Cardiff University at SemEval-2020 Task 6: Fine-tuning BERT for Domain-Specific Definition Classification

Shelan S. Jeawak, Luis Espinosa-Anke and Steven Schockaert
School of Computer Science and Informatics
Cardiff University, UK
{jeawakss,espinosa-ankel,schockaerts1}@cardiff.ac.uk

Abstract

We describe the system submitted to SemEval-2020 Task 6, Subtask 1. The aim of this subtask is to predict whether a given sentence contains a definition or not. Unsurprisingly, we found that strong results can be achieved by fine-tuning a pre-trained BERT language model. In this paper, we analyze the performance of this strategy. Among others, we show that results can be improved by using a two-step fine-tuning process, in which the BERT model is first fine-tuned on the full training set, and then further specialized towards a target domain.

1 Introduction

Definitions are central to the way in which humans convey knowledge about the meaning of concepts. Accordingly, a large number of general and domain-specific dictionaries have been created. As new concepts emerge, and the meaning of existing concepts shifts, these dictionaries need to be updated. This continual process is traditionally carried out by linguists or domain experts, meaning that dictionaries are never fully up-to-date. In rapidly evolving scientific domains, among others, this is a clear limitation. An appealing alternative is to automatically identify and extract definitions expressed in free text. This task of extracting term-definition pairs from text corpora is known as Definition Extraction (DE).

Early attempts to solve this task relied on rule-based methods (Klavans and Muresan, 2001; Cui et al., 2005). However, such methods are typically only able to detect explicit, direct and structured definitions, which usually contain definitor verb phrases such as *means, is, is defined as*. Later, a large number of supervised and semi supervised machine learning models for DE have been proposed (Westerhout, 2009; Reiplinger et al., 2012; Jin et al., 2013). While being able to identify a wider range of definitions, these approaches cannot be adapted to new domains efficiently, as they crucially rely on carefully designed features, which might not be available, or be less effective, in the new domain. More recently, the focus has shifted to neural network based models (Espinosa-Anke and Schockaert, 2018; Veyseh et al., 2019).

The method we analyze in this paper is based on fine-tuning a pre-trained BERT language model (Devlin et al., 2018). This strategy has recently proven successful across a wide range of Natural Language Processing (NLP) tasks. In particular, we focus on SemEval-2020 Task 6: DeftEval: Extracting term-definition pairs in free text (Spala et al., 2020). We participated in Subtask 1: Sentence Classification. This task required participants to predict whether a given sentence contains a definition. The associated dataset contains documents from seven different domains, including biology, history and economics. In our analysis, we focus on comparing two different strategies for fine-tuning the pre-trained BERT model:

1. fine-tuning a single BERT model based on all the available training data;
2. fine-tuning a separate BERT model for each of the 7 domains, each time only relying on the training data that is available for that domain.

The first strategy has the advantage that all training data can be exploited. However, our hypothesis is that this strategy may struggle to optimally capture the different definition styles that are used in different domains. The second strategy avoids confusing the classifier with different definition styles, but it implies

that only a limited amount of training data is available for each domain. We also experiment with a third approach, which is aimed at combining the best of both worlds:

3. fine-tuning a domain-specific BERT model in two steps, by first fine-tuning the model on all training data, and subsequently specializing it to the target domain in an additional fine-tuning step.

2 Data

We used the DEFT corpus (Spala et al., 2019), which was made available as part of the SemEval-2020 Task 6 competition. This dataset was collected from a repository of English open source textbooks[1] and annotated by five annotators using the brat annotation framework[2]. For Subtask 1, the available sentences are split into 17,819 training sentences, 872 development sentences and 859 testing sentences. Each split contains data from 7 domains: biology, history, physics, psychology, economic, sociology, and government. The number of sentences in each split for each domain are listed in Table 1.

Domain	Train	Dev.	Test
Biology	5,041	216	232
Economic	2,500	91	89
Government	3,520	238	255
History	1,301	54	50
Physics	1,699	94	83
Psychology	2,440	103	97
Sociology	1,318	76	53

Table 1: The number of sentences in each domain.

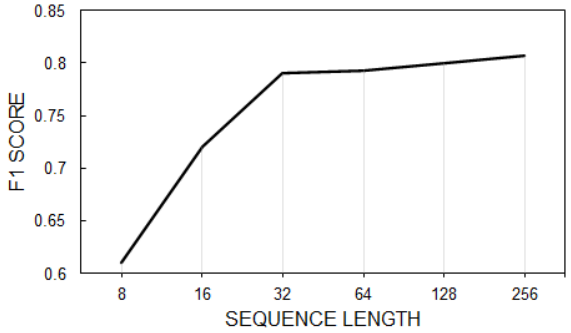

Figure 1: Impact of padding/truncating sentences to different lengths, for 'BERT - Fine-tune all domains'. The reported F1 score is for the positive class.

3 Methods

Given the success of BERT (Devlin et al., 2018) across a wide range of NLP tasks, we decided to focus on analyzing its performance in the context of definition extraction. We considered the following variants.

Fine-tuning strategies. We experimented with the BERT-base model, using the pytorch huggingface implementation named *BertForSequenceClassification*[3]. Essentially, this method corresponds to adding a classification layer on top of the pre-trained BERT model, and fine-tuning the BERT model while training the classification layer. We will specifically compare the following fine-tuning strategies:

- **BERT-all:** We fine-tune the model based on all the training data (i.e. from all domains). This is our official submission to the competition, which ranked 16 out of 56 submissions.
- **BERT-target:** We fine-tune the model only on training data for a given target domain. In other words, for each of the 7 domains, we train a separate model.
- **BERT-double:** We fine-tune the pre-trained BERT model twice. For the first fine-tuning step, we used the training data from all domains. Subsequently, we fine-tune the resulting model, based on the training data from the considered target domain.

As a baseline strategy, we also explored the following variant:

- **BERT-name:** In this case, we used all the available training data, but we add the domain name at the start of the input sentence as an additional token, to condition the model.

[1] https://cnx.org
[2] http://brat.nlplab.org/
[3] https://huggingface.co/transformers/v2.2.0/model_doc/bert.html#bertforsequenceclassification

		P	R	F1
BERT-all	0	0.90	0.90	0.90
	1	0.79	0.80	0.80*
	Macro avg.	0.85	0.85	0.85
	Weighted avg.	0.87	0.87	0.87
BERT-name	0	0.89	0.90	0.89
	1	0.79	0.76	0.78
	Macro avg.	0.84	0.83	0.84
	Weighted avg.	0.86	0.86	0.86
BERT-target	0	0.89	0.89	0.89
	1	0.77	0.77	0.77
	Macro avg.	0.83	0.83	0.83
	Weighted avg.	0.85	0.85	0.85
BERT-double	0	0.92	0.90	0.91
	1	0.81	0.84	**0.83**
	Macro avg.	0.87	0.87	0.87
	Weighted avg.	0.88	0.88	0.88

* This is our official submission to the competition.

		P	R	F1
LSTM-base	0	0.82	0.89	0.85
	1	0.73	0.6	0.66
	Macro avg.	0.77	0.74	0.75
	Weighted avg.	0.79	0.79	0.79
LSTM-BERT-pre	0	0.86	0.91	0.89
	1	0.79	0.71	0.75
	Macro avg.	0.83	0.81	0.82
	Weighted avg.	0.84	0.84	0.84
LSTM-BERT-ft	0	0.85	0.93	0.89
	1	0.83	0.67	0.74
	Macro avg.	0.84	0.80	0.82
	Weighted avg.	0.85	0.85	0.84

Table 2: Results of sentence classification task.

LSTM based strategies. Apart from the standard strategy of fine-tuning a BERT model, we also experimented with using LSTMs (Hochreiter and Schmidhuber, 1997), using contextualised word vectors from BERT as input. We again compare several strategies:

- **LSTM-base:** We used BertTokenizer[4] to tokenize the sentence. For this baseline model, we then trained the LSTM, including the corresponding token embeddings, from scratch. We used 300-dimensional word embeddings and two hidden layers of 256 dimensions.
- **LSTM-BERT-pre:** We used the same LSTM architecture as before, but instead of learning the token embeddings, we used the last hidden of the pre-trained 'bert-base-uncased' model.
- **LSTM-BERT-ft:** In this case, we used the final layer of the fine-tuned **BERT-all** model as the word embedding layer to the LSTM model.

4 Results and Discussion

For all experiments, we used Google Colab free GPU [5] to train 'bert-base-uncased' models for 4 epochs. For the **BERT-double** method, we used 4 epochs for each of the two fine-tuning steps. We set the batch size to 16 and we pad the sentences to the 256 sequence length, which gave the best performance for the **BERT-all** model based on the development set, as shown in Figure 1. We used the Adam optimizer and a learning rate of $2 \cdot 10^{-5}$ and 10^{-3} for fine-tuning BERT models and LSTM based models respectively.

The results of the considered methods are summarized in Table 2 in term of precision, recall and F1 score. We show the performance of each model for predicting the positive (1) and negative (0) classes as well as their macro and weighted average. The official score in the competition was the F1 score for the positive class. The results show that fine-tuning BERT outperforms the LSTM based strategies. When comparing the different fine-tuning strategies, we found that specifying the domain name as an additional token in *BERT-name* failed to outperform the standard fine-tuning strategy. On average, the standard strategy also performed better than domain-specific fine-tuning. However, the double fine-tuning strategy led to the best results overall. A more detailed analysis of the main fine-tuning strategies is presented in Table 3, which shows the results for each of the 7 domains separately. One surprising finding is that the relative performance of the domain-specific fine-tuning strategy does not seem directly related to the amount of training data. In particular, this strategy outperforms the 'all domains' strategy on the History domain, despite the fact that far less training data is available for this domain than for most of the others. Conversely, despite the fact that Government is one of the largest domain, in terms of available training data, the domain specific fine-tuning strategy performs comparatively very poorly. Table 4 lists randomly selected examples of the incorrectly classified sentences from Government domain. Looking at these

[4] https://huggingface.co/transformers/v2.2.0/model_doc/bert.html#berttokenizer
[5] https://colab.research.google.com

		BERT-all			BERT-target			BERT-double		
		P	R	F1	P	R	F1	P	R	F1
Biology	0	0.87	0.85	0.86	0.84	0.86	0.85	0.89	0.87	0.88
	1	0.83	0.85	0.84	0.84	0.81	0.82	0.85	0.87	**0.86**
	Macro avg.	0.85	0.85	0.85	0.84	0.83	0.83	0.87	0.87	0.87
	Weighted avg.	0.85	0.85	0.85	0.84	0.84	0.84	0.87	0.87	0.87
Economic	0	0.97	1	0.98	0.95	1	0.97	0.97	1	0.98
	1	1	0.93	**0.96**	1	0.89	0.94	1	0.92	**0.96**
	Macro avg.	0.98	0.96	0.97	0.97	0.95	0.96	0.98	0.95	0.97
	Weighted avg.	0.98	0.98	0.98	0.97	0.97	0.97	0.98	0.97	0.97
Govern.	0	0.91	0.94	0.93	0.9	0.92	0.91	0.92	0.93	0.93
	1	0.68	0.56	0.61	0.57	0.51	0.54	0.67	0.62	**0.64**
	Macro avg.	0.79	0.75	0.77	0.74	0.72	0.72	0.79	0.78	0.79
	Weighted avg.	0.87	0.87	0.87	0.84	0.85	0.84	0.88	0.88	0.88
History	0	0.74	0.87	0.8	0.78	0.83	0.81	0.78	0.83	0.81
	1	0.73	0.55	0.63	0.72	0.65	**0.68**	0.72	0.65	**0.68**
	Macro avg.	0.74	0.71	0.71	0.75	0.74	0.75	0.75	0.74	0.75
	Weighted avg.	0.74	0.74	0.74	0.76	0.76	0.76	0.76	0.76	0.76
Physics	0	0.87	0.87	0.87	0.85	0.87	0.86	0.89	0.85	0.87
	1	0.65	0.65	0.65	0.64	0.61	0.62	0.65	0.74	**0.69**
	Macro avg.	0.76	0.76	0.76	0.74	0.74	0.74	0.77	0.79	0.78
	Weighted avg.	0.81	0.81	0.81	0.79	0.80	0.79	0.83	0.82	0.82
Psycho.	0	0.92	0.83	0.87	0.91	0.83	0.86	0.93	0.88	0.90
	1	0.78	0.90	0.83	0.77	0.87	0.82	0.84	0.90	**0.86**
	Macro avg.	0.85	0.86	0.85	0.84	0.85	0.84	0.89	0.89	0.89
	Weighted avg.	0.86	0.86	0.86	0.85	0.85	0.85	0.89	0.89	0.89
Sociology	0	0.94	0.89	0.92	0.94	0.84	0.89	0.94	0.86	0.90
	1	0.78	0.88	**0.82**	0.70	0.88	0.78	0.74	0.88	0.80
	Macro avg.	0.86	0.88	0.87	0.82	0.86	0.83	0.84	0.87	0.85
	Weighted avg.	0.89	0.89	0.89	0.87	0.85	0.85	0.88	0.87	0.88

Table 3: Performance of sentence classification for each domain when fine-tuning BERT model.

sentences, we can see that some gold definitions are either incorrectly labeled or very difficult to classify even for a human. For instance, the first sentence contains a definition (of "ideology"), but the sentence as a whole is not a definition. Surprisingly, we found that some of these sentences are also present in the training set but with opposite labels as in the test set.

5 Conclusions

We have described our participation in SemEval-2020 Task 6 on Extracting Definitions from Free Text in Textbooks. In particular, we participated in Subtask 1, where the aim was to classify a given sentence as definitional or not. We evaluated to use of LSTMs and we compared different strategies for fine-tuning a pre-trained BERT language model. We found the latter to be more effective, especially when we fine-tuned the model twice. In particular, the BERT model is then first fine-tuned on the full training set and then fine-tuned further towards the target domain.

Acknowledgements. This work was supported by ERC Starting Grant 637277.

References

Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2005. Generic soft pattern models for definitional question answering. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 384–391.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Luis Espinosa-Anke and Steven Schockaert. 2018. Syntactically aware neural architectures for definition extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 378–385.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–1780.

Sentence	Label
5847 . While some Americans disapprove of partisanship in general , others are put off by the ideology — established beliefs and ideals that help shape political policy — of one of the major parties .	1
6246 . The current relationship between the U.S. government and Native American tribes was established by the Indian Self - Determination and Education Assistance Act of 1975 .	1
For example , Senator Ted Cruz (R - TX) announced his 2016 presidential bid at Liberty University , a fundamentalist Christian institution .	1
6264 . Conservative governments attempt to hold tight to the traditions of a nation by balancing individual rights with the good of the community .	1
6858 . Through their own constitutions and statutes , states decide what to require of local jurisdictions and what to delegate .	1
The wall was erected in 1963 by East Germany to keep its citizens from defecting to West Berlin .	1
5978 . The federal government responded by enacting the Force Bill in 1833 , authorizing President Jackson to use military force against states that challenged federal tariff laws .	1
In 1895 , in United States v. E. C. Knight , the Supreme Court ruled that the national government lacked the authority to regulate manufacturing .	1
5812 . For example , food , clothing , and housing are provided in ample supply by private businesses that earn a profit in return .	1
According to the doctrine of comparable worth , people should be compensated equally for work requiring comparable skills , responsibilities , and effort.	0
Through a talk program or opinion column , the elite commentator tells people when and how to react to a current problem or issue .	0
6525 . The Democratic Party emphasized personal politics , which focused on building direct relationships with voters rather than on promoting specific issues .	0
Volunteers in Service to America was a type of domestic Peace Corps intended to relieve the effects of poverty .	0
In Schechter Poultry Corp. v. United States (1935) , the Supreme Court found that agency authority seemed limitless .	0
6984 . In top - down implementation , the federal government dictates the specifics of the policy , and each state implements it the same exact way .	0
In bottom - up implementation , the federal government allows local areas some flexibility to meet their specific challenges and needs .	0
6851 . Despite the Constitution 's broad grants of state authority , one of the central goals of the Anti - Federalists , a group opposed to several components of the Constitution , was to preserve state government authority , protect the small states , and keep government power concentrated in the hands of the people .	0
6854 . Just three decades later , during the 1964 presidential election campaign , incumbent President Lyndon B. Johnson declared a " War on Poverty , " instituting a package of Great Society programs designed to improve circumstances for lower - income Americans across the nation .	0

Table 4: Examples of the incorrectly classified sentences from the Government domain.

Yiping Jin, Min-Yen Kan, Jun Ping Ng, and Xiangnan He. 2013. Mining scientific terms and their definitions: A study of the acl anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790.

Judith L Klavans and Smaranda Muresan. 2001. Evaluation of the definder system for fully automatic glossary construction. In *Proceedings of the AMIA Symposium*, page 324. American Medical Informatics Association.

Melanie Reiplinger, Ulrich Schäfer, and Magdalena Wolska. 2012. Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 55–65. Association for Computational Linguistics.

Sasha Spala, Nicholas A Miller, Yiming Yang, Franck Dernoncourt, and Carl Dockhorn. 2019. Deft: A corpus for definition extraction in free-and semi-structured text. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 124–131.

Sasha Spala, Nicholas Miller, Franck Dernoncourt, and Carl Dockhorn. 2020. Semeval-2020 task 6: Definition extraction from free text with the deft corpus. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2019. A joint model for definition extraction with syntactic connection and semantic consistency. *arXiv preprint arXiv:1911.01678*.

Eline Westerhout. 2009. Definition extraction using linguistic and structural features. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 61–67. Association for Computational Linguistics.