expand this approach to multilingual settings without using any off-the-shelf linking system to include low-resource languages without annotated data and pretrained NLP tools. The process to create a distantly supervised and weakly-labeled multilingual relation extraction dataset, RELX-Distant, is summarized in Chapter 4. We create pairs from RELX-Distant for Matching the Multilingual Blanks (MTMB) objective.

The final model is pretrained over mBERT with two objectives. The first one is Masked Language Model, which is proposed in [9]. The second one is Matching the Multilingual Blanks (MTMB). Similar to the English relation classification approach in [6], positive and negative sentence pairs from different languages are constructed from RELX-Distant for MTMB. The objective aims to find whether a pair, containing a sentence in the English language and another sentence in a different language, has the same relation or different relations. We show that a model with this objective learns relation representation in different languages better.

Positive sentence pairs are selected to share the same entities, which result in having the same relation by the distant supervision scheme. (S_{en}, S_{es}) in Figure 5.2 is a positive pair because both sentences include the E1 (Hexapla) and E2 (Origen) entities that have the P50 (Author) relation.

Negative pairs are constructed in a way that English and non-English sentences include entities with different relations. As positive pairs contain sentences with the same entities and relations, they tend to have similar contexts. In order to make the model distinction in relation, not in context, we use *strong* negative pairs, following [6]. While the relations in the two sentences are different, one of the entities in the sentences is common in strong negative pairs. For example, in Figure 5.2, English and Turkish sentence pair, (S_{en}, S_{tr}), is a strong negative pair as both of them has the entity E2 (Origen), while English sentence has P50 (Author) relation, and Turkish sentence has P19 (Place of Birth) relation.

In order to capture relation patterns and avoid memorization of entities during pretraining, each entity in sentence pairs is replaced by a special token, called [BLANK], with 0.7 probability. By following the mentioned flow, we gather 20 million pairs from RELX-Distant for the MTMB objective. These pairs are uniformly distributed between negative and positive classes, as well as the languages in RELX-Distant.



Example Sentence
Pair with Masking

Sentence #1	In the 3rd century, Origen wrote his Hexapla
Blank Sentence #1	In the 3rd century, [BLANK] wrote his [BLANK]
Sentence #2	Esta es la obra de Orígenes llamada Hexapla
Blank Sentence #2	Esta es la obra de [BLANK] llamada Hexapla

Figure 5.3. The flow of Matching the Multilingual Blank.

mBERT pretraining with MTMB objective is summarized in Figure 5.3. As an example, we illustrate a positive pair, an English sentence, Sentence #1 with Origen and Hexapla entities, and a Spanish sentence, Sentence #2 with Orígenes and Hexapla entities. As we discussed above, entities are replaced by [BLANK] special token, which is learned from scratch during pretraining, with 0.7 probability (which results in re-

placing all entities except the Hexapla entity in Sentence #2). Finally, this pair is separated by the [SEP] token and given to an already pretrained mBERT model. The representation of the [CLS] token is fed into a binary classifier to detect whether this pair contains the same relation or not. As a result, mBERT is further pretrained by 20 million pairs with 4 different language pairs (English - French, German, Spanish, Turkish).

Implementation details are similar to the fine-tuning process described for mBERT and $\text{BERT}_{\text{base}}$. However, we further pretrain the mBERT model with two objectives before fine-tuning over multi-way relation classification. These objectives are Masked Language Model [9] and Matching the Multilingual Blanks. While fine-tuning of mBERT or $\text{BERT}_{\text{base}}$ over a relation classification dataset is comparatively inexpensive (each epoch takes around 10 minutes on a Tesla V100), one epoch of MTMB takes around 10 days on a Tesla V100 GPU because of the high number of pairs. We release the weights of MTMB publicly in our GitHub repository and HuggingFace's model hub.

5.4. Evaluation Metric

Following [59], our evaluation metric is (18+1)-way evaluation with directionality taken into account. Each relation except no_relation is evaluated separately by a micro average of F1 scores of both directions. Eventually, the macro average of micro-averaged F1 scores of each relation is presented as the final score.

6. RESULTS

We compare English relation classification results by KBP-37 and the cross-lingual results by RELX. As Dodge and co-workers [74] present, the results of BERT models may vary with just different seeds. The results are reported by taking the average scores of 10 runs to decrease the variance.

After presenting the results, we discuss our findings with error analysis. We show that the MTMB model is more effective on both mono-lingual relation classification and cross-lingual relation classification tasks than $\text{BERT}_{\text{base}}$ and mBERT models, consecutively. Furthermore, we show that MTMB's performance gain over mBERT is higher when the training data in the source domain is scarce.

6.1. KBP-37

English relation classification results on KBP-37 development and test sets are presented in Table 6.1. It can be seen that our proposed model outperforms both mBERT and $\text{BERT}_{\text{base}}$ models in the English language by 1.6% and 1.1% point absolute improvement, consecutively. Even MTMB is a multilingual transformer model, it outperforms the $\text{BERT}_{\text{base}}$ model which is pretrained explicitly as an English transformer model with a more diverse English corpus including BookCorpus [53] and more suitable tokens for the English language. We also report that $\text{BERT}_{\text{base}}$ outperforms the mBERT model as expected due to different pretraining schemes between them even though they have similar complexity in terms of model size. We perform randomization tests [75] and show that Matching the Multilingual Blanks significantly ($p - value < 0.05$) outperforms both mBERT and $\text{BERT}_{\text{base}}$.

Table 6.1. F1 scores of our models compared to the state-of-the-art models on the development and test sets of KBP-37 (*English*).

Model	**Development**	**Test**
$\text{BERT}_{\text{large}}$ from [6]	69.5	68.3
MTB from [6]	<u>70.3</u>	<u>69.3</u>
$\text{BERT}_{\text{base}}$	66.0	65.4
mBERT	65.5	64.9
MTMB	<u>66.8</u>	<u>66.5</u>

$\text{BERT}_{\text{large}}$ and MTB from [6] outperform our approach as well as proposed baselines, $\text{BERT}_{\text{base}}$ and mBERT. We believe that the main difference stems from the model size and the languages that the transformer models are pretrained on. Both $\text{BERT}_{\text{large}}$ and MTB have around 340 million parameters, while our three models have around 110 million parameters with fewer attention heads and layers. The difference between model complexity can be observed better between $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ models. Both of them are pretrained in the English language, with the same corpora and same vocabulary but still $\text{BERT}_{\text{large}}$ outperforms $\text{BERT}_{\text{base}}$ with a significant margin.

6.2. RELX

Cross-lingual relation classification results on RELX development and test sets are presented in Table 6.2. It can be seen that Matching the Multilingual Blanks improves mBERT for five languages, including RELX-English. According to randomization tests, Matching the Multilingual Blanks significantly ($p-value < 0.05$) outperforms mBERT both on English (monolingual) and cross-lingual relation classification.

Table 6.2. F1 scores of mBERT and MTMB evaluated on RELX.

Model	**English**	**French**	**German**	**Spanish**	**Turkish**
mBERT	61.8	58.3	57.5	57.9	55.8
MTMB	63.6	59.9	59.9	62.4	56.2

We conduct further experiments by varying the size of the training data by 10%, 20% , 50%, and 100%. The results are illustrated in Figure 6.1. It shows that the performance gain (the absolute difference between F1 scores) of MTMB over mBERT is higher when the amount of training data is lower. Furthermore, MTMB, finetuned with 20% of the training data is able to obtain the same performance of mBERT with full training for the Spanish language. For the other evaluated languages (except Turkish), half of the training data are sufficient for MTMB to reach the same level as full training with mBERT. Therefore, human annotation costs in the source language can be reduced thanks to MTMB.

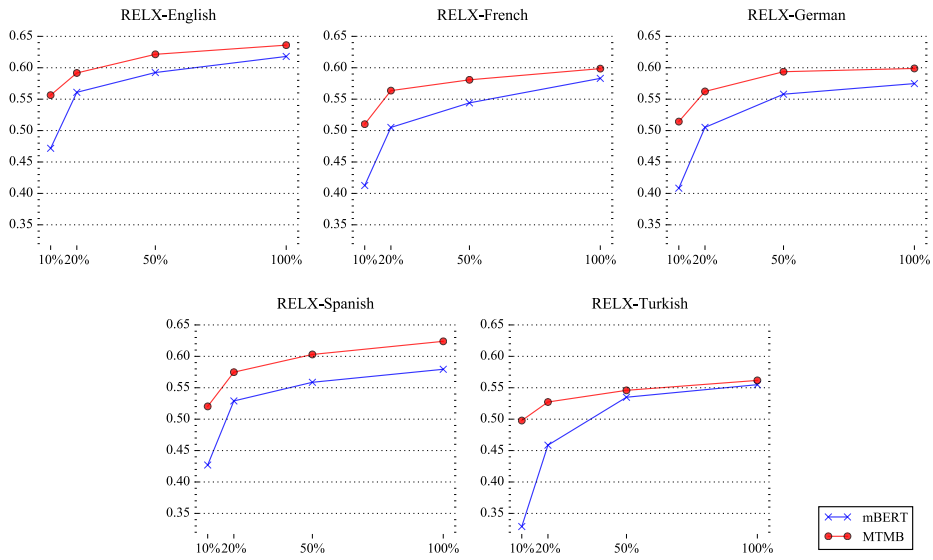


Figure 6.1. Cross-lingual relation classification performance (F1 score *y-axis*) of mBERT and MTMB with varying amounts of training data (*x-axis*).

6.3. Error Analysis

As it can be seen in Table 6.2, the Spanish language performs best cross-lingual performance among other languages. We see that these results are on par with prior studies on other cross-lingual tasks like natural language inference and question answering [11] that also report higher performance for Spanish. We also see that the worse cross-lingual performance is happened for Turkish. In [56], it is stated that the performance of the multilingual transformer model is affected by word order, and the best performing languages in the cross-lingual setup are the ones in the target languages that have the highest typological similarity to the source language. In order to investigate the effects of language similarity in our work, we compare the source language (English) to the target languages (French, German, Spanish, Turkish) by a subset related to grammatical order out of the World Atlas of Language Structures (WALS) [57] features[2] as in [56]. In terms of these features, Turkish is the least similar language to English among the target languages in RELX. We see that our results support the claim presented in [56].

We conduct further error analysis to show future directions. It reveals that an important portion (120) of 176 mispredicted sentences in RELX-English have also mispredicted in all target languages. Moreover, relation classes with less than 600 samples in the English training data have a 60% more error rate among these common errors, implying that increasing the sample size may help in all languages.

RELX and KBP-37 include directional relations except for the no_relation class. We analyze relation direction errors which is the predicted relation (e.g. founded(e1, e2)) is the same as the gold class (e.g. founded(e2, e1)), while the predicted direction is wrong. We observe 79 relation direction errors for Turkish, while other languages have less than 15. While the source language (English) has prepositions and Subject-Verb-Object (SVO) word order, the Turkish language has postpositions and usually a

[2] *81A*: Order of Subject, Object and Verb, *85A*: Order of Adposition and Noun Phrase, *86A*: Order of Genitive and Noun, *87A*: Order of Adjective and Noun, *88A*: Order of Demonstrative and Noun, *89A*: Order of Numeral and Noun

Subject-Object-Verb (SOV) word order. The differences in grammatical ordering features between Turkish and English are potential roots for direction errors, as observed in [56]. Finally, we do not observe any notable difference in errors across languages respecting the sentence length.

7. CONCLUSION

In this thesis, we propose new datasets and models for the cross-lingual relation classification task. First, we explained relation classification and cross-lingual tasks in NLP, giving the current status of cross-lingual works in the domain, and the motivations behind them. Pretrained transformer models reshaped not only monolingual studies in NLP but also the approach to cross-lingual NLP. Finally, we explained cross-lingual works with transformer models and prior architectures.

We proposed two types of contributions to the cross-lingual relation classification task: datasets and models.

Two publicly datasets are released:

(i) RELX: Human annotated, cross-lingual relation classification benchmark for English, French, German, Spanish, and Turkish languages. This dataset is constructed by human translation of the already-existing English relation classification dataset, KBP-37. Therefore, this dataset also includes parallel sentences. In total, 2,510 sentences in five languages with 37 relations are proposed.

(ii) RELX-Distant: Weakly-labeled, large-scale multilingual relation classification dataset. This dataset is collected from Wikipedia by linking entities by their hyperlinks to Wikidata. Relations are constructed from Wikidata via distant supervision. In total, 2,575,582 sentences in five languages with 24 relations are proposed.

For the modeling, a new pretraining objective called Matching the Multilingual Blanks (MTMB) is introduced and compared with baseline models, mBERT and $\text{BERT}_{\text{base}}$.

(i) Matching the Multilingual Blanks (MTMB) objective is successfully included in the pretraining objective. The public checkpoint of mBERT is further pretrained with the MTMB objective on the RELX-Distant dataset.

(ii) $\text{BERT}_{\text{base}}$, mBERT, and MTMB transformer models are finetuned with the KBP-37 dataset and evaluated on KBP-37 and RELX. To the best of our knowledge, we presented the first transformer approach to the cross-lingual relation classification task.

In our experiments, we showed that MTMB significantly outperforms mBERT (by 1.6% point absolute improvement) and BERT (by 1.1% point absolute improvement) baselines on the English dataset (KBP-37), and significantly improves the F1 score of mBERT (by a 2.14% point absolute improvement for the average of 5 languages) on the cross-lingual dataset (RELX). When the training data are lower in the source language, the improvement of MTMB is higher. MTMB's absolute improvement of the average of 5 languages is 10.58%, relative improvement is 25.5% over mBERT.

In error analysis, we observed that target languages which similar to the source language, typologically, achieve better results for cross-lingual relation classification. While the performance of MTMB in Spanish (a target language without any training data and typologically similar to English) is comparable to English (a source language with training data), it obtains the lowest F1 score in Turkish.

Furthermore, we observed that directional errors mostly occur in the Turkish language. While there are 79 relation direction errors (prediction relation type is correct, but predicted direction is incorrect) for Turkish, there are less than 15 for other languages. We believe that these errors mostly occur in Turkish because Turkish is the least similar language to English according to the WALS features, word order, adpositions.

For future work, we plan to expand RELX-Distant to all available languages on Wikipedia. Furthermore, we might investigate different entity linking and distant-

supervision schemes to improve RELX-Distant. We also would like to explore a different approach for the pretraining objective (specific to multilingual relation classification) and investigate the directional errors by increasing the number of languages in RELX and RELX-Distant. Finally, we plan to observe the effects of the MTMB transformer model in different cross-lingual tasks such as named entity recognition, natural language inference, and question answering.

REFERENCES

1. Xu, K., S. Reddy, Y. Feng, S. Huang and D. Zhao, "Question Answering on Freebase via Relation Extraction and Textual Evidence", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2326–2336, Association for Computational Linguistics, Berlin, Germany, Aug. 2016.

2. Kambhatla, N., "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction", *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 178–181, Association for Computational Linguistics, Barcelona, Spain, Jul. 2004.

3. Mikolov, T., K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", *arXiv preprint arXiv:1301.3781*, 2013.

4. Zhang, D. and D. Wang, "Relation Classification via Recurrent Neural Network", *arXiv preprint arXiv:1508.01006*, 2015.

5. Xu, Y., L. Mou, G. Li, Y. Chen, H. Peng and Z. Jin, "Classifying Relations via Long Short Term Memory Networks Along Shortest Dependency Paths", *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1785–1794, 2015.

6. Soares, L. B., N. FitzGerald, J. Ling and T. Kwiatkowski, "Matching the Blanks: Distributional Similarity for Relation Learning", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2895–2905, 2019.

7. Upadhyay, S., *Exploiting Cross-lingual Representations for Natural Language Processing*, Ph.D. Thesis, University of Pennsylvania, 2019.

8. Indurkhya, N., "Emerging Directions in Predictive Text Mining", *WIREs Data*

Mining and Knowledge Discovery, Vol. 5, No. 4, pp. 155–164, 2015.

9. Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, Jun. 2019.

10. Conneau, A. and G. Lample, "Cross-lingual Language Model Pretraining", *Advances in Neural Information Processing Systems*, pp. 7059–7069, 2019.

11. Artetxe, M., S. Ruder and D. Yogatama, "On the Cross-lingual Transferability of Monolingual Representations", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Association for Computational Linguistics, Online, Jul. 2020.

12. Liu, J., Y. Lin, Z. Liu and M. Sun, "XQA: A Cross-lingual Open-domain Question Answering Dataset", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2358–2368, Association for Computational Linguistics, Florence, Italy, Jul. 2019.

13. Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Association for Computational Linguistics, Online, Jul. 2020.

14. Wu, S. and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 833–844, 2019.

15. Kim, S. and G. G. Lee, "A Graph-based Cross-lingual Projection Approach for Weakly Supervised Relation Extraction", *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 48–53, Association for Computational Linguistics, Jeju Island, Korea, Jul. 2012.

16. Faruqui, M. and S. Kumar, "Multilingual Open Relation Extraction Using Cross-lingual Projection", *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1351–1356, Association for Computational Linguistics, Denver, Colorado, May–Jun. 2015.

17. Ni, J. and R. Florian, "Neural Cross-Lingual Relation Extraction Based on Bilingual Word Embedding Mapping", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 399–409, Association for Computational Linguistics, Hong Kong, China, Nov. 2019.

18. Köksal, A. and A. Özgür, "The RELX Dataset and Matching the Multilingual Blanks for Cross-Lingual Relation Classification", *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 340–350, Association for Computational Linguistics, Online, Nov. 2020.

19. Weizenbaum, J., "ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine", *Communications of the ACM*, Vol. 9, No. 1, pp. 36–45, 1966.

20. Lesk, M., "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", *Proceedings of the 5th Annual International Conference on Systems Documentation*, pp. 24–26, 1986.

21. Pollard, C. and I. A. Sag, *Head-driven Phrase Structure Grammar*, University of Chicago Press, 1994.

22. Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, R. L. Mercer and P. Roossin, "A Statistical Approach to Language Translation", *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*, 1988.

23. Nigam, K., J. Lafferty and A. McCallum, "Using Maximum Entropy for Text Classification", *IJCAI-99 Workshop on Machine Learning for Information Filtering*, Vol. 1, pp. 61–67, Stockholm, Sweden, 1999.

24. Zhou, G. and J. Su, "Named Entity Recognition Using an HMM-based Chunk Tagger", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 473–480, 2002.

25. Giménez, J. and L. Marquez, "Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited", *Recent Advances in Natural Language Processing III*, pp. 153–162, 2004.

26. Zhang, Y., V. Zhong, D. Chen, G. Angeli and C. D. Manning, "Position-aware Attention and Supervised Data Improve Slot Filling", *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pp. 35–45, 2017.

27. Pennington, J., R. Socher and C. D. Manning, "Glove: Global Vectors for Word Representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

28. Conneau, A., R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk and V. Stoyanov, "XNLI: Evaluating Cross-lingual Sentence Representations", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Association for Computational Linguistics, Brussels, Belgium, Oct.-Nov. 2018.

29. Xie, J., Z. Yang, G. Neubig, N. A. Smith and J. Carbonell, "Neural Cross-Lingual Named Entity Recognition with Minimal Resources", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 369–379, Association for Computational Linguistics, Brussels, Belgium, Oct.-Nov. 2018.

30. Bahdanau, D., K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *arXiv preprint arXiv:1409.0473*, 2014.

31. Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation", *arXiv preprint arXiv:1609.08144*, 2016.

32. Duh, K., A. Fujino and M. Nagata, "Is Machine Translation Ripe for Cross-lingual Sentiment Classification?", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 429–433, 2011.

33. Banko, M. and O. Etzioni, "The Tradeoffs Between Open and Traditional Relation Extraction", *Proceedings of ACL-08: HLT*, pp. 28–36, Association for Computational Linguistics, Columbus, Ohio, Jun. 2008.

34. Augenstein, I., D. Maynard and F. Ciravegna, "Distantly Supervised Web Relation Extraction for Knowledge Base Population", *Semantic Web*, Vol. 7, No. 4, pp. 335–349, 2016.

35. Zeng, D., K. Liu, S. Lai, G. Zhou and J. Zhao, "Relation Classification via Convolutional Deep Neural Network", *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2335–2344, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, Aug. 2014.

36. Nguyen, T. H. and R. Grishman, "Relation Extraction: Perspective from Convolutional Neural Networks", *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 39–48, Association for Computational Linguistics, Denver, Colorado, Jun. 2015.

37. Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, "Deep Contextualized Word Representations", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, Association for Computational Linguistics, New Orleans, Louisiana, Jun. 2018.

38. Kim, S., M. Jeong, J. Lee and G. G. Lee, "A Cross-lingual Annotation Projection Approach for Relation Detection", *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 564–571, Coling 2010 Organizing Committee, Beijing, China, Aug. 2010.

39. Zou, B., Z. Xu, Y. Hong and G. Zhou, "Adversarial Feature Adaptation for Cross-lingual Relation Classification", *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 437–448, Association for Computational Linguistics, Santa Fe, New Mexico, USA, Aug. 2018.

40. Abdou, M., C. Sas, R. Aralikatte, I. Augenstein and A. Søgaard, "X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension", *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 265–274, 2019.

41. Nothman, J., N. Ringland, W. Radford, T. Murphy and J. R. Curran, "Learning Multilingual Named Entity Recognition from Wikipedia", *Artificial Intelligence*, Vol. 194, pp. 151–175, 2013.

42. Chinchor, N. and E. Marsh, "Muc-7 Information Extraction Task Definition", *Proceeding of the Seventh Message Understanding Conference (MUC-7), Appendices*,

pp. 359–367, 1998.

43. Blaschke, C. and A. Valencia, "The Frame-based Module of the SUISEKI Information Extraction System", *IEEE Intelligent Systems*, Vol. 17, No. 2, pp. 14–20, 2002.

44. Zelenko, D., C. Aone and A. Richardella, "Kernel Methods for Relation Extraction", *Journal of Machine Learning Research*, Vol. 3, No. Feb, pp. 1083–1106, 2003.

45. Miwa, M. and M. Bansal, "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1105–1116, 2016.

46. Joshi, M., D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer and O. Levy, "Spanbert: Improving Pre-training by Representing and Predicting Spans", *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 64–77, 2020.

47. "Wikipedia: Size of Wikipedia", `https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia`, accessed in May 2021.

48. Simonyan, K. and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition", *arXiv preprint arXiv:1409.1556*, 2014.

49. He, K., X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

50. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All You Need", *arXiv preprint arXiv:1706.03762*, 2017.

51. Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language Models are Unsupervised Multitask Learners", *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.

52. Sennrich, R., B. Haddow and A. Birch, "Neural Machine Translation of Rare Words with Subword Units", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016.

53. Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba and S. Fidler, "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 19–27, 2015.

54. Wang, A., A. Singh, J. Michael, F. Hill, O. Levy and S. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

55. "Wikipedia: List of Wikipedias", `https://meta.wikimedia.org/wiki/List_of_Wikipedias`, accessed in June 2021.

56. Pires, T., E. Schlinger and D. Garrette, "How Multilingual is Multilingual BERT?", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, 2019.

57. Dryer, M. S. and M. Haspelmath (Editors), *WALS Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

58. Walker, C., S. Strassel, J. Medero and K. Maeda, "ACE 2005 Multilingual Training Corpus", Philadelphia: Linguistic Data Consortium.

59. Hendrickx, I., S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano and S. Szpakowicz, "SemEval-2010 Task 8: Multi-Way

Classification of Semantic Relations between Pairs of Nominals", *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 33–38, Association for Computational Linguistics, Uppsala, Sweden, Jul. 2010.

60. Ni, J. and R. Florian, "Neural Cross-Lingual Relation Extraction Based on Bilingual Word Embedding Mapping", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 399–409, Association for Computational Linguistics, Hong Kong, China, Nov. 2019.

61. Vrandečić, D. and M. Krötzsch, "Wikidata: A Free Collaborative Knowledgebase", *Communications of the ACM*, Vol. 57, No. 10, pp. 78–85, 2014.

62. Tanon, T. P., G. Weikum and F. Suchanek, "Yago 4: A Reason-able Knowledge Base", *European Semantic Web Conference*, pp. 583–596, Springer, 2020.

63. Mintz, M., S. Bills, R. Snow and D. Jurafsky, "Distant Supervision for Relation Extraction without Labeled Data", *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011, 2009.

64. Bollacker, K., C. Evans, P. Paritosh, T. Sturge and J. Taylor, "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge", *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250, 2008.

65. Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, "Dbpedia: A Nucleus for a Web of Open Data", *The Semantic Web*, pp. 722–735, Springer, 2007.

66. Shin, J., S. Wu, F. Wang, C. De Sa, C. Zhang and C. Ré, "Incremental Knowledge Base Construction Using Deepdive", *Proceedings of the VLDB Endowment*

International Conference on Very Large Data Bases, Vol. 8, p. 1310, NIH Public Access, 2015.

67. Kadlec, R., O. Bajgar and J. Kleindienst, "Knowledge Base Completion: Baselines Strike Back", *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 69–74, 2017.

68. Riedel, S., L. Yao and A. McCallum, "Modeling Relations and Their Mentions without Labeled Text", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163, Springer, 2010.

69. Honnibal, M. and I. Montani, "Spacy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing", , 2017, to appear.

70. Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush, "HuggingFace's Transformers: State-of-the-art Natural Language Processing", *ArXiv*, Vol. abs/1910.03771, 2019.

71. Loshchilov, I. and F. Hutter, "Decoupled Weight Decay Regularization", *International Conference on Learning Representations*, 2018.

72. Kingma, D. P. and J. Ba, "Adam: A Method for Stochastic Optimization", Y. Bengio and Y. LeCun (Editors), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

73. Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An Imperative Style, High-performance Deep Learning Library", *Advances in Neural Information Processing*

Systems, pp. 8026–8037, 2019.

74. Dodge, J., G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi and N. Smith, "Fine-tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping", *arXiv preprint arXiv:2002.06305*, 2020.

75. Yeh, A., "More Accurate Tests for the Statistical Significance of Result Differences", *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, 2000.