Chemical entity recognition and MeSH normalization in PubMed full-text literature using BioBERT

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, L. Alfonso Ureña-López and M. Teresa Martín-Valdivia
Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain

Abstract—**This paper presents the participation of the SINAI group in Track 2 - NLM-CHEM Full-text Chemical Identification and Indexing in PubMed articles. This challenge aims to identify chemicals located in texts and subsequently assign a unique identifier using the MeSH vocabulary. Detecting chemicals automatically in biomedical texts is a crucial task because current health information systems are not prepared to analyze and extract this knowledge due to the time and cost involved in manual processing. To address this challenge, our group presents different machine learning approaches such as deep learning and Transformer-based models such as BERT. Specifically, we compared the BiLSTM-CRF neural network with different language model inputs and the BERT architecture using the pre-trained BioBERT model. Moreover, for MeSH indexing we have used the Metamap tool. We achieve 87.98% F1 in the Named Entity Recognition task and 75.51% F1 in the Named Entity Normalization task, both in approximate matching.**

Keywords—Natural Language Processing;* Named Entity Recognition*;* Named Entity Normalization*; deep learning; BERT; Metamap; PubMed articles.

I. INTRODUCTION

Chemical and drug Named Entity Recognition (NER) is a fundamental step for further biomedical text mining and has received much attention recently. This task aims to automatically detect chemical and drug mentions in biomedical literature and is a great challenge to the scientific community for several reasons: there are different ways to refer to the same chemical or drug, abbreviations and acronyms are commonly used, symbols are often included in scientific publications and new chemicals and drugs are constantly and rapidly reported [1].

Drug and chemical name recognition seeks to recognize these types of mentions in unstructured medical texts and classify them into predefined categories. This is a fundamental task of medical information extraction and medical relation extraction systems [2] and is the key to linking entities with vocabularies and terminologies available in the biomedical domain such as MeSH.

The clinical Natural Language Processing (NLP) community organized a series of open challenges with the focus on identifying chemical and drug entities from narrative clinical notes, including the chemical compound and drug name recognition tasks such as CHEMDNER [3] and PharmaCoNER [4], and the extraction of drug-drug interactions from biomedical texts task such as DDIExtraction [5]. These workshops are very useful because the participants use innovative and updated systems, offering a state-of-the-art approach to the tasks.

Contributing to the participation in challenges, this paper describes the system presented by the SINAI team for the Track 2 - NLM-CHEM at BioCreative VII workshop [6]. This challenge focuses on extracting information from full-text articles in PubMed. Specifically, the challenge included two tasks: Chemical Identification and Chemical Indexing. The Chemical Identification task evaluation consisted of two subtasks: Chemical Mention Recognition and Chemical Normalization. On the one hand, the Chemical Mention Recognition subtask can be considered as a NER task, on the other hand, the Chemical Normalization subtask aims to normalize and assign to each entity a unique MeSH identifier.

We have been involved in the Chemical Identification task. In the first subtask (NER) our proposal aims at deep learning models and pre-trained models based on the Transformer architecture using BERT (Bidirectional Encoder Representations from Transformers). More specifically, we employ the BioBERT model trained on a large corpus from the biomedical domain. For the second task, we use the indexing tool MetaMap.

II. DATASET

NLM-CHEM corpus [7] consists of 150 full-text articles with chemical entity annotations from human experts for ~5000 unique chemical names, mapped to ~2000 MeSH identifiers.

The corpus has been distributed in two different formats: JSON and XML, so the names of the chemicals are annotated in the corpus in a structured way as shown in the following example:

```
<annotation id="0">
   <infon key="type">Chemical</infon>
   <infon key="identifier">MESH:D012721</infon>
   <location offset="43" length="8"/>
   <text>Carbaryl</text>
</annotation>
```

According to the example shown, the entities are labeled with the *Chemical* tag, have a unique MeSH identifier, contain the starting position in the text (offset) and the length of the entity, and finally have the entity mentioned.

Moreover, the corpus offered by the organizers to train and fine-tune our models is divided into three sets: training, development, and test. Subsequently, the official and unlabelled test set have been deliberated for evaluation.

On the one hand, the training set is composed of 80 documents whose most frequent MeSH code is D006859 with 438 occurrences. In the training set, the entity with the code D006859 contains different descriptions such as *hydrogen*, *H*, *1H*, *4H*, among others. With this simple example, we can highlight how challenging the task is given the number of descriptions that a single MeSH identifier can have.

On the other hand, 20 are the documents that make up the development set. The most frequent MeSH code in this set is D009569 (376 occurrences) with descriptions such as *nitric oxide* and *NO*.

Finally, another annotated corpus called test set has been offered by the organizers of the challenge. This set contains 50 documents with 328 occurrences of the MeSH identifier D017239 with descriptions such as *paclitaxel* and *taxol*.

III. CHEMICAL ENTITY RECOGNITION

A. Deep Neural Network

To address the chemical detection task (first subtask), we focus on recognizing and extracting specific types of chemical entities in the text. Specifically, we follow a methodology proposed by Huang et al. [8] implementing a BiLSTM-CRF model for the NER task.

Recurrent Neural Networks (RNNs) are powerful deep learning models for application in NLP. These models usually use a vector representation for each token by reading token by token and "remembering" important information. In other words, they are loop networks that allow for the persistence of information and are capable of handling sequential data such as text sequences [9].

As input to the RNN employed, we have used two contextual word embeddings: ELMo and BioBERT. Nowadays, contextual word embeddings such as Embeddings from Language Models (ELMo) and BERT have emerged. These techniques generate embeddings for a word according to the context in which the word appears. ELMo is derived from a bidirectional LSTM that is trained with a coupled language model objective on a large text corpus [10], in this way, ELMo looks at the entire sentence before assigning a vector to each word. Otherwise, BERT representations are jointly conditioned on both the left and right context and use the Transformer [11], a neural network architecture based on a self-attention mechanism. In our methodology, we used the pre-trained Transformer-based model called BioBERT [12] because it is the first domain-specific BERT-based model pre-trained on a large-scale biomedical corpus.

B. Transformers-based approach

Although RNNs have obtained high results and a wide range of related literature on the NER task in recent years, the pre-training of Transformer-based language models such as BERT [13] has also led to impressive gains in NER systems.

Transformer architectures are designed to handle sequential data such as natural language for tasks like NER and text classification. This architecture is based on an attention mechanism that allows us to focus on certain words either on the left or on the right to deal with the current word according to the NLP task we are dealing with.

Some pre-trained models based on BERT are even specific to the biomedical domain such as BioBERT and ClinicalBERT. In this study, we used the BioBERT (*biobert-base-cased*) included in the Hugging Face repository.

IV. CHEMICAL ENTITY NORMALIZATION

The second step after carrying out the NER task is to assign a unique identifier to each recognized entity, also known as Named Entity Normalization (NEN) or entity linking. For this purpose, MeSH controlled vocabulary is used. Mesh provides uniformity and consistency in the indexing and cataloging of biomedical literature.

Moreover, MeSH is currently used by the National Library of Medicine (NLM) indexes to describe the subject content of journal articles for MEDLINE.

In order to normalize the entities with the appropriate identifier, we used Metamap [14]. Metamap is a highly configurable application developed by NLM to map biomedical text to the UMLS Metathesaurus or, equivalently, to identify Metathesaurus concepts referred to in an English text.

Metamap offers a variety of important modules to carry out the normalization, two of which stand out: part-of-speech tagging and Word-Sense Disambiguation (WSD) module.

Since Metamap returns the UMLS unique identifier (CUI - Concept Unique Identifier), it is necessary to perform the conversion between UMLS and MeSH. For this purpose, the original UMLS Metathesaurus files provide information related to CUI such as synonyms, relations, or unique identifiers from other vocabularies such as MeSH. Thus, we can link from a CUI to a MeSH identifier.

Fig. 1 illustrates the workflow carried out for the NEN task in Track 2 of Biocreative VII.

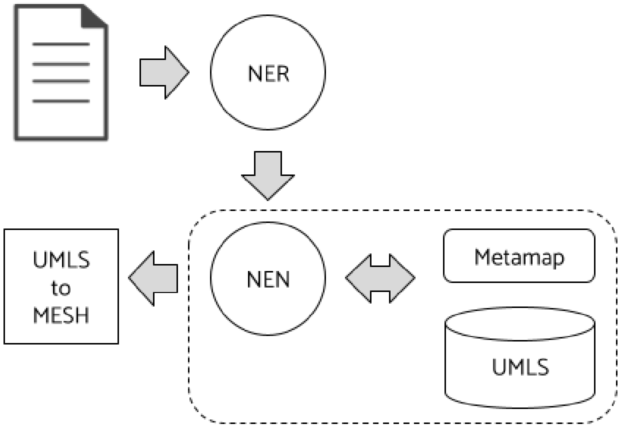

Fig. 1. Workflow conducted for the NEN task.

V. RESULTS

The metrics defined by the BioCreative VII Track 2 challenge to evaluate the submitted experiments are those commonly used for some NLP tasks such as NER or text classification, namely precision (P), recall (R), and F1-score (F1) considering strict and approximate match.

Regarding the results achieved, Table 1 shows the results obtained by the SINAI team for each run submitted in the chemical mention recognition task (NER subtask). As we can see, we obtain the worst results using the RNN approach with the embeddings extracted from ELMo. Specifically, we obtain 66.9% of F1 in a strict matching and 77.29% of F1 in an approximate matching. Taking into account the embeddings extracted from BioBERT to introduce them into the RNN, we improved the results a little bit, reaching 72.59% and 84.4% F1 for the strict and approximate evaluation respectively. These results make sense as BioBERT has been previously trained on biomedical domain texts. Finally, we obtain our best results using the Transformer-based approach with BioBERT, specifically 81.36% and 87.98% F1, reaching 90% precision in the approximate evaluation.

In the shaded part of Table I, we have also included the results achieved by other challenge participants as well as a reference proposed by the organizers [6]. Firstly, taking into account our best result and the median obtained by all participants, we did not manage to pass this assessment. Secondly, we are close to the benchmark proposed by the organizers, namely, we exceeded the proposed recall.

TABLE I. Results of Chemical Mention Recognition

Model	Strict			Approximate		
	P	R	F1	P	R	F1
RNN - ELMo (*)	75.41	60.11	66.90	86.82	69.64	77.29
RNN - BioBERT (*)	76.76	68.86	72.59	88.81	80.41	84.40
BERT-BioBERT	**83.12**	**79.67**	**81.36**	**90.09**	**85.96**	**87.98**
Median	84.76	81.36	83.73	92.20	86.82	89.51
Benchmark	84.40	78.77	81.49	91.56	84.92	88.11

(*) - unofficial

On the other hand, Table II shows the results according to the second subtask (NEN task). For this, we use Metamap and assign a unique MeSH identifier to each previously recognized entity of subtask one. In this case, we also achieved the lowest results using RNNs approaches. However, with the BERT approach, we obtain 77.63% F1 in strict matching and 75.51% F1 in approximate matching.

In this scenario, we outperformed the median of all task participants, specifically in precision and F-score metrics. Compared to the benchmark offered, we improved in recall and F1 of the strict evaluation and in recall of the approximate evaluation.

TABLE II. Results of Chemical Normalization to MeSH IDs

Model	Strict			Approximate		
	P	R	F1	P	R	F1
RNN - ELMo (*)	76.34	70.50	73.30	70.12	73.23	70.94
RNN - BioBERT (*)	78.36	72.43	75.27	70.83	74.99	72.27
BERT-BioBERT	**78.86**	**76.44**	**77.63**	**73.09**	**79.17**	**75.53**
Median	71.20	77.60	77.49	67.82	84.02	75.52
Benchmark	81.51	76.44	78.89	79.17	78.89	78.57

(*) - unofficial

VI. CONCLUSION

The SINAI group presents its participation in the NLM-CHEM challenge at BioCreative VII. We have participated in the first task of the challenge called Chemical Identification. In addition, this task was composed of two subtasks that we have designated as NER task and NEN task. The NER task aims to find chemical mentions in the full-text PubMed articles and assign a specific label and in the NEN task, the main objective was to index each found entity with a MeSH ID.

For the identification task, we have developed two systems with different deep learning approaches using a BiLSTM-CRF with two different embeddings: ELMo which is trained on a domain-general corpus, and BioBERT which is a pre-trained model based on Transformer. Moreover, we use the BERT architecture and the BioBERT pre-training model to evaluate its effectiveness. In order to assign a unique identifier to each detected concept, we have used the Metamap indexing tool. Metamap provides UMLS identifiers for each detected entity, so we subsequently link UMLS codes to MeSH identifiers.

The results obtained are not as expected; in the NER subtask, we did not exceed the mean of the participants and obtained 87.98% F1 in the approximate evaluation using BioBERT. In the NEN subtask, we improved the mean of the participants but we did not exceed the reference offered by the organizers.

For future work, we plan to perform in-depth error analysis on our approximations in order to see the weaknesses of each system. Moreover, we will study the performance of using linguistic features such as Part-Of-Speech tags as an input in the BERT model, as well as the use of the description of each concept in the MeSH vocabulary.

Acknowledgment

This work has been partially supported by the LIVING-LANG project [RTI2018-094653-B-C21] of the Spanish Government and the Fondo Europeo de Desarrollo Regional (FEDER).

References

1. Liu, S., Tang, B., Chen, Q., & Wang, X. (2015). Drug name recognition: approaches and resources. Information, 6(4), 790-810.
2. Warrer, P., Hansen, E. H., Juhl‐Jensen, L., & Aagaard, L. (2012). Using text‐mining techniques in electronic patient records to identify ADRs from medicine use. British journal of clinical pharmacology, 73(5), 674-684.
3. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2015). CHEMDNER: The drugs and chemical names extraction challenge. Journal of cheminformatics, 7(1), 1-11.
4. Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., & Krallinger, M. (2019, November). Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In Proceedings of The 5th Workshop on BioNLP Open Shared Tasks (pp. 1-10).
5. Segura Bedmar, I., Martínez, P., & Herrero Zazo, M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.
6. Robert Leaman, Rezarta Islamaj and Zhiyong Lu. (2021). Overview of the NLM-Chem BioCreative VII track: Full-text Chemical Identification and Indexing in PubMed articles. Proceedings of the seventh BioCreative challenge evaluation workshop.
7. Rezarta Islamaj, Robert Leaman, David Cissel, Meng Cheng, Cathleen Coss, Joseph Denicola, Carol Fisher, Rob Guzman, Preeti Kochar, Nicholas Miliaras, Zoe Punske, Keiko Sekiya, Dorothy Trinh, Deborah Whitman, Susan Schmidt and Zhiyong Lu. (2021). The chemical corpus of the NLM-Chem BioCreative VII track: Full-text Chemical Identification and Indexing in PubMed articles. Proceedings of the seventh BioCreative challenge evaluation workshop.
8. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.
9. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
10. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
14. Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association.