Exploring the zero-shot limit of FewRel

Alberto Cetoli
QBE Europe
30 Fenchurch St,
Billingsgate, London
EC3M 3BD
alberto.cetoli@uk.qbe.com

Abstract

This paper proposes a general purpose relation extractor (RE) that uses Wikidata descriptions to represent the relation's surface form. The results are tested on the FewRel 1.0 dataset, which provides an excellent framework for training and evaluating the proposed zero-shot learning system in English. This relation extractor architecture exploits the implicit knowledge of a language model through a question-answering approach[1].

1 Introduction and related works

The FewRel dataset (Han et al., 2018) has been tailored for the task of *few shot learning*, where the model is presented with a limited sample of candidates in English for each relation (Table 1). The evaluation is performed on a set of relations unseen during training, thus forcing the system to abstract and generalize some core language features from the examples. A plethora of methods has flourished to address the challenge posed by the FewRel set (Ye and Ling, 2019; Gao et al., 2019), eventually achieving super-human performance in the work of (Baldini Soares et al., 2019). Understandably, these methods greatly benefit from recent progress on pre-trained language models like BERT (Devlin et al., 2018), leveraging on the model's implicit semantic knowledge. Building extractors with a broad purpose is of utmost importance in information extraction, both at a theoretical and application level. In this spirit, the zero-shot learning approach implements the scenario where the model is never presented with any examples. An early implementation of this technique can be found in the *open information extraction* framework (Banko et al., 2007; Fader et al., 2011) which represents the relations with their surface forms. More recently the work by (Levy et al., 2017) builds on Question-Answering techniques to build new datasets and models for RE: The relations are represented as questions and the answers are the connected entities. Another work by (Obamuyide and Vlachos, 2018) uses Textual Entailment for relation extraction. Since many sentences can express the same relation, typically zero-shot relation extractors are limited in their ability to generalize and do not perform as well as few-shot learning models.

This work aims to improve upon prior systems by leveraging on the advancements of question answering models and by representing relations using their surface forms as they appear on Wikidata. The FewRel dataset is an ideal playground for this task: It forces the system to generalize by evaluating on unseen relations while at the same time mapping every relation to a Wikidata identifier. The contributions of this paper are as in the following

- Exploring the limitations and establishing a benchmark for zero-shot relation extraction on FewRel 1.0. The outcome shows how close zero-shot RE can be to one-shot learning.
- Introducing a new technique to exploit the implicit knowledge of a language model fine-tuned on SQUAD.
- Building a general purpose system that only needs a Wikidata-style description as a surface form to extract relations.

[1] The code for this paper is available at https://github.com/fractalego/fewrel_zero_shot.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1447–1451
Barcelona, Spain (Online), December 8-13, 2020

Wikidata ID	Description	Aliases
P140	religion of a person, organization or religious building, or associated with this subject	religious affiliation, faith, life stance, denomination
P931	territorial entity or entities served by this transport hub (airport, train station, etc.)	serves city, city served, train station serves

Table 1: A sample of relations used in the train set of FewRel 1.0.

2 Problem statement

The overarching goal is to classify a relation that connects two entities, i.e. to build a system that - given a pair of entities e_1, e_2, and a sentence s - predicts the probability of the entities being connected by a relation r. In this work, the relation is represented through its surface form. Consider the example sentence "John Smith receives an OBE" with entities e_1 as "John Smith" and e_2 as "OBE". With the proposed method the correct relation is found by iterating over all the available surface forms and choosing the most probable one:

$$r^* = \text{argmax}_r \mathcal{P}\left(r, e_1, e_2, s\right) \,. \tag{1}$$

Zero-shot relation extraction is notoriously difficult, due to the challenge of generalizing the surface form to every possible semantic equivalent. In order to improve generalization, this paper aims to utilize the implicit knowledge contained in a pre-trained language model fine-tuned on SQUAD 1.1, a question-answering dataset (Rajpurkar et al., 2016). The challenge is therefore to implement a model that is compatible with Eq. (1) and the question-answering task. A key insight is to notice that relation extraction can be understood as a question answering problem (Levy et al., 2017), where the surface form is the query and the entities connected by the relation are the "answer" to the query.

In this work, however, the entities e_1 and e_2 are an input of the model. Our solution is to signal that something is different about the relevant words by doubling the entities' tokens[2]: For example, the entity "John Smith" becomes "John John Smith Smith" and "OBE" appears as "OBE OBE". Please note that - as the results show - the system does not just learn to associate repeating words to boundaries: In a nutshell *we are asking the model if the entities are in the right place given a relation*. Using this technique the pre-trained question answering model is seen to vastly improve the generalization in the task at hand.

An adversarial approach to the training is shown in Fig. 1: The model is taught to generate the entities given a sentence and a surface form (left of Fig. 1); in the "fake" examples the system should return an *adversarial configuration* (right of Fig. 1), where no entity is found: In this scenario all the output items are vanishing except the first. The system thus learns to quantify an *Adversarial Score* $\mathcal{A}$. The probability in Eq. (1) can then be defined as

$$\mathcal{P} = 1 - \mathcal{A}\,. \tag{2}$$

Notice that *the real output of the system is this Adversarial Score*. The tagging of relevant entities is an expedient used to exploit the implicit knowledge of a pre-trained question answering system.

3 Model

The model is based on a straightforward extension of the BERT architecture for question answering on SQUAD (Devlin et al., 2018) (Fig. 1): As an input the system is presented with a text. This text is a concatenation of the relation's surface form (e.g. "A person has a title") and the sentence we want to extract the relation from (e.g. "John Smith receives an OBE").

[2] The author tried different methods, including parenthesis around the entities or other special characters. The method presented here provides the highest score.

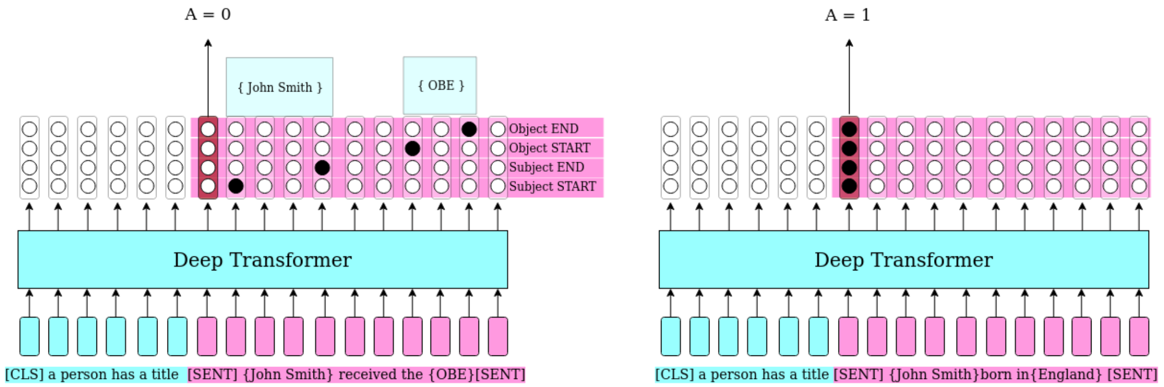
A = 0
{ John Smith }
{ OBE }
Object END
Object START
Subject END
Subject START
Deep Transformer
[CLS] a person has a title [SENT] {John Smith} received the {OBE}[SENT]
A = 1
Deep Transformer
[CLS] a person has a title [SENT] {John Smith}born in{England} [SENT]

Figure 1: On the left: During training the system is fed the relation's surface form and a sentence. If a relation is genuine the model outputs the relation's entities boundaries. On the right: For adversarial inputs no entity is tagged. All the columns in the output should be vanishing except the first. During evaluation the only output is the adversarial score $\mathcal{A}$ (see text).

The input is first tokenized according to the WordPiece algorithm (Wu et al., 2016), then passed onto a pre-trained BERT model (Wolf et al., 2019). The last hidden layer of BERT is directed onto a dense layer with a 4-dimensional output per token. Finally, a softmax filter is applied *horizontally*. In this way each row gives a unique position for the start and end of the relation's subject and object. The softmax is only applied for the length of the sentence, consistently with the task of marking the entities' boundaries.

As shown in Fig. 1 if the relation does not connect the relevant entities, the first token (corresponding to a [SENT] tag) is trained to be 1 for all the rows of the output. Ideally this method enforces that when no relation is found in the sentence no entity is tagged. Each row of the output is therefore trained to give an adversarial score $\mathcal{A}_i$. The *Adversarial Score* $\mathcal{A}$ for Eq. (2) is then defined as the minimum of all the four adversarial scores.

$$\mathcal{A} = \min(\mathcal{A}_i)\,. \tag{3}$$

4 Training

The model is implemented in PyTorch (Paszke et al., 2019). The original FewRel set is augmented with adversarial examples, generated by choosing a random relation from the training set excluding the correct one. The surface form itself is chosen randomly from a list made by concatenating the relation *description* and its *aliases*, as seen in Table 1.

Four different pre-trained models are used to investigate the performance of the current architecture. The models differ in size and on whether they have been fine-tuned on the SQUAD set. For all of them the Adam optimizer is employed:

1. Distillbert: This is the smallest model. We fine-tune the system with 2×10^{-5} step size. The adversarial set is three times the size of the original set.

2. Bert large: Step size 10^{-5}. The adversarial set is two times the size of the original set.

3. DistillBert fine-tuned on SQUAD: Step size 2×10^{-6}. The adversarial set is three times the size of the original set.

4. Bert large fine-tuned on SQUAD: Step size 3×10^{-6}. The adversarial set is two times the size of the original set.

5 Results and discussion

The zero-shot accuracy is compared with prior 1-shot results: While the 5-shot regime scores better, the number of aliases per Wikidata relation is inconsistent, thus there cannot be a guarantee that a relation will have 5 different surface forms to choose from. Only the Wikidata description is used in validation and testing.

	0-shot 5-ways	**0-shot 10-ways**	**1-shot 5-ways**	**1-shot 10-ways**
(Ye and Ling, 2019)	-	-	79.0 ± 0.2	67.4 ± 0.2
(Gao et al., 2019)	-	-	85.66	76.84
(Baldini Soares et al., 2019)	-	-	88.9	$-$
(1) Distillbert	70.1 ± 0.5	55.9 ± 0.6	-	-
(2) Bert	80.8 ± 0.4	69.6 ± 0.5	-	-
(3) Distillbert + SQUAD	81.3 ± 0.4	70.0 ± 0.2	-	-
(4) Bert + SQUAD	86.0 ± 0.6	76.2 ± 0.4	-	-

Table 2: Model accuracy on the FewRel 1.0 Validation set. These results are an average of 5 runs, using as error the difference between the best and worst estimate.

	0-shot 5-ways	**0-shot 10-ways**	**1-shot 5-ways**	**1-shot 10-ways**
(Ye and Ling, 2019)	-	-	82.98	73.59
(Gao et al., 2019)	-	-	88.32	80.63
(Baldini Soares et al., 2019)	-	-	93.9	89.2
(4) Bert + SQUAD	82.72 ± 0.02	67.9 ± 0.7	-	-

Table 3: Model accuracy on the FewRel 1.0 Test. These evaluations are an average of two results computed on the test set while the error is their difference.

Validation: As shown in Table 2 the vanilla Distillbert (1) model is seen with the lowest score. On a closer inspection most of the errors are due to the model confusing the relations "part of" ($P361$) and "member of" ($P463$): The system does not have enough granularity to distinguish between the two. This issue seems to disappear when using the larger pre-trained BERT (2).

Both the Distillbert and BERT models improve after being fine tuned on SQUAD (3 and 4). It is worth noticing that the smaller model (3) improves dramatically, achieving better accuracy than the vanilla BERT model (2): Clearly the question answering dataset provides a relevant template for zero-shot relation extraction. This phenomenon suggests a clear way to increase the accuracy: fine-tune the question answering model on a dataset with more examples than SQUAD 1.1.

In the end, the best results on the validation set are achieved with the larger BERT model (4) after fine tuning on SQUAD. Remarkably, the accuracy is shown to be competitive with 1-shot results from recent models, although the limited number of relations on the validation set - only 16 - cannot guarantee generalization on the set of all possible relations.

Testing: The results on the test set are less competitive (Table 3). It is more difficult to inspect the reason behind the performance drop since the test set is not released publicly. Likely this set contains relations whose surface form is semantically close, as for the smallest model in validation. Even so, the 5-ways result nears prior 1-shot accuracies. The 10-ways result seems to suffer more from lack of generalization, ostensibly because the relations that are semantically close are more likely to appear in the same batch.

6 Conclusions and future work

This paper built and tested a general purpose system that only needs a Wikidata-style description as a surface form to extract relations. The results are evaluated on the FewRel 1.0 set and compared to recent works. The model seems competitive on the validation set, whereas it is found struggling on the test set due to the intrinsic challenge of generalizing from a relation's surface form.

An intriguing future line of research is to combine the zero-shot description with few-shot examples. This hybrid approach should be able to leverage on this paper's findings and achieve competing results in the task of relation extraction.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July. Association for Computational Linguistics.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. 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 6251–6256, Hong Kong, China, November. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, October-November. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, August. Association for Computational Linguistics.

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, November. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2872–2881, Florence, Italy, July. Association for Computational Linguistics.