LasigeBioTM team at CLEF2020 ChEMU evaluation lab: Named Entity Recognition and Event extraction from chemical reactions described in patents using BioBERT NER and RE

Pedro Ruas, Andre Lamurias, and Francisco M Couto

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon 1749-016, Portugal
psruas@fc.ul.pt
alamurias@lasige.di.fc.ul.pt
fcouto@di.fc.ul.pt

Abstract. This manuscript describes the participation of the LasigeBioTM team in the NER and EE tasks of the ChEMU evaluation lab. We have fine-tuned the BioBERT NER model to locate and tag named entities and the BioBERT RE model to detect relations between trigger words and named entities. For the NER task, we obtained a F1-score of 0.9392 (exact matching) and 0.9630 (relaxed matching), which was an improvement over the baseline approach and achieving the 3rd best team result. For the EE task, we were not able to produce all the required annotation files due to the dimension of the test set and, consequently, we did not obtain results in time to submit to the competition. However, we obtained an accuracy of 0.9849 when we applied the BioBERT RE model on the development set.

1 Introduction

Chemical patents are a valuable source of information for chemical research. Every year, thousands of patents are registered, increasing the already large wealth of information available. Considering only the European Patent Office (EPO) and the year 2019, there were 7697 new patents filled in the "Pharmaceuticals" category and 6197 in the "Organic fine chemistry" category[1]. The manual analysis of these documents is costly, both in terms of time and effort, so it is necessary to develop text mining approaches to extract information in a more efficient way.

There are some challenges associated with patent text, like the presence of longer sentences, the use of specific terminology, the complexity of the syntactic

[1] http://documents.epo.org/projects/babylon/eponet.nsf/0/BC45C92E5C077B10C1258527004E95C0/

Category	Entity type
Chemical entity	"REACTION_PRODUCT"
	"STARTING_MATERIAL"
	"REAGENT_CATALYST"
	"SOLVENT"
	"OTHER_COMPOUND"
Reaction property	"TEMPERATURE"
	"TIME"
	"YIELD_PERCENT"
	"YIELD_OTHER"
Reaction label	"EXAMPLE_LABEL"

Table 1. Entity types and the respective broad category in which they are included.

structure, which limits the performance of general text mining models, usually developed for news text [3]. In addition, chemical patents typically contain a large number of chemical compounds with a complex structure, which originates ambiguity if, for example, we aim to link those entities to a reference knowledge base [1].

The ChEMU evaluation lab [7] proposed the chemical named entity recognition (NER) task and the chemical reaction event extraction (EE) task on a large corpus of chemical patents.

Language representation models, like Bidirectional Encoder Representations from Transformers (BERT) [2], have shown state-of-the-art performance across several natural language processing tasks, including Named Entity Recognition (NER) and Relation Extraction (RE). The goal of NER is to find in a given text the location of named entities and to classify them according to pre-defined types. Several works have described the application of BERT to the NER task [6, 9]. In turn, the goal of the RE task is the detection and classification of semantic relations between two given entities in a text, and BERT has also been applied to this task [8].

We describe here our approach to Task 1 (NER) and Task 2 (EE) of ChEMU, which consisted of the application of the BioBERT NER model for Task 1 and modelling Task 2 as a joint NER + RE task, in which we applied the BioBERT NER and the BioBERT RE models.

2 Methodology

2.1 Task 1 - NER

Data preparation The objective of the task was to recognise named entities related to chemical reactions in patents and to tag them according to ten different types:

We used a rule-based tokenizer proposed by one of the BioBERT authors[2] which uses regular expressions. We started by tokenizing the text of the docu-

[2] `https://github.com/dmis-lab/biobert/issues/107#issuecomment-615558492`

ments belonging to the released train (900 documents) and development (225 documents) sets. Each token was then tagged according to the IOB2 notation. The separator between each sentence was an empty line. For each set, there was a file that included all tokens, one per line, and the respective tags. These files were the input for fine-tuning the model we have used, which is detailed in the next section. When the test set was released, we applied the same process to the respective documents.

Model We used BioBERT [5], which consists of the BERT model pre-trained on several general corpora (English Wikipedia and BooksCorpus) and additionally in biomedical-specific corpora (PubMed abstracts and PMC full-text articles). The results reported by the authors show that BioBERT achieves better performance in the NER of biomedical entities in comparison with the classical implementation of BERT [5].

We fine-tuned the BioBERT NER model using the files corresponding to the competition datasets converted to the IOB2 notation, more concretely, 900 documents in the train set and 113 randomly chosen documents in the development set. We changed the training batch size from 32 to 24 to lower the memory requirements. To lower the required time for training, we used the pre-trained weights "BioBERT-Base v1.0 (+ PubMed 200K)", which corresponds to the smallest vocabulary available for BioBERT. The number of training epochs, the learning rate and the maximum sequence length were set to the default values, respectively, 10.0, 1^{-5} and 128. We did not have time to explore different values for the referred hyper-parameters, so we opted for the default values as it was the safest approach (with exception of the training batch size).

Then, we applied the fine-tuned model to recognise the entities and to predict the respective type in the remaining 112 documents of the development set that had not been previously used (inference mode). We used the same values for the hyper-parameters. We developed a module to process the BioBERT NER output and to generate the respective annotation files according to the BRAT standoff format[3]. We submitted the resulting annotation files to the competition page and obtained a F1-score of 0.9524 using the exact matching criterion and a F1-score of 0.9904 using the relaxed matching criterion. Given these results, we fine-tuned again the model, but this time using all documents belonging to the train (900) and the development (225) sets.

With the release of the test set, we applied the fine-tuned model to the converted file containing the tokenized text. We maintained the previous values used in inference mode for all hyper-parameters, with the exception of the maximum sequence length, which we changed from 128 to 384, due to the presence of longer sentences in the test set when comparing with the train and the development sets.

[3] `https://brat.nlplab.org/standoff.html`

2.2 Task 2 - Event extraction

Data preparation Our approach was to model the Task 2 as a joint NER and RE task. The goal was to detect trigger words and to recognise arguments involving trigger words and the entities described in Task 1. We followed a similar approach for the Task 1 and converted both the documents of the train and the development sets to the IOB2 format. But in this case, we only considered the annotations relative to event trigger words, i.e., words associated with individual steps in the context of the reaction. These event trigger words belonged to two additional entity types not present in Task 1: "REACTION_STEP" and "WORKUP". Like in the Task 1, for each set there was a file that included all tokens and the respective tags.

For the RE part of the task, there were two labels for a relation between an event trigger words and a chemical entity: "ARG1", to label a relation between a trigger word and a chemical entity, and "ARGM", to label a relation between a trigger word an adjunct entity, like temperature, yield or time. First, we performed sentence segmentation of the text present in the documents of the train and the development sets and, for each sentence containing at least a trigger word and an entity, we assumed that it could potentially contain a relation. For sentence segmentation, we used the same script referred in the Task 1, which consists of a rule-based model proposed by one of the BioBERT authors. In each sentence, the trigger word and the entity were replaced, respectively, by the tags "@TRIGGER$" and "@LABEL$". If the trigger word and the entity were effectively part of an argument in the gold standard annotations, the sentence would be assigned the label "1", otherwise the label would be "0". Besides, if in a given sentence, for example, there were present a trigger word and two different entities, the sentence would appear in two different lines of the final file, each one associated with a trigger word - entity pair. At the end, for each set we obtained a file containing a sentence per line, with the respective relation label and the tags @TRIGGER$ and @LABEL$ in the correct position.

Model For the NER step, we fine-tuned the BioBERT NER model using the files corresponding to the documents of the train (900) and development (225) sets converted to the IOB2 notation. We used the same hyper-parameter values for fine-tuning as described in Task 1. The documents of the test set were also converted into a single file according to the IOB2 notation and the approach was similar to that of Task 1.

For the RE step, we considered the files containing the sentences of the train in the format described above to fine-tune the BioBERT RE model. We used the pre-trained weights "BioBERT-Base v1.0 (+ PubMed 200K) and changed the batch size from 32 to 24. The number of training epochs, the learning rate and the maximum sequence length were set to the default values, respectively, 3.0, 2^{-5} and 128. We evaluated the fine-tuned model on the development set and obtained an accuracy of 0.9849. We fine-tuned again the model using all the sentences in the train and development sets. We applied the fine-tuned model to predict the relation labels in the converted documents of the test set only

	Exact evaluation			Relaxed evaluation		
Model	Precision	Recall	F1-score	Precision	Recall	F1-score
Baseline	0.9071	0.9071	0.8893	0.9219	0.8723	0.9053
LasigeBioTM	**0.9327**	**0.9457**	**0.9392**	**0.9590**	**0.9671**	**0.9630**

Table 2. Task 1 (NER) evaluation using the exact and the relaxed matching criteria considering all entity types. The performance of the baseline approach and our approach are shown in terms of Precision, Recall and F1-score. The highest values for each metric in the exact and relaxed evaluations are highlighted.

	Exact evaluation			Relaxed evaluation		
Entity type	Precision	Recall	F1-score	Precision	Recall	F1-score
EXAMPLE_LABEL	0.9577	0.9742	0.9659	0.9718	0.9885	0.9801
OTHER_COMPOUND	0.9529	0.9534	0.9531	0.9648	0.9658	0.9653
REACTION_PRODUCT	0.8387	0.8877	0.8625	0.9204	0.9498	0.9349
REAGENT_CATALYST	0.8887	0.8869	0.8878	0.9145	0.9091	0.9118
SOLVENT	0.9410	0.9696	0.9551	0.9433	0.9720	0.9574
STARTING_MATERIAL	0.8920	0.9058	0.8988	0.9460	0.9421	0.9440
TEMPERATURE	0.9674	0.9706	0.9690	0.9870	0.9934	0.9902
TIME	0.9846	0.9889	0.9868	0.9799	0.9955	0.9876
YIELD_OTHER	0.9732	0.9909	0.9820	0.9799	0.9955	0.9876
YIELD_PERCENT	**0.9897**	**0.9923**	**0.9910**	**0.9974**	**1.0000**	**0.9987**

Table 3. Task 1 (NER) evaluation using the exact and relaxed matching criteria according to each entity type. The performance of our approach is shown in terms of Precision, Recall and F1-score. The highest values for each metric in the exact and relaxed evaluations are highlighted.

changing the maximum sequence length from 128 to 384. At last, we developed a module to import the output of the RE model of BioBERT and to determine the type of the argument: "ARG1" if the relation was between a trigger word and a chemical entity, "ARGM" if the relation was between a trigger word and an adjunct entity.

3 Results

The evaluation results for Task 1 (NER) are available in Table 2 and the Table 3 shows the results for the task according to each entity type.

For Task 2 (EE) we did not obtain any results for the test set, but we obtained an accuracy of 0.9849 when we applied the BioBERT RE model on the dev set.

4 Discussion

4.1 Task 1 (NER)

Overall, the results obtained for this task, both using exact (F1-score of 0.9392) and relaxed matching (F1-score of 0.9630) criteria were positive. Comparing with the baseline approach, we obtained a higher F1-score, both considering the exact matching (+0.0499) and the relaxed matching (+0.0577) criteria. These results represent the 3rd best position in terms of team results and the 5th best position in the overall submission rank.

The good performance of the BioBERT model is due to the fact that it produces contextualised word representations that consider both the left and the right contexts of the words. The meaning of the words is usually related with the context where they appear, i.e., a given word can have two (or more) different meanings in different contexts. This is particularly relevant for chemical compounds, which can have different roles according to the reaction (i.e. the context) in which they participate. The BioBERT NER model obtained higher F1-score when recognising the entities of the type "YIELD_PERCENT", both using the exact (0.9910) and the relaxed matching (0.9987) criteria. This is related with the fact that this type of entities always included the character "%" in their surface form after a numerical value (for example, "53%"), which was an immutable pattern easily recognisable. In the other hand, the model obtained the lowest results when recognising the entities of the type "REACTION_PRODUCT" using the exact matching criteria (F1-score of 0.8625) and the entities of the type "REACTION_CATALYST" using the relaxed matching (F1-score of 0.9118). There was more ambiguity associated with these two types of entities (and also with the entities of the type "STARTING_MATERIAL") since they were related with chemical compounds, which can have different roles. This means that, for example, the chemical compound "4-nitropyridin-2-amine" can be the reaction product of a given reaction but in other reaction can participate as the reaction catalyst or as the starting chemical compound. The BioBERT NER model was able to correctly recognise almost all entities belonging to these types, however, the fact that it was originally pre-trained on scientific articles and general corpora and not on patents including chemical compounds, prevented an even higher performance.

4.2 Task 2 (EE)

The test set was too large (10000 documents) when comparing with the other sets, and the text of the documents was significantly larger too: the average number of characters present in the documents of the test set was 61002, whereas in the documents of the train and development sets was 835 and 772, respectively. This difference increased the required time to fine-tune and apply our model. Due to the limited time participants had to submit the results, we were only able to produce annotations for 1126 documents out of the necessary 10000 comprising the test set.

The entities and events to extract were located in the description of chemical reactions within the patents text. So the inclusion of a module able to detect the chemical reactions within the text would filter out irrelevant text and, consequently, would allow a faster application of our approach, which would be specially relevant for Task 2.

As it was previously referred, instead of using BioBERT, a model pre-trained directly over chemical patents text would possibly obtain higher performance in both tasks.

5 Conclusion

In Task 1 (NER), we obtained a F1-score of 0.9392 (exact matching) and 0.9630 (relaxed matching), which corresponds, respectively, to an increase of 0.0499 and 0.0577 over the baseline performance and to the 3rd best performing system. For Task 2 we were not able to produce results due to the lack of time to apply our approach over the entire test set. Consequently, our future work will focus mainly on the resolution of the problems associated with the poor performance in this task. First, we will explore other RE systems, like for example "BO-LSTM" [4]. Second, we will apply a module like Yoshikawa et al. [10] to extract the specific snippets describing chemical reactions within the patents text. The machine learning model (BiLSTM-CRF) proposed by the authors obtained a significantly higher performance in the extraction of chemical reactions comparing with simpler baseline approaches, including a rule-based model, which we expect will enhance our approach.

Acknowledgements

This project was supported by FCT through funding of the DeST: Deep SemanticTagger project, ref. PTDC/CCI-BIO/28685/2017, and the LASIGE ResearchUnit, ref. UIDB/00408/2020

References

1. Akhondi, S.A., Klenner, A.G., Tyrchan, C., Manchala, A.K., Boppana, K., Lowe, D., Zimmermann, M., Jagarlapudi, S.A., Sayle, R., Kors, J.A., Muresan, S.: Annotated chemical patent corpus: A gold standard for text mining. PLoS ONE **9**(9), 1–8 (2014). https://doi.org/10.1371/journal.pone.0107477
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (oct 2018), `http://arxiv.org/abs/1810.04805`
3. Hu, M., Cinciruk, D., Walsh, J.M.: Improving Automated Patent Claim Parsing: Dataset, System, and Experiments (2016), `http://arxiv.org/abs/1605.01744`
4. Lamurias, A., Sousa, D., Clarke, L.A., Couto, F.M.: BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies. BMC Bioinformatics **20**(10), 1–12 (2019). https://doi.org/10.1186/s12859-018-2584-5

5. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020). https://doi.org/10.1093/bioinformatics/btz682
6. Moon, T., Awasthy, P., Ni, J., Florian, R.: Towards Lingua Franca Named Entity Recognition with BERT. Tech. rep. (2019)
7. Nguyen, D.Q., Zhai, Z., Yoshikawa, H., Fang, B., Druckenbrodt, C., Thorne, C., Hoessel, R., Akhondi, S.A., Cohn, T., Baldwin, T., Verspoor, K.: ChEMU: Named entity recognition and event extraction of chemical reactions from patents. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12036 LNCS**, 572–579 (2020). https://doi.org/10.1007/978-3-030-45442-5_74
8. Papanikolaou, Y., Roberts, I., Pierleoni, A.: Deep Bidirectional Transformers for Relation Extraction without Supervision pp. 67–75 (2019). https://doi.org/10.18653/v1/d19-6108
9. Peng, Y., Yan, S., Lu, Z.: Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets (2019). https://doi.org/10.18653/v1/w19-5006, `https://www.mendeley.com/catalogue/2347f426-b409-3772-9174-688480ed2a76/`
10. Yoshikawa, H., Nguyen, D.Q., Zhai, Z., Druckenbrodt, C., Thorne, C., Akhondi, S.A., Baldwin, T., Verspoor, K.: Detecting Chemical Reactions in Patents. Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association (3), 100–110 (2019), `https://www.aclweb.org/anthology/U19-1014`