Project Report

Incorporating Deep Contextual Embeddings in Neural Relation Extraction for Bangla

Taufiquzzaman Peyash, M.J. Darad, Fahad Muntasir
[taufiquzzaman.peyash, m.j.darad, mohammad.muntasir]@northsouth.edu

Advisor: Dr. Nabeel Mohammed

North South University

Abstract

Being the seventh largest language in the world information extraction from Bangla language is still a very tough problem. We propose to build an application which will generate question along with answers from given Bangla paragraph. This will not only help the young students who just joined the schools but also help the deprived, physically challenged people. Performing this task is still difficult since it belongs to Bangla Language's domain. Solv- ing this problem will require implementation of Natural Language Processing (NLP) task such as Information Extraction from Bangla Language. Information Extraction being the core part of our project, we plan to achieve the goal by using state-of-the-art architectures based on Deep Learning such as BERT, ELMo etc. Information Extract on the other hand constists of two subtasks: NER- Named Entity Recognition and RE- Relation Extraction. NER, RE, DL,BERT, ELMo, NLP.

Contents

1

Introduction

1.1 Background

Bangladesh has brought enormous success in the education sector and an increased literacy rate is clearly showing it. By raising awareness and integrating technology literacy rate can be increased more significantly. According to UNESCO, the adult literacy rate has reached 72.89

1.2 Motivation

The government of Bangladesh is focusing on integrating technology in the field of the education sector. Bangladesh has become a role model in the asia pacific region for this. We also want to contribute in the era of digitization by deploying educational systems which will be helping students for a better understanding of the language. There hasn't been any work on Relation Extraction in Bangla language. This inspired us to explore this unsolved problem and contribute to the research happening in the domain of Bangla Language.

1.3 Problem Statement

Our proposed project will extract relation from Bangla text which will further help te get useful insights from Bangla paragraph. This system can be used in academic, legal and many other forms of documents. Given an input sequence of Bangla characters/tokens, the job of our model is to find out pairs of entities that may hold a relationship where an entity is defined to be a thing with distinct and independent existence. To illustrate the idea of this application : "কুইরুক ইয়াগি যা ভেঁড়ার লেজের চর্বি অনেক সময় কেবাব এবং মাংসের পদ তৈরীতে ব্যবহৃত হয়" Here the entities are "কুইরুক ইয়াগি", "ভেঁড়ার লেজের চর্বি", "কেবাব এবং মাংসের পদ"
The relationships between "কুইরুক ইয়াগি" and "কেবাব এবং মাংসের পদ" is used_for

1.4 Project Study

In the field of Natural Language Processing, Relation Extraction is a widely researched topic that enables the machine to comprehend, translate and control natural languages. But when it comes to Bangla Language it is not a trivial task to do even in a small domain because Bangla is a free word language and does not have any capitalization information which plays a major role in information extraction. The task of information extraction from unstructured data consists of two-sub-tasks. The first part is Named Entity Recognition and the second one is Relation Extraction. Relation extraction plays a vital role in retrieve structured information from unstructured text by extracting the semantic relationships between two entities (person, location, organization, and object).

1.5 Significance of the study

- Educational Impact: Since we plan to Extract Relationships from different kinds of textual data, this will help the education sector. Students from junior classes will understand the meaning of different relationships between different entities with the help of our application.
- Education for Physically Challenged Students: Physically challenged students, who need special attention for their studies will take the help of our application to better understand any lesson.
- Impact on Women Empowerment: Even though our government is working hard for women's empowerment, this is still a difficult job to deal with. Many deprived women will be able to learn and test themselves with the help of the idea we are proposing.

2

Literature Review

Due to the fact that most of the problem in real-world are nonlinear in nature, we need to come with a solution which can map these complex problems to a nonlinear function to solve it in a tractable time. Neural Networks, in particular, are very good at learning non-linear functions, for which in recent times almost all of the solutions with state-of-the-art performance are based on Neural Networks. Many problems from different domains such as – Computer Vision, Pattern Recognition, Natural Language Processing (NLP), etc. has surpassed human-level performance due to the use of Artificial Neural Networks. In this review, we will only focus on the tasks related to NLP.

At the early age of Natural Language Processing, hand-crafted features were used in different NLP tasks, but with these hand-crafted features, the problem was the curse of dimensionality. A large number of hand-crafted features were used to present a sparse representation of linguistic information. In word2vec, Mikolov (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) proposed a CBOW and Skip Gram models to calculate the distributed representation of words. The objective of this (ibid.) paper is to compute the conditional probability of a target word given the context words in predefined window size using Neural Networks. GloVe or Global Vectors for Word Representation is another way of computing word embeddings (Pennington, Socher, & Manning, 2014), which exploits the advantages of existing two major model families – global matrix factorization and local context window methods. In this manner, a dense representation of words is computing a lower dimension containing syntactic and semantic information.

Yoon Kim (Kim, 2014) has conducted a series of experiments with Convolutional Neural Networks (CNN) trained on top of pre-trained word vectors for sentence-level classification tasks. The model proposed in (ibid.) achieved state-of-the-art performance on four out of 7 tasks, including Sentiment Analysis and Question Classification. In this paper author, Yoon Kim showed without fine-tuning the proposed model exploiting the pre-trained word vectors can achieve excellent results on multiple benchmarks, which suggests that the static word vectors play a role of universal feature extractor, and fine-tuning for a task-specific problem can lead to further improvements.

The task of Information Extraction (IE) is composed of two different sub-tasks – Named Entity Recognition and Relation Extraction, as mentioned previously. An End-to-End relation extraction model for both recognition of the named entities and extraction of relations between those entities were introduced by Miwa and Bansal (Miwa & Bansal, 2016), which uses two stacked networks, a Bidirectional LSTM for entity recognition and a Tree-based LSTM for extracting the relation that links the recognized entities. There are very few models that are End-to-End. Most of the works have been done separately, either for the Named Entity Recognition (NER) or Relation Extraction (RE) for the entities that are already present in the Knowledge Base (KB).

2.1 Neural Networks

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated. Neural networks help us cluster and classify. You can think

of them as a clustering and classification layer on top of the data you store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on. (Neural networks can also extract features that are fed to other algorithms for clustering and classification; so you can think of deep neural networks as components of larger machine-learning applications involving algorithms for reinforcement learning, classification and regression.)

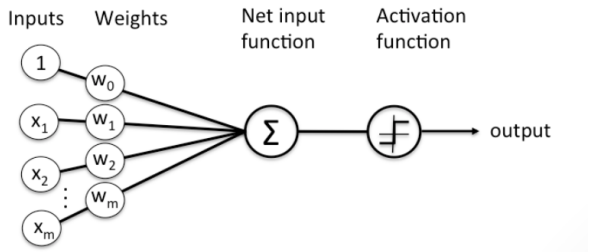

Figure 2.1: Neural Network

2.2 Recurrent Neural Networks

They are networks with loops in them, allowing information to persist. In the above diagram, a chunk of the

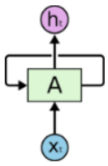

Figure 2.2: A recurrent neural network

neural network, A, looks at some input xt and outputs a value ht. A loop allows information to be passed from one step of the network to the next. A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor. This chain-like nature reveals that recurrent neural networks are intimately related to sequences and lists.

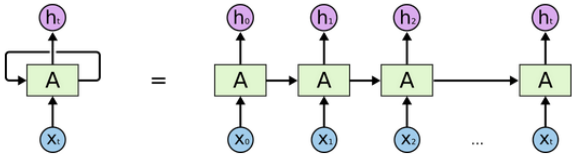

Figure 2.3: An unrolled recurrent neural network

2.3 Long-Short Term Memory

Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They were introduced in 1997 (Hochreiter & Schmidhuber, 1997) and were refined and popularized by many people in the following work. They work tremendously well on a large variety of problems and are now widely used.

LSTMs are explicitly designed to avoid long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

All recurrent neural networks have the form of a chain of repeating modules of the neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

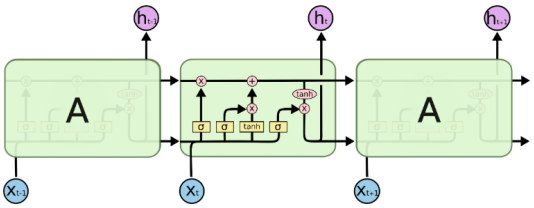

Figure 2.4: The repeating module in an LSTM contains four interacting layers.

2.4 Embeddings from Language Models(ELMO)

ELMo is a deep contextualized word representation (Peters et al., 2018) that models both complex characteristics of word use (e.g., syntax and semantics), and how these uses vary across linguistic contexts (i.e., to model polysemy). These word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus. They can be easily added to existing models and significantly improve the state of the art across a broad range of challenging NLP problems, including question answering, textual entailment and sentiment analysis. The architecture above uses a character-level

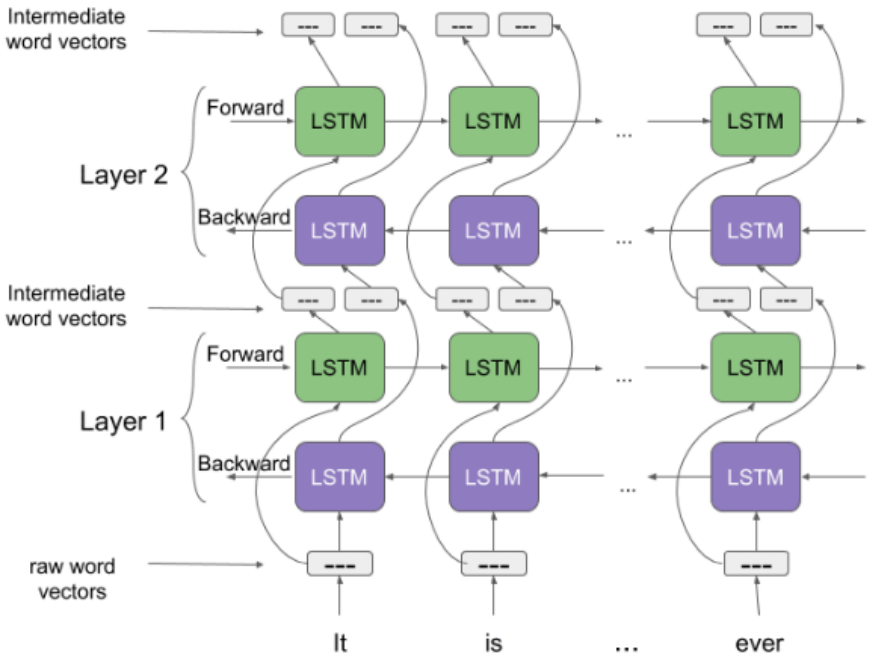

Figure 2.5: Architecture of ELMo

convolutional neural network (CNN) (LeCun, Bengio, et al., 1995) to represent words of a text string into raw word vectors. These raw word vectors act as inputs to the first layer of biLM. The forward pass contains information about a certain word and the context (other words) before that word. The backward pass contains information about the word and the context after it This pair of information, from the forward and backward pass, forms the intermediate word vectors. These intermediate word vectors are fed into the next layer of

biLM. The final representation (ELMo) is the weighted sum of the raw word vectors and the 2 intermediate word vectors.

2.5 Relation Extraction

Data set preparation is one of the biggest challenges in supervised learning. Mintz et al. proposed a distant supervision approach (Mintz, Bills, Snow, & Jurafsky, 2009) for automatically generating large amounts of training data to avoid the tedious work of manually labeling the dataset. In this approach, they aligned documents with Knowledge Base (KB) with an assumption that if a relationship exists in a pair of the entity in the current KB, then all the documents containing that entity pair will have that relationship. But the problem with this assumption is that there might be some documents which may contain those entity pairs but with different relation.

Another method (Riedel, Yao, & McCallum, 2010) proposed a more relaxed version of this distant supervision approach by formulating the problem as a multi-instance learning problem to alleviate this problem and reduce the noise. An End-to-End Convolutional Neural Network-based architecture was proposed (Liu, Sun, Chao, & Che, 2013) to extract relation between entities, which is one of the earliest models that does not require any hand-craft features, rather it encodes the input sentence using word vectors and lexical features, and outputs a probability distribution over all the relation classes using a softmax output layer.

Zeng et al. proposed a Piecewise Convolutional Neural Networks (PCNN) that uses the multi-instance learning paradigm to build a relation extractor using distant supervision in (Zhang & Wang, 2015), which outperforms the traditional non deep learning model based on distant supervision proposed by Mintz et al. (Mintz et al., 2009). A problem in the model by Zeng et al. (ibid.) is the information loss, which was addressed (Jiang, Wang, Li, & Wang, 2016) by using a cross-document max-pooling layer. But using the attention mechanism over the PCNN gives the best performance, which outperforms all other existing deep learning and non-deep learning models. In Universal Schema-based approaches there is an issue in generalizing to new open-domain relations. Toutanova et al. in (Toutanova et al., 2015) proposed a CNN based architecture which embeds the text between entities, which solves the previously stated issue in Universal Schema-based approach. Verga et al. used a similar kind of architecture in (Verga et al., 2016) like Toutanova et al. (Toutanova et al., 2015), rather than using just a CNN they tried CNN and LSTM-RNN both. They observed that LSTM-RNN outperforms the CNN architecture.

3

Methodology

There are several studies for solving relation classifi-cation task. Early methods used handcrafted features through a series of NLP tools or manually designing kernels. These approaches use high-level lexical and syntactic features obtained from NLP tools and manually designing kernels, but the classification models relying on such features suffer from propagation of implicit error of the tools.On the other hand, deep neural networks have shown outperform previous models using handcraft features. Especially, many researchers tried to solve the problem based on end-to-end models using only raw sentences and pre-trained word representations learned by Skip-gram and Continuous Bag-of-Words. Zeng et al. employed a deep convolutional neural net-work (CNN) for extracting lexical and sentence level features (Zeng, Liu, Chen, & Zhao, 2015). Dos Santos et al. proposed model for learning vector of each relation class using ranking loss to reduce the impact of artificial classes (Santos, Xiang, & Zhou, 2015). Zhang and Wang used bidirectional recurrent neural network (RNN) to learn long-term dependency between entity pairs.

We propose two methodologies to extract relations from Bangla. First one being CNN based and the second one is biLSTM based model.
In the CNN based model, the input to the CNN for relation extraction consists of sentences marked with the two entity mentions of interest. As CNN's can only work with fixed length inputs, we compute the maximal separation between entity mentions linked by relation and choose an input width greater than this distance. We ensure that every input (relation mention) has this length by trimming longer sentences and padding shorter sentences with a special token The network uses multiple window sizes for filters, position embeddings for encoding relative distances and pre-trained word embeddings for initialization in a non-static architecture. In our case, we will use ELMo deep contextualized embeddings.

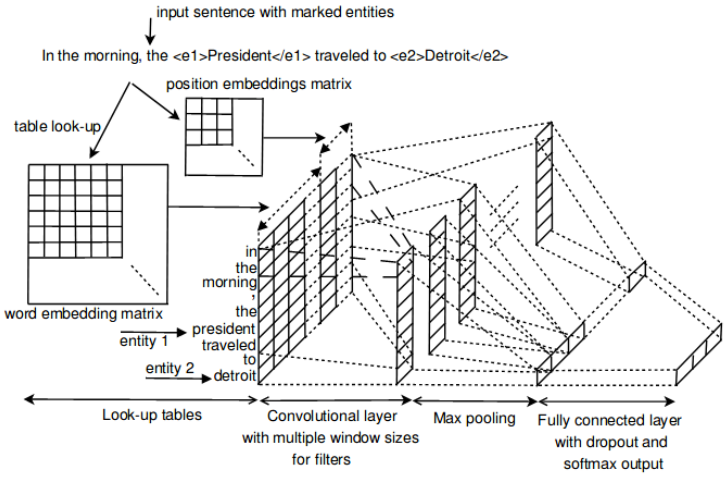

Figure 3.1: Relation Extraction with CNN

For the biLSTM based model, we will be using Entity-Aware Attention and Latent Entity Typing (Lee, Seo, & Choi, 2019).This model consist of four main components:

1. Word Representation that maps each word in a sentence into vector representations. We will use deep contextualized embeddings as it captures the contextual meaning.
2. Self-Attention that captures the meaning of the correlation between words based on multi-head attention.
3. BLSTMwhich sequentially encodes the representations of the self-attention layer.
4. Entity-aware Attentionthat calculates attention weights with respect to the entity pairs, word positions relative to these pairs, and their latent types obtained by LET.

After that, the features are averaged along with the time steps to produce the sentence-level features.

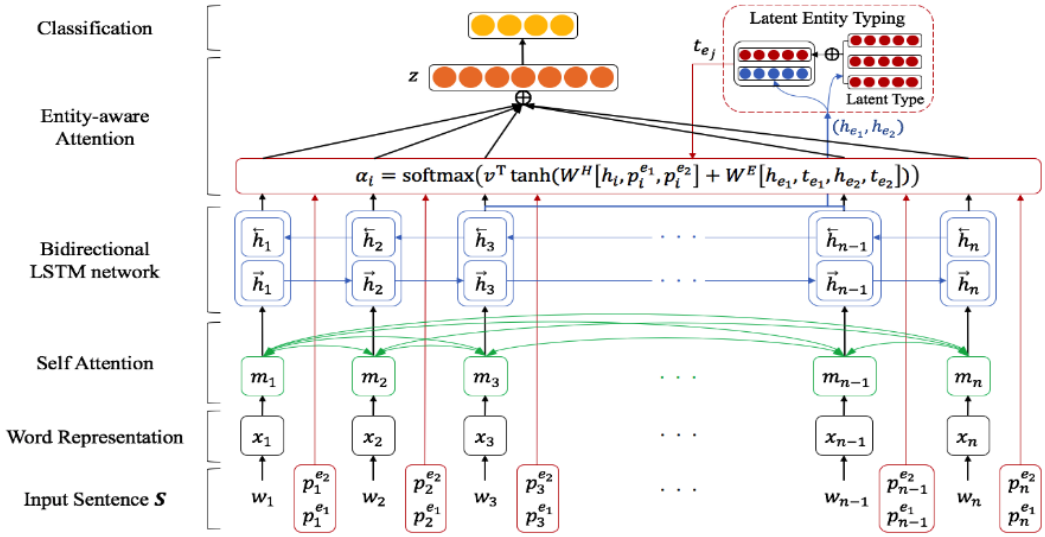

Figure 3.2: Relation Classification with Entity-aware Attention and Latent Entity Typing

3.1 Data Collection

Articles from online newspapers such as – Prothom Alo, Bangladesh Protidin etc., and Blogs are written in Bangla and Bangla Wikipedia are the main source of Bangla text. We will collect these Bangla articles using an open-source scrapping library named Spacy, which is widely popular for data scrapping from such sources.

3.2 Data Processing

This step is composed of two separate subtask – Data cleaning and Text Tokenization. All the articles collected from online contains lots of garbage words (i.e. words from another language) and punctuations. It is not possible to use these raw texts as inputs to our system directly, so to get rid of these noises we need to clean these scrapped texts frst. We will be using a Script written in Python 3.5 to clean the raw data. Now when the text is clean, these clean text will be tokenized into words. By tokenization, we mean that each article will be divided into smaller chunks of the word. So that we can represent each word with a unique token id. Tokenization will also be done using the same library used for scrapping i.e. Spacy. After completing data cleaning and tokenization we are ready to move into the next step of our pipeline.

3.3 Data Annotation

We developed an web based annotation tool to annotate our scraped dataset. This annotated dataset will be used for the training, testing, and validation. The web interface will be displaying a number of relations from a pre-defined set of relations. Field experts will be annotating each sentence by assigning an appropriate relation between two named entities. All these annotated data will be saved in a database, from which the whole dataset will be retrieved later on. The annotation will be a double-blinded process to ensure that all the sentences are annotated correctly. By the term ”Double-Blinded Process” we mean that we are going to establish such a protocol which will ensure that each sentence is annotated at least twice by two di□erent feld experts, and in case of con□ict between the assignment of any relationship we will take a third opinion and resolve it manually.

3.4 Novelty

There are no works on Relation Extraction have been done for the Bengali Language. Mostly because there are no annotated datasets, and annotating dataset is expensive and time consuming. The novelty of our work is that we annotated scrapped data and applied deep learning model to extract relation. Besides, we used deep contextualized embeddings instead of word2vec, GloVe or fasttext embeddings which doesn't take contenxt into consideration.

4

Deliverables

After successful implementation of relation extraction model we will be able to:

- Extract useful information from unstructured data.
- Query information from structured data.
- Build question answering systems.
- Build machine reading comprehension model.
- Build chatbots that could be used in E-learning platforms, Company websites/apps, Healthcare related apps etc.

5

Work Done

Since there were no existing annotated Bengali datasets, we built a corpus by scraping online newspapers and blogs website. Then we manually annotated the dataset. As manual annotation is time-consuming, we were able to annotate only around 5000 sentences before the deadline.

To use these annotated sentences for relation extraction, we needed to preprocess the data and incorporate with the relation extraction pipeline according to the model we decided to use.

Unfortunately, we couldn't finish the preprocessing stage before the deadline. So we decided to use an English language dataset for relation extraction, naming SemEval-2010 Task 8. It contains a total of 10,717 annotated sentences. We split the dataset into 8000 and 2717 for training and testing accordingly. The idea was to create a pipeline for relation extraction so that when we are done with the annotation and preprocessing phase, we can use the same pipeline for Bangla dataset right away.

As for the model, we decided to implement the model described in Lee et al. The model contains bidirectional LSTM networks which leverage entity-aware attention using latent entity typing. In the paper, for the embedding layer, they used GloVe embeddings. And before passing it to the biLSTM layer, it applies Self Attention. With these settings, they achieved an F1 score of 85.2 with latent entity typing and 84.7 without latent entity typing.

We used contextual embeddings, ELMo in our case. Since ELMo capture the contextual essence, we saw an immediate improvement in F1 score. In the paper, they achieved 85.2 with 100 epochs, and without tuning any hyperparameters, we achieved 85.3 in just 50 epochs.

Table 5.1: Comparison of relation extraction models

Models	F1
Non-Neural Model	
SVM	82.2
SDP-based Models	
MVRNN	82.4
FCM	83.0
DepNN	83.6
DepLCNN+NS	**85.6**
SDP-LSTM	83.7
DRNNs	**86.1**
End-to-End Models	
CNN	82.7
CR-CNN	84.1
Attention-CNN	84.3
+POS, WN, WAN	**85.9**
BLSMT	82.7
+ PF, POS, NER, DEP, WN	84.3
Attention-BLSTM	84.0
Hier-BLSTM	84.3
Proposed model (Lee et al.)	**85.2**
Proposed model + ELMo	**85.3**

References

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735--1780.

Jiang, X., Wang, Q., Li, P., & Wang, B. (2016). Relation extraction with multi-instance multi-label convolutional neural networks. In Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers (pp. 1471--1480).

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.

Lee, J., Seo, S., & Choi, Y. S. (2019). Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. Symmetry, 11(6), 785.

Liu, C., Sun, W., Chao, W., & Che, W. (2013). Convolution neural network for relation extraction. In International conference on advanced data mining and applications (pp. 231--242).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111--3119).

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp: Volume 2-volume 2 (pp. 1003--1011).

Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using lstms on sequences and tree structures. arXiv preprint arXiv:1601.00770.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp) (pp. 1532--1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Riedel, S., Yao, L., & McCallum, A. (2010). Modeling relations and their mentions without labeled text. In Joint european conference on machine learning and knowledge discovery in databases (pp. 148--163).

Santos, C. N. d., Xiang, B., & Zhou, B. (2015). Classifying relations by ranking with convolutional neural networks. arXiv preprint arXiv:1504.06580.

Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., & Gamon, M. (2015). Representing text for joint embedding of text and knowledge bases. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1499--1509).

Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1753--1762).

Zhang, D., & Wang, D. (2015). Relation classification via recurrent neural network. arXiv preprint arXiv:1508.01006.