Bf3R at SemEval-2018 Task 7: Evaluating Two Relation Extraction Tools for Finding Semantic Relations in Biomedical Abstracts

Mariana Neves[1], **Daniel Butzke**[1], **Gilbert Schönfelder**[1,2], **Barbara Grune**[1]
[1] German Federal Institute for Risk Assessment (BfR)
Diedersdorfer Weg 1, 12277, Berlin, Germany
[2] Charité - Universitätsmedizin Berlin, Institute of Clinical Pharmacology and Toxicology,
Charitéplatz 1, 10117 Berlin, Germany
mariana.lara-neves@bfr.bund.de

Abstract

Automatic extraction of semantic relations from text can support finding relevant information from scientific publications. We describe our participation in Task 7 of SemEval-2018 for which we experimented with two relations extraction tools - jSRE and TEES - for the extraction and classification of six relation types. The results we obtained with TEES were significantly superior than those with jSRE (33.4% vs. 30.09% and 20.3% vs. 16%). Additionally, we utilized the model trained with TEES for extracting semantic relations from biomedical abstracts, for which we present a preliminary evaluation.

1 Introduction

Finding relevant publications for a certain topic is an important task daily carried out by most researchers in various domains, such as computer science or biomedicine. However, most information retrieval methods usually consider only words and terms (named entities) and do not usually profit from semantic relationships between these entities (Lu, 2011). Many approaches frequently consider words and entities as bags of words but do not take advantage from intrinsic properties of scientific texts, such as subsections (e.g., introduction, methods, results), common concepts (e.g., task, material) and relations between these concepts (e.g., model-feature, part-whole). However, extracting semantic relations from scientific text can potentially support finding relevant information for a certain topic by focusing on particular terms which participate in those relations. In addition, the relation type and the corresponding arguments provide further information regarding the role that a certain entity plays in the text.

We describe the experiments that we carried out during our participation in Subtask 2 of SemEval-2018 Task 7[1] (Gábor et al., 2018). The task consisted on the extraction of six semantic relations from scientific abstracts, namely: "USAGE", "RESULT", "MODEL", "PART_WHOLE", "TOPIC" and "COMPARISON". While the entities were given (and all belong to the general type "ENTITY"), participants of subtask 2 were required to identify the relations and classify these into one of the six types. All relations were asymmetrical (regarding their direction), except for "COMPARISON", and the identification of the direction of the relations was mandatory. The documents came from the ACL Anthology[2], thus belonged to the domain of computational linguistic, and were derived from a more comprehensive corpus which includes more relations than the ones under evaluation in the challenge (Gábor et al., 2016).

Our contribution in this work is two-fold: (a) we experimented with two available relation extraction (RE) tools in the context of the Subtask 2 of SemEval-2018 Task 7; and (b) we evaluated the models trained on the task data for the extraction of the relations mentioned above from biomedical abstracts. In the next section, we present a short overview of related work, followed by a detailed description of our methods in section 3, the results that we obtained both during the development and official evaluation phases (section 4) and the discussion of our results and the preliminary experiments with biomedical publications.

2 Related Work

Despite the importance of the task, few previous work has focused on the identification of semantic elements in publications. Document zoning is probably the task that more attention received in

[1] https://competitions.codalab.org/competitions/17422
[2] http://aclweb.org/anthology/

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 816–820
New Orleans, Louisiana, June 5–6, 2018.

the last years and covered the identification of sections both in abstracts (Hirohata et al., 2008) and full text (Liakata et al., 2012). A more comprehensive study and comparison of different schemes for zoning was carried out in (Guo et al., 2010).

Regarding automatic extraction of scientific relations, many researchers have proposed various scheme based on either (or both) concepts or relations. For instance, (Gupta and Manning, 2011) proposed the annotation of the focus, technique and domain in scientific publications. A more comprehensive schema was proposed by (Tateisi et al., 2016), who developed an ontology of semantic structures in research articles. More recently, the ScienceIE task in SemEval-2017 (Augenstein et al., 2017) proposed the automatic extraction of both entities (Task, Process and Material) and relations from scientific abstracts. Finally, (Gábor et al., 2016) recently proposed a schema of about 20 relations, some of which are considered in the challenge.

The current task focuses on relation extraction and classification, for which many approaches have been proposed in the past years and for which some tools are readily available, including the ones we describe here. Similar to other natural language processing (NLP) tasks, recent work has shifted to the use of neural networks, as reported in (Nguyen and Grishman, 2015) and (Sorokin and Gurevych, 2017).

3 Methods

We evaluated the performance of two existing tools for relation extraction, namely, jSRE (Giuliano et al., 2006) and TEES (Björne et al., 2012). Both tools can be trained for any RE task, provided that a corpus in the appropriate format is available. The methods behind jSRE utilize kernel methods, features derived from shallow linguistic information and both global (sentence level) and local (regarding the relations) contexts (Giuliano et al., 2006). TEES trains support vector machines algorithms using a variety of features derived from the sentence, tokens and dependency chains (Björne et al., 2012).

The workflow of our experiments is shown in Figure 1. After parsing the corpus provided by the organizers, we performed standard NLP preprocessing, followed by preparing input files in the appropriate format required by the two RE tools. This included the generation of negative examples, which are necessary for the jSRE tool. After training the models with each tool and classifying the test documents, we merged predictions (only for jSRE) and printed the predicted relations in the format required by the challenge.

Corpus reader. The main corpus was provided in two files: (a) one XML file which includes the text and entities (all belonging to the general format "ENTITY"); and (b) one file in plain text format with the list of the positive relations and one of the corresponding types listed above. For reading the data, we utilized the BioC format (Comeau et al., 2013). Our model also includes the identification of the direction of the relation (which is necessary for the task).

Pre-processing. We processed all documents using the Stanford CoreNLP library[3] (Manning et al., 2014), including sentence splitting, tokenization, part-of-speech (POS) tagging, chunking, constituency parsing and dependency parsing. While jSRE is based on shallow parsing, TEES relies on both dependency and constituency parsing.

Corpus preparation. We prepared the input format required by each RE tool as specified in their documentation. For jSRE, we generated plain text files for each relation type. These included the original tokens, lemma, POS tagging, indicative of participation in the relation (T or O, otherwise) and the relation category. For multi-token entities, the corresponding tokens should be merged into one, e.g., "storage_media_and_networks" instead of the four individual tokens. The relation category were the following: -1 (unknown), 1 (positive), (2 positive reverse) and 0 (negative). For TEES, we generated a combined XML file which included complete pre-processing analysis (sentences, tokens, full parse tree, constituents and dependency parsing), as well as entities and relations (including their types). The indicative for the direction of the relation is provided as an attribute.

Negative examples generation. We automatically generated negative examples for jSRE. We produced negative examples for each pair of entities following the following guidelines: (a) the entities should belong to the same sentence (according to the sentence splitting analysis); (b) just one

[3] https://stanfordnlp.github.io/CoreNLP/

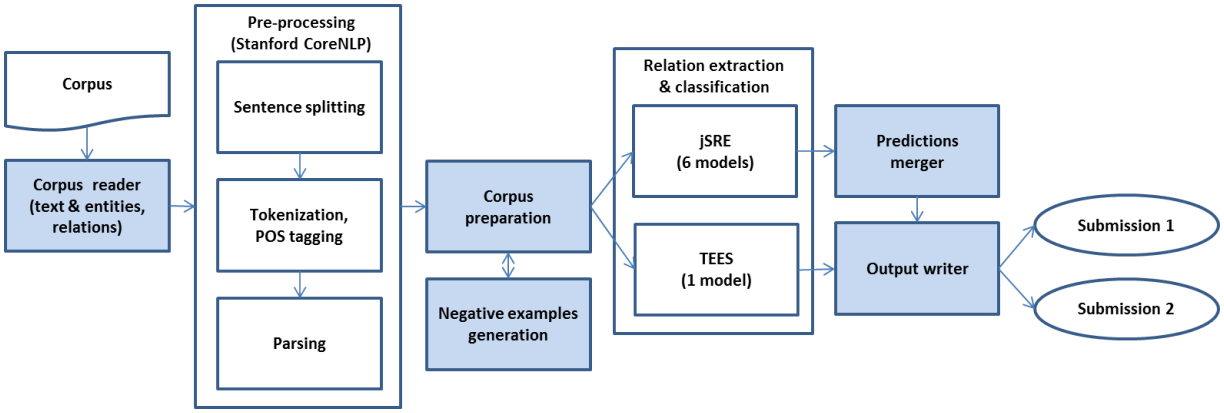

Figure 1: Workflow of the components in our approach. The components that we developed are displayed in blue.

training example (negative or positive) for each pair (always in the order that these appear in the sentence). We generated negatives examples also for sentences which contain no positive example at all. This step was not necessary for TEES, as the tool includes it in its training procedure.

Relation extraction and classification. Provided the training and development files in the format required by each tool, we trained the system according to their documentation. In the case of jSRE, we build six models, one for each relation type. Each model from jSRE calculates scores for one of the categories: 1 (positive), 2 (positive reverse) and 0 (negative). We tried the three kernels included in jSRE (LC - Local Context, GC - Global Context and SL - Shallow Linguistic), and we obtained best results with the later. In the case of TEES, we only trained one model, which performs the both the automatic extraction and classification into a category.

Predictions merger. This component is only necessary for jSRE and it consisted on reading the prediction for each of the categories (0, 1 or 2) from each of the six models and choosing the one that scored higher.

Output writer. We converted the output from both tools to the output (submission) format required by the challenge. We also checked whether a reverse relation was predicted for the "COMPARISON" type and avoided printing it, given that this is a symmetrical relation.

4 Data and Results

The training set released by the organizers consisted of 350 documents which we split in the following datasets: 250 for training, 50 for tuning (only for TEES) and 50 for development test. The whole dataset contained the following distribution of relation types which appear in 342 (out of 350) documents: 483 for "USAGE", 326 for "MODEL", 234 for "PART_WHOLE", 95 for "COMPARE", 72 for "RESULT" and only 18 for "TOPIC". Our evaluation during the development of the system (over the development test dataset) is shown in Table 1. We did not obtain predictions for the "TOPIC" from none of the RE tools, given the low frequency of this relation in the training set. Indeed, only three instances of this relation type are present in the development set.

Regarding the official test set, the organizers released 152 documents for this aim. All our submissions were based on models trained only on the 250 documents, i.e., we did not train models based on the totality of the 350 documents. Our official results for Subtask 2 is shown in Table 2.

5 Discussion

Relation extraction and classification. We tried two available RE tools for extracting semantic relations from scientific publications. TEES performed significantly superior to jSRE and we chose to use this tool for our further experiments with biomedical publications (cf. below). However, the performance of TEES is rather low in comparison to the best results in the challenge (cf. Table 4). Finally, we did experiment with a simple union of predictions generated by both tools, but adding the predictions from jSRE only harmed the performance of TEES (cf. Table 1).

With respect to both tools, we found TEES easier to use and run, also given our previous experience with it (Thomas et al., 2013). Addition-

Tool	USAGE	RESULT	MODEL	PART_WHOLE	COMPARISON	TOPIC
TEES (t)	29.63%	38.10%	26.23%	26.32%	28.57%	0.00%
jSRE (j)	16.67%	12.50%	24.56%	6.67%	20.00%	0.00%
(t) + (j)	19.58%	28.57%	24.62%	20.00%	28.57%	0.00%

Table 1: Results for each category for the development set (50 documents).

Tool	Extraction		Classification	
	D	T	D	T
TEES (t)	44.69%	33.4%	25.45%	20.3%
jSRE (j)	22.32%	30.9%	15.03%	16.0%
(t) + (j)	37.63%	-	20.88%	-

Table 2: Results for Sub-task 2 of SemEval-2018 Task 7, for the extraction and classification tasks, for both development (D) and official test (T) sets. The highest F1 in the official test set were 50% and 49.3% for the extraction and classification tasks, respectively.

ally, we found the input format from jSRE harder to process. On the other hand, TEES requires full parsing while jSRE is based on shallow parsing. Finally, TEES is readily available for download while we needed to contact the developers of jSRE in order to get a copy of it and needed to do some changes on the code in order to run it. Changes on the code were also necessary in order to obtain scores (probabilities) for the various categories and thus, obtain the predicted relation type. TEES, on the other hand, supports both relation extraction and classification by default.

Semantic relations in biomedical abstracts. We experimented with the model trained on TEES to extract the same semantic relations from biomedical abstracts. Our aim was to evaluate whether the predicted relations is part of either the research goal or the methods in the publication. In particular, we were interested in assessing whether the relations could potentially support the automatic extraction of either an animal experiment or an alternative method to animal experiment (e.g., in vitro or in silico experiments) (Liebsch et al., 2011). This information could later support the automatic identification of abstracts which describe either of the two experiments (animal or alternative to animal).

We processed a set of 161 abstracts retrieved from PubMed[4]. We followed the same workflow showed on Figure 1 only that we performed NER on the abstracts using the Metamap tool (Aronson and Lang, 2010). We used exactly the same model (from TEES) that we used to predict relations for the official test set of the challenge. We obtained a total of 241 relations from 108 abstracts (out of a total of 161). The number of relations detected for each type were the following: 99 for "MODEL", 87 for "USAGE", 30 for "PART_WHOLE", 22 for "RESULT", two for "COMPARISON" and one for "TOPIC".

We manually checked 28 relations detected from a sample of 13 abstracts. During these attempts, the definitions of the semantic relations as provided by the organizers gave much room for individual interpretations by the evaluating researcher. Being aware of this possible pitfall, however, we judged 9 out of 28 suggested relations as correct.

6 Conclusions

During our participation in the SemEval-2018 Task 7, we experimented with two available relation extraction tools - jSRE and TEES. As future work, we plan to run additional experiments with the current tools, such as using the totality of the training data as well as combination of the systems, as carried out in (Thomas et al., 2013). Additionally, we plan to use additional tools, such as ones based on neural networks (Nguyen and Grishman, 2015).

We applied the generated model from TEES for extraction of semantic relations from biomedical abstracts. Our manual evaluation of some of those relations shows that these have the potential to support the identification of the methods that are part of the research goal. We now plan to run a comprehensive evaluation based on a larger collection of biomedical abstracts as well as a task-specific assessment.

Acknowledgments

We would like to thank Jari Björne for support with TEES, Alberto Lavelli for providing a copy of jSRE and Roland Roller for fruitful discussions on the generation of negative examples.

[4] http://www.ncbi.nlm.nih.gov/pubmed/

References

Alan R Aronson and Franois-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17(3):229–236.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 546–555.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of Turku in the Bionlp'11 Shared Task. *BMC Bioinformatics* 13(11):S4.

Donald C. Comeau, Rezarta Islamaj Doan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. 2013. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database* 2013:bat064.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA.

Kata Gábor, Haifa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the acl anthology corpus for the automatic analysis of scientific literature. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*. Trento, Italy.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, BioNLP '10, pages 99–107.

Sonal Gupta and Christopher D. Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *In Proceedings of IJCNLP*.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *In Proc. of the IJCNLP 2008*.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28(7):991–1000.

Manfred Liebsch, Barbara Grune, Andrea Seiler, Daniel Butzke, Michael Oelgeschläger, Ralph Pirow, Sarah Adler, Christian Riebeling, and Andreas Luch. 2011. Alternatives to animal testing: current status and future perspectives. *Archives of Toxicology* 85(8):841–858.

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database* 2011:baq036.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Thien Huu Nguyen and Ralph Grishman. 2015. *Relation extraction: Perspective from convolutional neural networks*.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1784–1789.

Yuka Tateisi, Tomoko Ohta, Sampo Pyysalo, Yusuke Miyao, and Akiko Aizawa. 2016. Typed entity and relation annotation on computer science papers. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Philippe Thomas, Mariana Neves, Tim Rocktäschel, and Ulf Leser. 2013. Wbi-ddi: Drug-drug interaction extraction using majority voting. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, pages 628–635.