Zero-shot Relation Classification as Textual Entailment

Abiola Obamuyide
Department of Computer Science
University of Sheffield
avobamuyide1@sheffield.ac.uk

Andreas Vlachos
Department of Computer Science
University of Sheffield
a.vlachos@sheffield.ac.uk

Abstract

We consider the task of relation classification, and pose this task as one of textual entailment. We show that this formulation leads to several advantages, including the ability to (i) perform zero-shot relation classification by exploiting relation descriptions, (ii) utilize existing textual entailment models, and (iii) leverage readily available textual entailment datasets, to enhance the performance of relation classification systems. Our experiments show that the proposed approach achieves 20.16% and 61.32% in F1 zero-shot classification performance on two datasets, which further improved to 22.80% and 64.78% respectively with the use of conditional encoding.

1 Introduction

The task of determining the relation between various entities from text is an important one for many natural language understanding systems, including question answering, knowledge base construction and web search. Relation classification is an essential part of many high-performing relation extraction systems in the NIST-organised TAC Knowledge Base Population (TAC-KBP) track (Ji and Grishman, 2011; Adel et al., 2016). As a result of its wide application, many approaches and systems have been proposed for this task (Zelenko et al., 2003; Surdeanu et al., 2012; Riedel et al., 2013; Zhang et al., 2017).

A shortcoming common to previous proposed approaches, however, is that they identify only relations observed at training time, and are unable to generalize to new (unobserved) relations at test time. To address this challenge, we propose to formulate relation classification as follows: Given a unit of text T which mentions a subject X and a candidate object Y of a knowledge base relation $R(X, Y)$, and a natural language description d of R, we wish to evaluate whether T expresses $R(X, Y)$. We formulate this task as a textual entailment problem in which the unit of text and the relation description can be considered as the premise P and hypothesis H respectively. The challenge then becomes that of determining the truthfulness of the hypothesis given the premise. Table 1 gives examples of knowledge base relations and their natural language descriptions.

This formulation brings a number of advantages. First, we are able to perform zero-shot classification of new relations by generalizing from the descriptions of seen training relations to those of unseen relations at test time. Given a collection of relations, for instance, *spouse(X,Y)* and *city_of_birth(X,Y)* together with their natural language descriptions and training examples, we can learn a model that can classify other instances of these relations, as well as instances of other relations that were not observed at training time, for instance *child(X,Y)*, given their descriptions. In addition to being able to utilize existing state-of-the-art textual entailment models for relation classification, our approach can use distant supervision data together with data from textual entailment as additional supervision for relation classification.

In experiments on two datasets, we assess the performance of our approach in two supervision settings: in a *zero-shot* setting, where no supervision examples are available for new relations, and in a *few-shot* setting, where our models have access to limited supervision examples of new relations. In the former setting our approach achieves 20.16% and 61.32% in F1 classification performance in the two datasets considered, which further improved to 22.80% and 64.78% respectively with the use of conditional encoding. Similar improvements hold in the latter setting as well.

Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 72–78
Brussels, Belgium, November 1, 2018.

Relation	Subject (**X**)	Object (**Y**)	Text (Premise)	Description (Hypothesis)
religious_order	Lorenzo Ricci	Society of Jesus	**X** (August 1, 1703 – November 24, 1775) was an Italian Jesuit, elected the 18th Superior General of the **Y**.	***X** was a member of the group **Y***
director	Kispus	Erik Balling	**X** is a 1956 Danish romantic comedy written and directed by **Y**.	*The director of **X** is **Y***
designer	Red Baron II	Dynamix	**X** is a computer game for the PC, developed by **Y** and published by Sierra Entertainment.	***Y** is the designer of **X***

Table 1: Examples of relations, entities, sample text instances, and relation descriptions.

2 Related Work

Most recent work, including Adel et al. (2016) and Zhang et al. (2017), proposed models that assume the availability of supervised data for the task of relation classification. Rocktäschel et al. (2015) and Demeester et al. (2016) inject prior knowledge in the form of propositional logic rules to improve relation extraction for new relations with zero and few training labels, in the context of the *universal schema* approach (Riedel et al., 2013). They considered the use of propositional logic rules, which for instance, can be mined from external knowledge bases (such as Freebase (Bollacker et al., 2008)) or obtained from ontologies such as WordNet (Miller, 1995). However, the use of propositional logic rules assumes prior knowledge of the possible relations between entities, and is thus of limited application in extracting new relations.

Levy et al. (2017) showed that a related and complementary task, that of entity/attribute relation extraction, can be reduced to a question answering problem. The task we address in this work is that of zero-shot *relation classification*, which determines if a given relation exists between two given entities in text. As a result the output of our approach is a binary classification decision indicating whether a given relation exists between two given entities in text. The task performed by Levy et al. (2017) is that of zero-shot entity/attribute *relation extraction*, since their approach returns the span corresponding to the relation arguments ("answers") from the text.[1] In addition, our approach for zero-shot relation classification utilizes relation descriptions, which is typically available in relation ontologies, and is thus not reliant on crowd-sourcing.

Our approach also takes inspiration from various approaches for leveraging knowledge from a set of source tasks to target tasks, such as recent transfer learning methods in natural language processing (Peters et al., 2018; McCann et al., 2017). Closest to our work is that of Conneau et al. (2017), who showed that representations learned from natural language inference data can enhance performance when transferred to a number of other natural language tasks. In this work, we consider the task of zero-shot relation classification by utilizing relation descriptions.

[1]Note that for this reason, a direct comparison between the two approaches is not straightforward, as this would be akin to comparing a text classification model and a question answering model.

3 Model

Our approach takes as input two pieces of text, a sentence containing the subject and object entities of a candidate relation, and the relation's description, and returns as output a binary response indicating whether the meaning of the description can be inferred between the two entities in the sentence. See Table 1 for some examples. The problem of determining whether the meaning of a text fragment can be inferred from another is that of natural language inference/textual entailment (Dagan et al., 2005; Bowman et al., 2015).

We take as our base model the Enhanced Sequential Inference Model (*ESIM*) introduced by Chen et al. (2017), one of the commonly used models for text pair tasks (**?**). *ESIM* utilizes Bidirectional Long Short-Term Memory (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) (BiLSTM) units as a building block and accepts two sequences of text as input. It then passes the two sequences through three model stages - input encoding, local inference modelling and inference composition, and returns the class c with the highest classification score, where c in textual entailment is one of *entailment*, *contradiction* or *neutral*. In our experiments, for each (sentence, relation description) pair we return a 2-way classification prediction instead.

In this section we briefly describe the input en-

coding and inference composition stages, which we adapt using conditional encoding as described in the following subsection. The *input encoding* and *inference composition* stages operate analogously, and each receives as input two sequences of vectors, $\{\boldsymbol{p}_i\}$ and $\{\boldsymbol{h}_j\}$, or more compactly two matrices $\boldsymbol{P} \in \mathbb{R}^{I\times d}$ for the premise and $\boldsymbol{H} \in \mathbb{R}^{J\times d}$ for the hypothesis, where I and J are respectively the number of words in the premise and hypothesis, and d is the dimensionality of each vector representation. In the case of the *input encoding* layer, P and H are word embeddings of words in the premise and hypothesis respectively, while in the case of *inference composition*, P and H are internal model representations derived from the preceding local inference modelling stage. Then the input sequences are processed with BiLSTM units to yield new sequences $\bar{\boldsymbol{P}} \in \mathbb{R}^{I\times 2d}$ for the premise and $\bar{\boldsymbol{H}} \in \mathbb{R}^{J\times 2d}$ for the hypothesis:

$$\bar{\boldsymbol{P}}\ ,\ \overrightarrow{\boldsymbol{c}}_p, \overleftarrow{\boldsymbol{c}}_p = BiLSTM(\boldsymbol{P}) \quad (1)$$

$$\bar{\boldsymbol{H}}\ ,\ \overrightarrow{\boldsymbol{c}}_h, \overleftarrow{\boldsymbol{c}}_h = BiLSTM(\boldsymbol{H}) \quad (2)$$

where $\overrightarrow{\boldsymbol{c}}_p\ , \overleftarrow{\boldsymbol{c}}_p \in \mathbb{R}^d$ are respectively the last cell states in the forward and reverse directions of the BiLSTM that reads the premise. $\overrightarrow{\boldsymbol{c}}_h\ , \overleftarrow{\boldsymbol{c}}_h \in \mathbb{R}^d$ are similarly defined for the hypothesis.

3.1 Conditional encoding for ESIM

When used for zero-shot relation classification, *ESIM* encodes the sentence independently of the relation description. Given a new target relation's description, it is desirable for representations computed for the sentence to take into account the representations for the target relation description. Therefore we explicitly condition the representations of the sentence on that of the relation description using a conditional BiLSTM (cBiLSTM) (Rocktäschel et al., 2016) unit. Thus, Equation 1 is replaced with:

$$\bar{\boldsymbol{P}} = cBiLSTM(\boldsymbol{P}, \overrightarrow{\boldsymbol{c}}_h, \overleftarrow{\boldsymbol{c}}_h) \quad (3)$$

where $\overrightarrow{\boldsymbol{c}}_h$ and $\overleftarrow{\boldsymbol{c}}_h$ respectively denote the last memory cell states in the forward and reverse directions of the BiLSTM that reads the relation description. This adaptation is made to both input encoding and inference composition stages. We refer to the adapted *ESIM* as the Conditioned Inference Model (*CIM*) in subsequent sections.

4 Datasets

We evaluate our approach using the datasets of Adel et al. (2016) and (Levy et al., 2017). The dataset of Adel et al. (2016) (*LMU-RC*) is split into training, development and evaluation sets. The training set was generated by distant supervision, and the development and test data were obtained from manually annotated TAC-KBP system outputs. We obtained the descriptions for the relations from the TAC-KBP relation ontology guidelines.[2] This resulted in a dataset of about 6 million positive and negative instances, each consisting of a relation, its subject and object entities, a sentence containing both entities and a relation description.

We applied a similar process to the relation extraction dataset of (Levy et al., 2017) (*UW-RE*). It consists of 120 relations and a set of question templates for each relation, containing both positive and negative relation instances, with each instance consisting of a subject entity, a knowledge base relation, a question template for the relation, and a sentence retrieved from the subject entity's Wikipedia page. We wrote descriptions for each of the 120 relations in the dataset, with each relation's question templates serving as a guide. Thus all instances in the dataset (30 million positive and 2 million negative ones) now include the corresponding relation description, making them suitable for relation classification using our approach.

In addition to the two datasets, we also utilize the *MultiNLI* natural language inference corpus (Williams et al., 2018) in our experiments as a source of supervision. We map its *entailment* and *contradiction* class instances to positive and negative relation instances respectively.

5 Experiments and Results

We conduct two sets of experiments. The first set of experiments tests the performance of our approach in the zero-shot setting, where no supervision instances are available for new relations (Section 5.1). The second set of experiments measures the performance of our approach in the limited supervision regime, where varying levels of supervision is available (Section 5.2).

Implementation Details Our model is implemented in Tensorflow (Abadi et al., 2016).

[2] https://tac.nist.gov/2015/KBP/ColdStart/guidelines/TAC_KBP_2015_Slot_Descriptions_V1.0.pdf

Dataset	Model	F1 (%)
LMU-RC	ESIM	20.16
	CIM	**22.80**
UW-RE	ESIM	61.32
	CIM	**64.78**

Table 2: Zero-shot relation learning results for *ESIM* and *CIM*.

Dataset	Supervision	F1 (%)
LMU-RC	TE	25.54
	TE+DS	**26.28**
UW-RE	TE	44.38
	TE+DS	**62.33**

Table 3: Zero-shot relation learning results for model CIM pre-trained on two sources of data: Textual Entailment (TE), or both Distant Supervision and Textual Entailment (TE+DS). The results in Table 2 correspond to DS only supervision.

We initialize word embeddings with 300D Glove (Pennington et al., 2014) vectors. We found a few epochs of training (generally less than 5) to be sufficient for convergence. We apply Dropout with a keep probability of 0.9 to all layers. The result reported for each experiment is the average taken over five runs with independent random initializations. In order to prevent overfitting to specific entities, we mask out the subject and object entities with the tokens *SUBJECT_ENTITY* and *OBJECT_ENTITY* respectively.

5.1 Zero-shot Relation Learning

For this experiment we created ten folds of each dataset, with each fold partitioned into train/dev/test splits along relations. In each fold, a relation belongs exclusively to either the train, dev or test split.

Table 2 shows averaged F1 across the folds for the models on the *LMU-RC* and *UW-RE* datasets. We observe that using only distant supervision for the training relations and without supervision for the test relations, the models were still able to make predictions for them, though at different performance levels. *CIM* obtained better performance compared to *ESIM*, as a result of its use of conditional encoding.

Table 3 shows F1 scores of model *CIM* pre-trained on only MultiNLI (referred to as TE) or a combination of MultiNLI and distant supervision (referred to as TE+DS) data in the zero-shot setting. We find that *CIM* pre-trained on only textual entailment data is already able to make predictions for unseen test relations, while using a combination of distant supervision and textual entailment data achieved improved F1 scores across both datasets, demonstrating the validity of our approach in this setting. We also note that using TE+DS data performs worse than DS data alone in the case of the *UW-RE* dataset, unlike in the case of *LMU-RC*. We hypothesize that this is because DS data performs much better for the former.

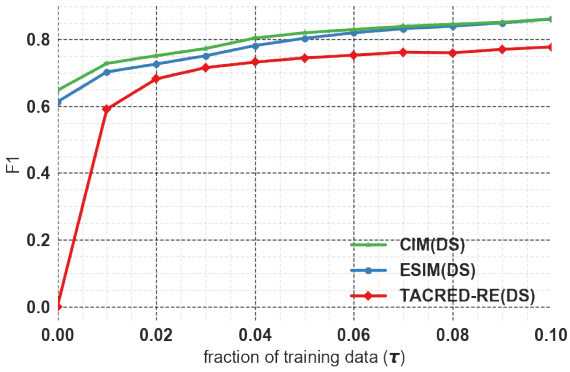


Figure 1: Limited supervision results: F1 scores on *UW-RE* as fraction of training data (τ) is varied. When τ=0, we get the zero-shot results in Table 2

5.2 Few-shot Relation Learning

For the experiments in the limited-supervision setting, we randomly partition the dataset along relations into a train/dev/test split. Similar to the zero-shot setting, a relation belongs to each split exclusively. Then for each experiment, we make available to each model a fraction τ of example instances of the relations in the test set as supervision. Note that the particular example instances we use are a disjoint set of instances which are not present in the development and evaluation sets. In addition to ESIM and our proposed model CIM, we also report results for the TACRED Relation Extractor (*TACRED-RE*), the position-aware RNN model that was found to achieve state-of-the-art results on the TACRED (Zhang et al., 2017) dataset. *TACRED-RE* is a supervised model that expects labelled data for all relations during training, and thus not applicable in the zero-shot setup.

Results for this set of experiments are shown in Figure 1 for the *UW-RE* dataset. We find that only about 5% of the training data is required for both *ESIM* and *CIM* to reach around 80% in F1

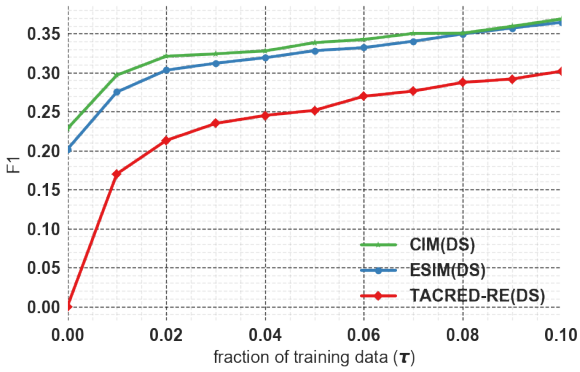


Figure 2: F1 scores on *LMU-RC* as fraction of training data (τ) is varied.

performance, with *CIM* outperforming *ESIM* in the 0-6% interval. However, beyond this interval, we do not observe any major difference in performance between ESIM and *CIM*, demonstrating that CIM performs well in both the zero-shot and limited supervision settings. For context, when given full supervision on the *UW-RE* dataset, *CIM* and *TACRED-RE* obtain F1 scores of 94.82% and 87.73% respectively. A similar trend is observed for the *LMU-RC* dataset, whose plot can be found in Figure 2.

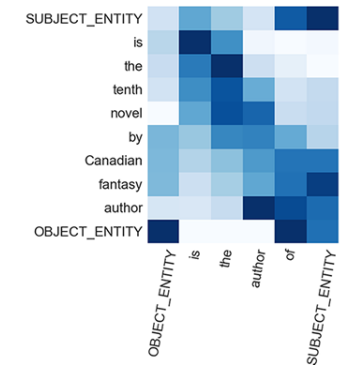
In general, all models obtain better results on *UW-RE* than on *LMU-RC*. We hypothesize that the performance difference is due to *UW-RE* being derived from Wikipedia documents (which typically have well-written text), while LMU-RC was obtained from different genres and sources (such as discussion forum posts and web documents), which tend to be noisier.

5.3 Qualitative Results

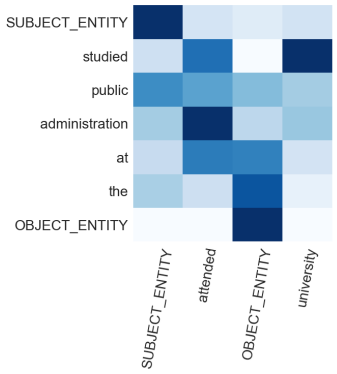
Figure 3 depicts a visualization of the normalized attention weights assigned by model *CIM* on instances drawn from the development set. We observe that it is able to attend to words that are semantically coherent with the premise (”novel” and ”author”, Figure 3a), (”studied” and ”university”, Figure 3b).

6 Conclusions

We show that the task of relation classification can be achieved through the use of relation descriptions, by formulating the task as one of textual entailment between the relation description and the piece of text. This leads to several advantages, including the ability to perform zero-shot relation classification and use textual entailment models and datasets to improve performance.



(a)



(b)

Figure 3: Attention visualization

Acknowledgments

We are grateful to Pasquale Minervini and Jeff Mitchell for helpful conversations and suggestions, and the Sheffield NLP group and anonymous reviewers for valuable feedback. This research is supported by the EU H2020 SUMMA project (grant agreement number 688139).

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, and Others. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI*, volume 16, pages 265–283.

Heike Adel, Benjamin Roth, and Hinrich Schütze. 2016. Comparing convolutional neural networks to traditional models for slot filling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 828–838, San Diego, California. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. *SIGMOD 08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190.

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2016. Lifted Rule Injection for Relation Embeddings. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, pages 1389–1399.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 2047–2052.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (June):74–84.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1119–1129.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance Multi-label Learning for Relation Extraction. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP '12*, (July):455–465.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45. Association for Computational Linguistics.