MULTI-INSTANCE LEARNING FOR RHETORIC STRUCTURE PARSING

S.S. Volkov[1,2,a], D.A. Devyatkin[1], A.V. Shvets[3]

[1] *Federal Research Center "Computer Science and Control" RAS, Moscow, Russia*

[2] *Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation*

[3] *Universitat Pompeu Fabra (UPF), Barcelona, Spain*

E-mail: [a] volkserg1@gmail.com

It would be helpful to consider various topic-independent features: syntax, semantics, and discourse relations between text fragments to accurately detect texts containing elements of hatred or enmity. Unfortunately, methods for identifying discourse relations in the texts of social networks are poorly developed. The paper considers the task of classification of discourse relations between two parts of the text. The RST Discourse Treebank dataset (LDC2002T07) is used to assess the performance of the methods. Since the size of this dataset is too small for training large language models, the work uses a model-pre fitting approach. Model pre-fitting is performed on a Reddit user comment dataset. Texts from this dataset are labeled automatically. Since automatic labeling is less accurate than manual marking, we use the multiple-instance learning (MIL) method to train models. A distinctive feature of modern language models is the large number of parameters. Using several models at different levels of such a text analyzer requires a lot of resources. Therefore, for the analyzer to work, it is necessary to use high-performance or distributed computing. The use of desktop grid systems can attract and combine computing resources to solve this type of problem.

Keywords: discourse analysis, multiple-instance learning, natural language processing, desktop grid

Sergey Volkov, Dmitry Devyatkin, Alexandr Shvets

1. Introduction

Social media are the most significant information source and the essential communication tool on the Internet, and this information can be aimed at discriminating against people. However, the topic of those messages can transform a lot, depending on the region and political factors. Therefore, pure topic-related text features like lexis are not helpful in building practical tools to filter those messages. At the same time, those messages often utilize some techniques of manipulation to develop hatred. We believe those techniques can be recognized with discourse features, which can be obtained via discourse analysis. The basis of that analysis is Rhetorical structure theory (RST). RST is a theory of text organization that describes relations that hold between parts of the text. Unfortunately, RST parsers are poorly developed, especially for social media text analysis. In this paper, we tackle discourse relation classification in social media that is a crucial part of discourse analysis. Namely, we answer the following research questions.

1. Can pre-training on automatically labeled social media corpora help to improve the accuracy of discourse analysis?
2. Does the multiple-instance approach on the pre-training step improve the accuracy of discourse analyzers?

2. Related work

The most studies on the discourse analysis use small labeled corpora such as Penn Discourse Treebank (PDTB) 2.0 and 3.0, and TED Multilingual Discourse Bank (parallel corpus) [1]. The peculiarity of these corpora is their small size, which is due to the complexity of the labeling and the large number of classes of discursive relations. The small size of these packages limits the applicability of complex models containing a large number of parameters. Therefore, approaches are being actively developed that use pre-training on related problems. For example, in [2], the author used a multilayer neural network trained on the PDTB corpus and tested on the TED corpus. In this case, the method of generating cross-language LASER embeddings was used [3]. In the paper [4], the generation of a text discourse scheme is proposed as a problem of pre-training a language model. This model is a multilayer network composed of bidirectional recurrent LSTM (Long-Short Term Memory) layers. The novelty of the work lies in the approach proposed for building the training corpus. First, a simple rule-based parser is used to detect explicit relations, then those relations are used to pre-train the models. Experiments have shown that both models, after training, also make it possible to reveal implicit discursive relationships. We believe the future development of that approach lies in applying a multi-instance learning (MIL) approach [5]. It is a type of supervised learning. Instead of receiving a set of instances which are individually labeled, model receives a set of labeled bags, each containing many instances. There are plenty of studies where MIL is utilized to tackle noisy labeling, for example, in case of distant learning [6]. Therefore, in this study we applied MIL-pre-training on automatically labeled social media corpus and tested the obtained models on the gold RST Discourse Treebank corpus.

3. Datasets and methods

3.1 Datasets

The primary dataset for evaluating the quality of the algorithms is RST Discourse Treebank (LDC2002T07) [7] – gold dataset. The dataset consists of 347 documents for training and 38 documents for testing. These documents were marked up manually. As a result, the data is a set of annotated parts of the text and their relationships (18537 samples for training and 2255 samples for test). The main disadvantage of this set is the small number of samples for training algorithms. For this reason, it will only be used for model fine-tuning.

For basic training of the models, we decided to use a large amount of automatically marked data from user's comments of Reddit news portal (2003-2018). These texts were divided into pairs of

connected discourse units using fast rhetorical theory discourse parser [8]. This way, we got about 16 million pairs of connected discourse units. Before training, this dataset was filtered to correct the class balance. Thus the huge part of the most common classes was discarded. After preprocessing the dataset contains 176677 records, which are balanced for 31 classes. The parser is able to retrieve more classes, but we decided to use only those classes that are represented in the gold dataset.

3.2 Methods

The main task of the discourse classification is the following. Model receives two clauses at the input and should predict the type of relationship between these clauses. During the research, several models were analyzed. The first model is based on Gated Recurrent Units (GRU) [9] layers. StanfordNLP tokenizer splited text into tokens for this model. As a result of tokenization, each word was represented as a lemma (a dictionary form of the word). After that, word2vec model vectorizes each lemma. So that way, we got a set of clauses which are represented as vectors array. The neural network receives two such arrays of vectors as input and must predict the type of discourse relationship between them. The model has the following architecture:

- Two input layers correspond to two input sentences (250 vectors of dimension 300).
- Add-layer. It takes as input two tensors and returns a single tensor with the same shape.
- GRU -layer. 256 neurons with dropout = 0.1.
- Self-attention layer. Attention width = 256, attention_activation = sigmoid.
- Second GRU -layer with return_sequences = False, dropout=0.1.
- Dense layer. 64 neurons, sigmoid activation function.
- Output layer which corresponds to considered discourse relation classes. Softmax activation function.

The second model is based on Bidirectional Encoder Representations from Transformers (BERT) [10]. For the experiments we used pre-trained google model – "bert_uncased_L-12_H-768_A-12". This model has 110 million parameters and was pretrained on a large corpus (Wikipedia + BookCorpus) by google. To preprocess the dataset for this model, a special bert-tokenizer was used. The model was built as a core of a new neural network with following architecture:

- Input layer, which receives a list of tokens ids and a list of token type ids. The model sequentially takes two lists of tokens related to the two clauses, which are separated by a special token. Maximum sequence length is 512. Token type ids is a list which for the first clause has all its tokens represented by a 0, whereas tokens, corresponding to the second clause, represented by a 1.
- Bert layer. BERT model represented as a layer.
- Dense layer. 256 neurons. ReLU activation function.
- Output layer which corresponds to considered discourse relation classes. Softmax activation function.

The third model is modification of the second model. For this case, we used a different approach to training the model – Multi-instance learning (MIL). In the MIL instead of receiving a set of instances, which are individually labeled, model receives a set of labeled bags, each containing many instances. The bag has the same label as most of the instances in the bag. The percentage of positive instances in a bag is adjusted using a coefficient. To train the model in this way, let's define a loss function (1):

$$loss = \gamma CCE(Y_{pred}, Y_{true}) + (1-\gamma)CCE(Y_{pred}, Y_{bag}) \quad (1)$$

where *CCE* – Categorical Cross-entropy, γ – balance coefficient, $Y_{pred}, Y_{true}, Y_{bag}$ – corresponding model predictions, instance labels and bag labels.

The main purpose of using MIL is to improve the quality of poorly balanced classification. Another important aspect is the fact that we use automatically labeled data to train the model; therefore, there is a chance that some labels will not actually correspond to the correct class. MIL can train the model on noisy data because it can handle a bag that contains a part of the negative instances.

4. Results

The first task is to compare the two main models - BERT and GRU. Both models were trained on automatically marked up dataset, and then fine-tuned on a gold dataset. Table 1 shows the F1-score of classification result on gold dataset (LDC2002T07) for top-5 classes.

Table 12. F_1 score of classification result for top-5 classes

GRU-based model		BERT-based model	
F_1	Class	F_1	Class
0.71	NS-Attribution	0.87	NS-Attribution
0.65	SN-Attribution	0.82	SN-Attribution
0.60	NS-Elaboration	0.73	SN-Condition
0.60	SN-Condition	0.63	NN-Same-Unit
0.48	NS-Enablement	0.54	NN-Joint

As a result of the initial comparison of the BI-GRU and BERT models, we can conclude that the BERT model performs better at this task. For this reason, the MIL approach will only be applied for the BERT model. Figure 1 shows the F1-score of classification result on gold dataset.

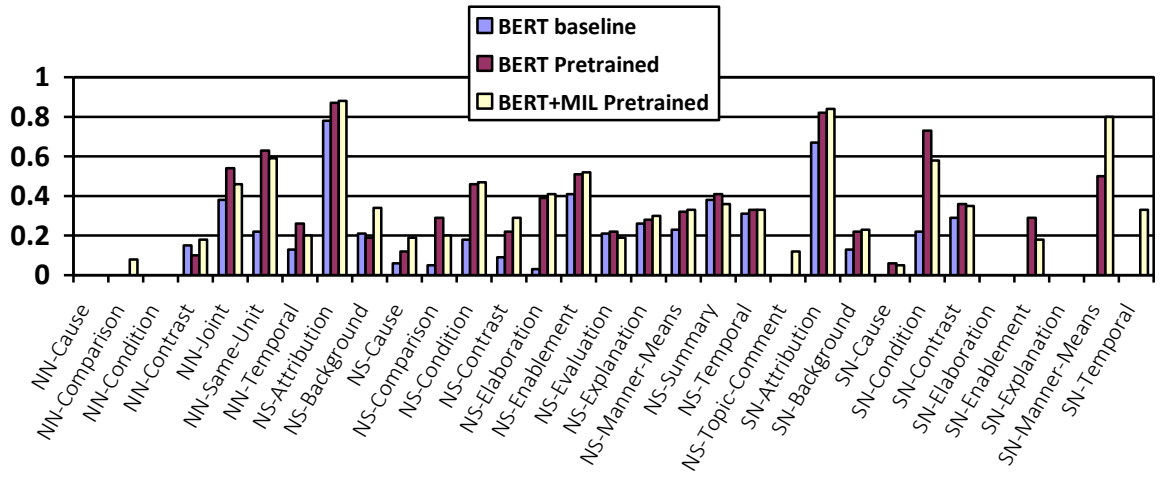

Figure 30. F_1-score of classification result on gold dataset

The figure 1 compares the 3 models. BERT baseline is the standard BERT model that has been only finetuned on gold dataset. Bert Pretrained is model that has been pretrained on automatically markuped data from Reddit, and then finetuned on gold dataset. The third model "BERT+MIL pretrained" has been pretrained on Reddit data using MIL algorithm, and then also finetuned on gold data without MIL. As we can see, for some classes MIL significantly increases the quality of the classification. Some classes that were impossible to define are now being detected, albeit with a slight accuracy.

5. Conclusion

In this paper, a comparison of several models for solving the problem of classifying discourse relations was presented. Experiments showed that the use of BERT-based models is most suitable for solving this problem. Also, as a result of experiments, it was found that the use of the MIL algorithm can increase the quality of the classification when using noisy data for pre-training. The improvement is especially noticeable in rare classes. Presented models of discourse relations classification can be used as part of the system for analyzing the emotional charge of the text. It is assumed that a text that contains an aggressive or negative context can be structured in such a way that the identification of discourse relations between its parts can increase the quality of detection of such texts. This text analysis system requires high-performance or distributed computing. The use of desktop grid systems can attract and combine computing resources to solve this problem.

6. Acknowledgement

This work was funded by RFBR according to the research project No. 21-011-44242.

References

[1] Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S. and Ogrodniczuk, M., 2019. TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. Language Resources and Evaluation, pp.1-27

[2] Kurfalı M., Östling R. Kurfalı, M. and Östling, R., 2019, September. Zero-shot transfer for implicit discourse relation classification. In Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (pp. 226-231).

[3] Artetxe, M. and Schwenk, H., 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7, pp.597-610.

[4] Nie A., Bennett E., Goodman N. DisSent: Learning sentence representations from explicit discourse relations //Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – 2019. – С. 4497-4510.

[5] Dietterich T. G., Lathrop R. H., Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles //Artificial intelligence. – 1997. – Т. 89. – №. 1-2. – С. 31-71.

[6] Le P., Titov I. Distant Learning for Entity Linking with Automatic Noise Detection //Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – 2019. – С. 4081-4090.

[7] Carlson L., Okurowski M. E., Marcu D. RST discourse treebank. – Linguistic Data Consortium, University of Pennsylvania, 2002.

[8] Heilman M., Sagae K. Fast rhetorical structure theory discourse parsing //arXiv preprint arXiv:1505.02425. – 2015.

[9] Cho K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation //arXiv preprint arXiv:1406.1078. – 2014.

[10] Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.