Relation Classification improved through Textual Entailment

Gabriele Brunini
University of Zurich
gbrunini@student.ethz.ch

Didem Durukan
University of Zurich
edurukan@student.ethz.ch

Christoph Mueck
University of Zurich
christophmueck@hotmail.com

Dominik Stammbach
ETH Zurich
dominik.stammbach@gess.ethz.ch

Abstract

In this course project, We propose to investigate the intersection of Natural Language Inference and Relation Classification. Relation Classification is the task of determining the relationship between two entities. It is considered one of the Natural Language Understanding building blocks and has critical applications, e.g., it is a sub-task of Relation Extraction and KB completion. We think about the task as a standard supervised learning task, which might have an interesting intersection with natural language inference. We could exploit this aspect to train better performing models – To investigate such, we conduct transformer-based experiments on an established benchmark in Relation Classification: the SemEVal2010 Task 8 dataset.

1 Introduction

A primary challenge in natural language understanding is relation extraction. Given that there is an enormous amount of natural language form, it is desirable to organize it via structured representations. Such representations, in turn, then have various applications in natural language understanding. Relation extraction provides the basic building blocks to create such structured representation, e.g., building knowledge graphs for a given text and populating knowledge bases (Grishman, 1997; Kumar, 2017; Zhang et al., 2017). Our aim in this project is to classify these relations, known as the Relation Classification task. We hypothesized that if we use Textual Entailment as a pre-training task, we can improve classification results.

1.1 Relation Classification

Relation extraction is about extracting triples of the form <*entity 1, relation, entity 2*> from unstructured text and relation classification is the task of classifying the relations, that is given a tuple <*entity 1, entity 2*>, we want to determine the type of *relation*. For example, let us have a sentence ”The company produces mugs”. Let *entity 1* be ”company” and let *entity 2* be ”mugs”. Then we want to determine the relation between the two entities in that sentence, i.e. assign it a label like ”Product-Producer(e2, e1)” or ”produces(x,y)”.

1.2 Textual Entailment

Textual entailment is a pairwise text classification task where we try to determine whether a given hypothesis can be inferred from the given sentence, i.e., we ask if, given the premise would be true, human reading of the hypothesis implies that the hypothesis also is true (Dagan et al., 2005; Bowman et al., 2015). In such cases, we would say the premise entails the hypothesis. Otherwise, it either contradicts the hypothesis, and certain datasets also include a ”neutral” class if neither of the two holds.

Let the *premise* be ”British oil executive was one of 16 people killed in the Saudi Arabian terror attack”, and let *hypothesis* be ”The bombers entered the embassy compounds”. This would be an example for non-entailment since it is trivial to see that hypothesis is not relevant to the given sentence. On the other hand, consider the *premise* be ”A Filipino hostage in Iraq was released”, and let *hypothesis 2* be ”A Filipino hostage was freed in Iraq”. This would hold as an example for **textually entailed** pair of *(premise,hypothesis)*.

Our objective is to improve relation classification performance using transformer-based models which have been pre-trained on textual entailment. The more related two tasks likely are, the better the transfer potential between two tasks (Ruder et al., 2019). We think that there exist various interesting links between Relation Classification and Textual Entailment, hence our investigation of the combination of the two tasks.

2 Datasets

The SemEval-2010 dataset (Hendrickx et al., 2009) is one of the benchmarks for relation classification. The dataset consists of 8000 training examples and 2717 test examples. The task focuses on the semantic relations between nominals and includes 19 classes (9 directed classes and one ”Other” class), e.g., Cause-Effect, Product-Producer, and others. The official metric is the macro-averaged F1 score for all classes except for the ”Other” class is ignored during evaluation.

The TACRED (TAC Relation Extraction Dataset) (Zhang et al., 2017) is a rather large-scale dataset for supervised relation classification. It contains 106,264 examples and 42 relations, where 79.5% of the examples are instances of the ”no_relation” class. The other 41 labels include relations like ”per:city_of_birth”, ”per:title” and ”org:founded_by”.

3 Goals

We plan to experiment with at least the SemEval-2010 datasets. We plan to re-implement R-BERT (Wu and He, 2019) and experiment with running a vanilla RoBERTa checkpoint and compare results to a RoBERTa checkpoint that has been pre-trained on MNLI (Bowman et al., 2015).

A typical entailment example from this dataset consists of the text ”A soccer game with multiple males playing.” and the hypothesis ”Some men are playing a sport.”. It is easy to see that both contain a variation of the relation play(men, soccer), whereas not entailed sentence pairs do not share such phenomena. Textual entailment, in general, is a good pre-training task (Conneau et al., 2018; Reimers and Gurevych, 2019) and relation classification can also be re-framed as zero-shot textual entailment (Obamuyide and Vlachos, 2018). These reasons combined make us believe that fine-tuning R-BERT from RoBERTa-Large-MNLI might provide a new state-of-the-art relation classification. If time allows and our hypothesis seems to hold on SemEval-2010 and R-BERT, we plan to also re-implement the current state of the art in relation classification (Cohen et al., 2020) and also conduct experiments on more datasets, e.g. TACRED.

Not surprisingly, transformer-based models have been introduced in Relation Classification (among all other conceivable NLP tasks), achieving stellar performance. The first and most popular model introducing BERT is R-BERT (Wu and He, 2019), which uses a BERT checkpoint. For each sentence, the model extracts three vectors, i.e. the CLS-token of the sequence and the (averaged) representation of both entities taking part in a relation. It then adds a non-linear layer on top of each of these representations and concatenates them together. This representation finally is used to classify relations in the different classes, and the loss is back-propagated through the whole transformer architecture. We have chosen this architecture because it is conceptually simple, and there exist various repositories on GitHub having implemented R-BERT, which keeps the scope of this project feasible.

Execution	roberta-large	roberta-large-mnli
1	0.8940	0.8921
2	0.8869	0.8920
3	0.8966	0.8921

Table 1: F1 scores

4 Experiments R-BERT on SemEval

4.1 Overall Results

As already mentioned, we are interested if R-BERT pre-trained on MNLI performs better than a vanilla R-BERT. To investigate this, we have run R-BERT on SemEval multiple times with different seeds (6 times in total, 3 times for both settings); one setting includes three runs with a vanilla Roberta-large checkpoint, the other setting consists of three runs with a Roberta-large checkpoint pre-trained on MNLI. We show the avg. macro-F1 (excluding the ”other” class of these three runs in table 1.

As we can see at first glance, the results do not seem to support our hypothesis. Moreover, the results do not support a consistent story – twice, Roberta-large slightly performs better than Roberta-large-MNLI, once it is the other way around. Differences in F1 remain marginal. Overall, we doubt that our proposed goal is in line with our expectations outlined in Section 3.

To ensure whether these results are statistically significant, we calculate a T-test for the means of two independent samples of scores to examine whether they are significantly different from each other. The standard T-test assumes equal variances of the two samples. We test for equal variances by calculating the Levene test statistics (p-value of 0.20). We cannot reject the null hypothesis, and

Table 2: Qualitative analysis

Sentence	True Label	Roberta Prediction	Roberta-MNLI prediction
Sentence: The $< e1 >$ song $< /e1 >$ was composed for a famous Brazilian $< e2 >$ musician $< /e2 >$.	Product-Producer(e1,e2)	Instrument-Agency(e1,e2)	Product-Producer(e1,e2)
In recent years, most $< e1 >$ floppies $< /e1 >$ have shipped pre-formatted from the $< e2 >$ factory $< /e2 >$ as DOS FAT12 floppies.	Product-Producer(e1,e2)	Product-Producer(e1,e2)	Entity-Origin(e1,e2)

thus, the two samples are assumed to have the same variance. The null hypothesis of the T-test is that two independent samples have identical average values. Resulting in a p-value of 0.89, we cannot reject the null hypothesis. Thus, we can reject that the two average scores are significantly different from each other.

4.2 In-depth analysis

After having not found statistically significant differences in performance of the two tested models on relation classification, we pursued a more fine-grained analysis of the model outcomes. First, we conducted a qualitative analysis of examples where our model's predictions did not agree. We show two examples of such examples in Table 2.

Table 3: F1 Per Class

Class	Roberta-vanilla F1	Roberta-MNLI F1
Cause-Effect(e1,e2)	0.94	0.92
Cause-Effect(e2,e1)	0.93	0.94
Component-Whole(e1,e2)	0.91	0.89
Component-Whole(e2,e1)	0.85	0.83
Content-Container(e1,e2)	0.92	0.90
Content-Container(e2,e1)	0.89	0.84
Entity-Destination(e1,e2)	0.95	0.95
Entity-Destination(e2,e1)	0.00	0.00
Entity-Origin(e1,e2)	0.90	0.87
Entity-Origin(e2,e1)	0.89	0.88
Instrument-Agency(e1,e2)	0.76	0.68
Instrument-Agency(e2,e1)	0.87	0.84
Member-Collection(e1,e2)	0.80	0.82
Member-Collection(e2,e1)	0.91	0.88
Message-Topic(e1,e2)	0.92	0.92
Message-Topic(e2,e1)	0.89	0.88
Other	0.69	0.65
Product-Producer(e1,e2)	0.85	0.89
Product-Producer(e2,e1)	0.88	0.89
micro-avg	0.87	0.86
macro-avg	0.83	0.81
weighted-avg	0.87	0.85

In total, we found 194 examples where our models would not agree. We looked through a handful of such examples but did not find any clear trends on which we could elaborate in detail. Given that these models are black boxes, it remains hard to speculate why the models would disagree without observing any clear trends in the examples we examined.

Furthermore, we checked whether we could find any large differences for specific sub-classes of the data. We show F1 score per class and model in table 3. Again, we do not find any striking quantitative differences.

Furthermore, these results differ across runs with different seeds. We had Roberta-MNLI performing better consistently across classes. For some runs, we have the picture outlined in Table 3. Overall, we see that our hypothesis **Textual Entailment is a good pre-training task for relation classification** unfortunately seemed to have failed.

5 Conclusion

In this course project, we investigated whether Textual Entailment (TE) would improve the performance of Relation Classification systems. We have described the task of Relation Classification and Textual Entailment and motivated our hypothesis, e.g. that TE is a good pre-training task in general, that in entailed TE examples, relations stay consistent, and that Relation Classification can be achieved as zero-shot Textual Entailment. Given these reasons combined, we thought it is likely that our hypothesis might be true. We then conducted experiments on the widely-established SemEval-2010 dataset, fine-tuned a Roberta-large checkpoint, and compared it with a Roberta-large checkpoint pre-trained on MNLI. We found that there does not seem to be any difference using the official evaluation metric of the dataset (macro-F1) in whether we should use one over the other. We then investigated in more detail whether we would find any salient differences, e.g. we manually looked through some of the examples where the two classifiers didn't agree but did not spot any clear trends. Lastly, we analyzed whether some classes would be greatly affected by using the vanilla Roberta model vs the one pre-trained on MNLI, but we have not found any differences there either. Hence, we report that our hypothesis was not supported in our experimental setting, and these findings concludes our project. Given these insights, we did not experiment further with different architectures and on different datasets.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation extraction as two-way span-prediction.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. Supervised learning of universal sentence representations from natural language inference data.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, SCIE '97, pages 10–27, London, UK, UK. Springer-Verlag.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

Shantanu Kumar. 2017. A survey of deep learning methods for relation extraction. *CoRR*, abs/1705.03645.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification.

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.