Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 203 (2022) 721–726

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Mobile4Medicine: Mobile Systems and Pervasive Computing for Personalized Medicine (M4Medicine)
August 9-11, 2022, Niagara Falls, Ontario, Canada

Named Entity Recognition and Knowledge Extraction from Pharmaceutical Texts using Transfer Learning

Nasi Jofche[a], Kostadin Mishev[a], Riste Stojanov[a], Milos Jovanovik[a], Eftim Zdravevski[a], Dimitar Trajanov[a,*]

[a] *Faculty of Computer Science and Engineering,*
Ss. Cyril and Methodius University in Skopje, 1000 Skopje, N. Macedonia

Abstract

The challenge of recognizing named entities in a given text has been a very dynamic field in recent years. This task is generally focused on tagging common entities, such as *Person*, *Organization*, *Date*, etc. However, many domain-specific use-cases exist which require tagging custom entities that are not part of the pre-trained models. This can be solved by fine-tuning the pre-trained models or training custom models. The main challenge lies in obtaining reliable labeled training and test datasets, and manual labeling would be a highly tedious task.

This paper presents a text analysis platform focused on the pharmaceutical domain. We perform text classification using state-of-the-art transfer learning models based on spaCy, AllenNLP, BERT, and BioBERT. We developed methodology that is used to create accurately labeled training and test datasets used for custom entity labeling model fine-tuning. Finally, this methodology is applied in the process of detecting *Pharmaceutical Organizations* and *Drugs* in texts from the pharmaceutical domain. The obtained F1 scores are 96.14% for the entities occuring in the training set, and 95.14% for the unseen entities, which is noteworthy compared to other state-of-the-art methods. The proposed approach implemented in the platform could be applied in mobile and pervasive systems since it can provide more relevant and understandable information to patients by allowing them to scan the medication guides of their drugs. Furthermore, the proposed methodology has a potential application in verifying whether another drug from another vendor is compatible with the patient's prescription medicine. Such approaches are the future of patient empowerment.

Keywords: Knowledge extraction; Natural language processing; Named entity recognition; Drugs; Drug data; spaCy; AllenNLP; BERT; BioBERT

[*] Corresponding author.
E-mail address: dimitar.trajanov@finki.ukim.mk

10.1016/j.procs.2022.07.107

1. Introduction

Considering the vast data volumes generated by variety of sources, including online social media platforms and news portals, individuals have hard time processing it and understanding it. This paper is concerned with using natural language processing (NLP) algorithms to perform intelligent knowledge extraction (KE) to process text from the pharmaceutical domain. Particularly, we aim to perform named entity recognition (NER) of *Pharmaceutical Organizations* and *Drugs*. NER takes a central place in many NLP systems, as a baseline task for information extraction, question answering, cyberbullying detection [1], text sumarization, topic modelling [2], etc. Our interest in this topic stems from a challenge in the LinkedDrugs dataset [3], where the collected drug products can have active ingredients (*Drug* entities) and manufacturers (*Pharmaceutical Organization* entities) written in a variety of ways, depending on the source, country of registration, language, etc. Our previous work shows promising results [4], and this paper builds upon it. As shown in [4], the ambiguity in entity naming in this dataset makes the data analysis process imprecise. Thus, using NER to normalize the name values for the active ingredients and manufacturers can significantly improve the quality of the dataset and the results from other downstream analytical tasks.

In this paper, we propose a methodology that can be used to automatically create labeled datasets for custom entity types showcased in texts from the pharmaceutical domain. In our case, this methodology is applied by tagging *Pharmaceutical Organizations* in pharmacy-related news. We prove that it can be extended to tagging other custom entities in different texts in the pharmaceutical domain by tagging *Drug* entities as well and assessing the obtained results. The main focus of this work is the automatic application of common language processing tasks, such as tokenization, dealing with punctuation and stop words, lemmatization. Additionally, we investigate the possibility of applying custom, business case-related text processing functions, like joining consecutive tokens to tag a multi-token entity or performing text similarity computations. The overall applicability and accuracy of this methodology is assessed by using two well-known language processing libraries, spaCy [5] and AllenNLP [6], which come with a pre-trained model based on convolutional layers with residual connections and a pre-trained model based on Elmo embeddings [7], respectively. The custom trained models which can tag the custom entity *Pharmaceutical Organization* indicate high tagging accuracy when compared to the initial pre-trained models' accuracy while tagging the more generic *Organization* entity over the same testing dataset. In addition, a model trained on the same dataset by fine-tuning the state-of-the-art BERT is used for gaining a better insight into the results. Lastly, a fine-tuned BioBERT [8], a model based on BERT architecture and pre-trained on biomedical text corpora, is also used to assess the results better.

A thorough explanation of the methodology used to generate the labeled datasets is given in the following sections, followed by custom model training and accuracy assessment. The extracted entities can help filter the documents and news that mention them, but this is not enough in the current era of data overflow. Therefore, we go one step further and integrate these results into a platform that then extracts and visualizes the knowledge related to these entities. This platform currently integrates state-of-the-art NLP models for co-reference resolution [7] and Semantic Role Labeling [9] to extract the context in which the entities of interest appear. This platform additionally offers convenient visualization of the obtained findings, which brings the relevant concepts closer to the people who use the platform.

2. PharmKE Knowledge Extraction Platform

This section describes our PharmKE platform [10, 11], which goes a step further in understanding pharmaceutical texts: on top of identifying *Drugs* and *Pharmaceutical Organizations*, it also extracts relations in the mentioned context and constructs a Knowledge Graph from them. The platform covers the entire process of understanding a document and its content – from its classification and filtering, i.e., does it belong to the pharmaceutical domain, all the way to visualization of the entities and their semantic relations.

Initially, the platform classifies whether a given text is from the pharmaceutical domain, and only the positively classified texts are accepted for further analysis. The classification model used in this step is a transferred BERT model, fine-tuned with a corpus of around 5,000 documents from the pharmaceutical domain as positive samples, and general news documents as negative samples (`https://www.kaggle.com/snapcrack/all-the-news`). 70% of these documents are used for fine-tuning of BERT's and XLNet's models, and their precision, recall and F1 measure are evaluated with the remaining 30% of the documents. The BERT model has a precision of 96.33%, recall of 95.28% and F1 of 95.8%. On the other hand, XLNet produced a precision of 99.83%, recall of 98.71%, and F1 of 99.26%.

Each correctly classified pharmaceutical text is further analyzed by recognizing combined entities through the proposed models, as well as by using BioBERT for the detection of BC5CDR (https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/) and BioNLP13CG (https://github.com/cambridgeltl/MTL-Bioinformatics-2016/tree/master/data) tags [12], which include Disease, Chemical, Cell, Organ, Organism, Gene, etc. Additionally, we use a fine-tuned BioBERT model in order to detect *Pharmaceutical Organizations* and *Drugs*, entity classes that are not covered by the standard NER tasks. We explain the fine-tuning process in more detail in Section 3. Tag collisions when combining the results from both models are avoided by applying precedence of the tags recognized by our fine-tuned model over the tags recognized by BioBERT's model (Simple Chemical). All of the recognized entities are visualized in the sentence, along with their respective tags.

The recognized entities serve as a baseline for finding all of their mentions in the entire text, by applying co-reference resolution in the background and replacing each mention ("it", "it's", "his", etc.) with their respective entity. Libraries such as AllenNLP, StanfordNLP [13] and NeuralCoref (https://github.com/huggingface/neuralcoref) provide implementations of the algorithms for co-reference resolution, focused on the CONLL2012 shared task [14]. Our platform utilizes the NeuralCoref library for co-reference resolution due to its high accuracy, ease of integration compared to StanfordNLP, and the capability to take into account user-specific information and the speakers in a conversation. Once the mentions in the text are replaced with their respective entities, the final task includes labeling the semantic roles in each sentence. This is performed by using the BERT-based algorithm for semantic role labeling [9]. Then, the concrete arguments, like subject and object, as well as modifier arguments like temporal, location, instrument, etc. are visualized in a sequential manner for quick understanding.

The result is a modular platform for pharmaceutical text analysis, which uses existing state-of-the-art models for entity recognition, as well as fine-tuned models for recognizing custom entities like *Pharmaceutical Organizations* and *Drugs*. The modular design of the platform enables a combination of results from multiple models which recognize a vast range of entities. It also allows for semantic role labeling and visualization for each entity and their respective mentions in the text, by using state-of-the-art algorithms implemented by popular libraries. The entire analysis can be exported in a JSON format, allowing it to be used for additional processing such as question answering, text summarization, fact extraction, etc.

As a final step, we annotate the entire text using the state-of-the-art knowledge extraction system DBpedia Spotlight [15]. The obtained results are then enriched with additional RDF facts which we construct from the identified *Pharmaceutical Organization* and *Drug* entities. This enriched knowledge graph is then available for further use within or outside the platform.

3. Entity Recognition for Pharmaceutical Organizations and Drugs

Our methodology starts with a text corpora from the pharmaceutical domain and a closed set of entities that belong to a given class. In our case, we are using entities that denote *Pharmaceutical Organizations* and *Drugs*. Using only these two prerequisites, we show that we can train models that can extract even unseen entities from the class of interest. First, we start with the text corpora from the pharmaceutical domain that potentially contains the entities from the class of interest. This text corpora consists of news collected from the following pharmacy-related websites: *FiercePharma* (www.fiercepharma.com), *Pharmacist* (www.pharmacist.com) and *Pharmaceutical Journal* (www.pharmaceutical-journal.com). Next, we tokenize the text such that we extract the words, and then we try to annotate each word with respect to the set of entities from the required type. We utilize cosine similarity and Levenshtein distance in particular [16], where we check if the word is similar to some of the entities. The annotation process assigns start and end positions for each token in the text, respectively. Once we are done with this phase, we have initialized a labeled dataset, denoted as *MD*.

One of the main challenges is that the Pharmaceutical Organization entity type can be found in a given text as multi-word phrases, such as **Sanofi Pharmaceuticals Ltd. Spain**, or as a single word: **Sanofi**. The name of the *Pharmaceutical Organization* can contain pharmacy-related keywords, such as **Pharmaceuticals**, **Pharma**, **Medical**, etc., which are not part of the core name of the organization, and can either be found along with it in the sentence or not at all. This means that we should not classify the countries, legal entities, and the pharmacy-related words as parts of the *Pharmaceutical Organization* type. Therefore, the annotation process sequentially performs use-case-specific token filtering during the creation of the *MD* dataset. This is done by using a non-entity list that contains all tokens

Table 1. Evaluation of the models on previously seen entities.

	PH_ORG		Organization	
Library	**Prec.**	**F1**	**Prec.**	**F1**
AllenNLP	95.57	90.3	49.41	48.26
spaCy	91.36	91.54	22.22	29.10
BERT	97.65	96.66	51.65	53.18
BioBERT	98.35*	96.86*	52.12*	53.38*

Table 2. Evaluation of the models on previously unseen entities.

	PH_ORG		Organization	
Library	**Prec.**	**F1**	**Prec.**	**F1**
AllenNLP	94.76	89.98	47.12	46.44
spaCy	90.95	88.51	21.98	28.01
BERT	97.45	97.68	51.51	55.68
BioBERT	97.52*	97.86*	52.42*	55.70*

that should be ignored. In our case, this list includes all countries in the world, together with the legal entity types for companies ("Ltd", "Inc", "GmbH", etc.) and pharmacy-related words. After filtering out the tokens from the non-entity list, only **Sanofi** will remain in our example, and we can be confident that the core name is thoroughly extracted. After matching the core name in the text, we use the same lists to detect neighbor tokens for multi-token names as part of the organization name using text similarity metrics. After the application of the custom, use-case-related filtering, the *MD* dataset consists of the core entities that have high text similarity. Only the entities with a similarity above the customized threshold are labeled as members of the target class. In our experiments, we use a similarity threshold of 0.9. Some *Pharmaceutical Organization* entities consist of multiple, consecutive tokens, such as *J & J*. We solve this by token concatenation of consecutive relevant tokens, using a custom function applied on the *MD*.

The *MD* dataset is then used to train a model which will be able to extract the named entities from the given class. Since NER models consider the context in which the entities appear in a sentence, the training dataset is not required to contain many diverse entities. Here we improve the general knowledge language model for the more specific task, using small or moderate amounts of labeled data. In our case, we fine-tune spaCy, AllenNLP, BERT and BioBERT models. However, each of these models requires a different data format. SpaCy requires an array of sentences with respective tagged entities for each sentence and their start- and end-positions. AllenNLP requires a dataset in BIOUL or BIO notations, which differentiate the following token annotations: multi-word entity beginning token *(B)*, multi-word entity inside tokens *(I)*, multi-word entity ending token: *(L)*, single-token entities: *(U)*, non-entity tokens: *(O)*. The dataset adapted for BERT and BioBERT labels the entities with *I - PH_ORG*, regardless of the number of tokens, while all other tokens are marked with *O*. The same methodology is used for creating labeled datasets for the *Drug* entity type. In this case, we use the same text corpora, but this time annotated with a somewhat larger set of *Drug* entities. Once we are done with the fine-tuning process, we have named entity recognition models able to extract the entities from a given type.

The accuracy of our proposed approach is assessed by using a pharmacy-related news dataset, which consists of around 5,000 news articles. The *Pharmaceutical Organization* entities set consists of 3,633 unique values, while the *Drug* entities set consists of 20,266 unique drug brand names. These sets were extracted and published as part of our previous work [3, 4]. The evaluation is performed in two distinct scenarios for both entity classes. In the first evaluation, we split the news dataset into training and test portions, with sizes of 70% and 30%, respectively, without considering the distribution of the inside entities. This scenario aims to check the overall precision of the fine-tuned model. In the second evaluation scenario, we evaluate the generalization ability of our approach. Here, we split the training and test portions based on the entities they contain, such that there will not be any entity overlap between them. To do so, we extract the documents that contain 30% of the entities as the testing portion, and the other news are used for training. However, the testing portion had more than 30% of the overall news. Therefore, to achieve a 70% - 30% ratio between the training and test portions, the test portion was reduced to contain exactly 30% of the news, while in the rest of the documents, the entities were replaced with other entities which do not belong to the entity set used in the testing portion.

The obtained fine-tuned models for detecting *Pharmaceutical Organization* entities using spaCy, AllenNLP, BERT and BioBERT were tested accordingly. The results were compared to the original models before their fine-tuning, where the task was the extraction *Organization* entities. The results are given in Table 1, indicating that the fine-tuned models can achieve a significantly higher F1 score compared to the original models. Also, we can outline that AllenNLP outperforms spaCy in this NER task. This result can be attributed to the different neural architectures used by both libraries, while the BERT model can outperform both. However, the pre-trained BioBERT on biomedical text can slightly outperform BERT in every evaluation.

Even though the pre-trained models consider the sentence context in which the entities appear, we can evaluate the fine-tuned model generalization capability by creating a test dataset that contains only entities that were not seen during the training. To achieve this, we use the joint dataset of the pharmacy-related news and generate a sample of entities in a random way to achieve a 70% - 30% split ratio between training and test datasets, where the test dataset contains entities not encountered in the training dataset. SpaCy, AllenNLP, BERT, and BioBERT models were also trained using these datasets, and the results are given in Table 2.

Table 3. Evaluation of the models on known entities.

Library	Precision	F1
AllenNLP	96.24	95.12
spaCy	90.95	94.87
BERT	98.86	95.98
BioBERT	98.92*	96.14*

Table 4. Evaluation of the models on previously unseen entities.

Library	Precision	F1
AllenNLP	92.65	89.85
spaCy	88.16	89.25
BERT	98.12	95.01
BioBERT	98.65*	95.14*

SpaCy, AllenNLP, BERT, and BioBERT models were also created for recognizing Drug entities in texts. The evaluation results are given in Table 3 for the scenario where the same *Drug* entity can be present in both the training and the test dataset. Table 4 shows the results when the test dataset does not contain any of the entities used in the training phase. Again, the train-test dataset ratio is 70% - 30%.

4. Discussion

The platform presented in this paper emphasizes a methodology for combining the best-performing NLP models and adapting them for use in a new domain. Furthermore, we use a modular approach, where each model is a separate phase in the knowledge extraction pipeline and allows for an easy upgrade with new and potentially superior models, therefore improving the performance of the entire platform. In contrast to [6, 5], the goal of our platform is to provide a knowledge extraction solution for the pharmaceutical domain that brings the state-of-the-art NLP achievements closer to the people which analyze large amounts of texts. The PharmKE platform is human-centric, meaning that it is designed to be used primarily by people who need to extract knowledge. The outcome from each phase is visualized, enabling the users to understand better the process of capturing and linking this knowledge. Since the web browser may not be the most convenient tool for domain experts to use in knowledge extraction, especially when they analyze texts from various sources, we are also publishing an Application Programming Interface (API) that exposes the results from our platform to other applications. With this, we enable the development of editor plugins that will potentially extract and visualize the knowledge in the tools that experts already use daily.

In the current version of the PharmKE platform, we fine-tuned the Named Entity Recognition module to extract two additional entity types, namely *Pharmaceutical Organization* and *Drug*, on top of the entity types already recognized by the superior BioBERT model. During the fine-tuning phase, we show a method for automatically creating the training set for the recognition of *Pharmaceutical Organizations* and *Drugs*, by using a text corpus from the pharmaceutical domain and a closed set of entity instances from the types of interest. The evaluation of the fine-tuned model showed that this methodology enables the recognition of entities that are not seen in the training set, which is a promising result. The PharmKE platform is open to continuous advancements in the NLP field. One of the crucial elements in the knowledge extraction process that the current models do not solve is the linking of the relations obtained by the SRL model with the corresponding properties in the knowledge graph. This is the challenge that our team will try to address in our future research and incorporate any model that will have better results in some of the current tasks. All of this is possible thanks to the modular design of the platform.

5. Conclusion

In this paper, we present a modular platform [10, 11] that incorporates state-of-the-art models for text categorization, pharmaceutical domain named entity recognition (NER), co-reference resolution (CRR), semantic role labeling (SRL) and knowledge extraction (KE). This platform is designed primarily for human users. PharmKE visualizes the results from each incorporated model, enabling pharmaceutical domain experts to better recognize the extracted knowledge from the input texts. Our strategic goal is to keep the PharmKE platform current and up-to-date, and its modular design enables easy incorporation of new and potentially superior models. One such step in this direction was our extension of the more recent BioBERT model for NER with the *Pharmaceutical Organization* and *Drug* entity type recognition. The platform is also publicly available [10] and is open-source [11], providing reproducibility of our results. This also means that other researchers can modify their copy of the platform, run their own instances of it, and even re-purpose it, thanks to its modular design. While training custom models for language understanding tasks in text, a common issue is the lack of labeled datasets for testing and training. To tackle this issue, we propose a methodology that can be used to automate the labeled dataset creation process for training models for custom entity tagging. The methodology was assessed by training custom models for named entity recognition using spaCy, AllenNLP, BERT, and BioBERT. The obtained results indicate that the newly trained models outperform the pre-trained models in detecting custom entities.

The proposed approach could be applied in mobile and pervasive systems because it can provide more relevant and understandable information to patients by allowing them to scan the medication guides of their drugs. Furthermore, the proposed methodology has a potential application in verifying whether another drug from another vendor is compatible with the patient's prescription medicine. Such approaches are the future of patient empowerment.

Acknowledgement

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss.Cyril and Methodius University, Skopje, Macedonia. This article is based upon work from COST Action CA 19121 (Good Brother: Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living) supported by COST (European Cooperation in Science and Technology). More information in https://www.cost.eu.

References

[1] F. Markoski, E. Zdravevski, N. Ljubešić, S. Gievska, Evaluation of recurrent neural network architectures for abusive language detection in cyberbullying contexts, Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science . . . , 2020.

[2] F. Markoski, E. Markoska, N. Ljubešić, E. Zdravevski, L. Kocarev, Cultural topic modelling over novel wikipedia corpora for south-slavic languages, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021, pp. 910–917.

[3] M. Jovanovik, D. Trajanov, Consolidating Drug Data on a Global Scale Using Linked Data, Journal of Biomedical Semantics 8 (1) (2017) 3.

[4] N. Jofche, M. Jovanovik, D. Trajanov, Named Entity Discovery for the Drug Domain, in: 16th International Conference on Informatics and Information Technologies, Faculty of Computer Science and Engineering, Skopje, 2019, pp. 1–10.

[5] M. Honnibal, I. Montani, spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing, To appear 7 (2017).

[6] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, L. S. Zettlemoyer, AllenNLP: A Deep Semantic Natural Language Processing Platform, 2017. arXiv:arXiv:1803.07640.

[7] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep Contextualized Word Representations, arXiv preprint arXiv:1802.05365 (2018).

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining, arXiv preprint arXiv:1901.08746 (2019).

[9] P. Shi, J. Lin, Simple BERT Models for Relation Extraction and Semantic Role Labeling, arXiv preprint arXiv:1904.05255 (2019).

[10] PharmKE Platform: Public instance, http://pharmke.env4health.finki.ukim.mk, accessed: 2022-04-15 (2022).

[11] PharmKE Platform: Source code, https://gitlab.com/jofce.nasi/pharma-text-analytics, accessed: 2022-04-15 (2022).

[12] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, J. Han, Cross-type Biomedical Named Entity Recognition with Deep Multi-task Learning, Bioinformatics 35 (10) (2018) 1745–1752.

[13] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.

[14] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Y. Zhang, CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, in: Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012, pp. 1–40.

[15] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia Spotlight: Shedding Light on the Web of Documents, in: Proceedings of the 7th international conference on semantic systems, ACM, 2011, pp. 1–8.

[16] W. H. Gomaa, A. A. Fahmy, A Survey of Text Similarity Approaches, International Journal of Computer Applications 68 (13) (2013) 13–18.