Extracting Drug-Drug Interactions with Character-Level and Dependency-Based Embeddings

Liliya Akhtyamova[1] and John Cardiff[2]

Abstract— **The DDI track of TAC-2018 challenge addresses the problem of an information retrieval of drug-drug interactions on structured product labeling documents with discontinuous and overlapping entities. In this paper, we present our participation for event extraction subtask (Task 1). We used a supervised long-short-term memory (LSTM) network with conditional random fields decoding (LSTM-CRF) approach with an automatic exploring of words and characters features. Additional dependency-based information was integrated into word embeddings to allow better word representation. Our system performed with above median score.**

I. INTRODUCTION

The tasks of developing automatic tools to extract mentions of adverse drug events or drug-drug interactions were aroused from the problem of a big proportion of deaths happening due to these events, which accounting one eighth only in the USA [1].

There are many sources, from where these data could be acquired and analyzed: social media, blog posts, medical articles, electronic health records, questionnaires, structured product labelings (SPLs), etc. While social media data are usually noisy, with lots of misspellings and informal, data from official sources lack this problem. However, the goal for research of the former is usually to get information and alarms on trends and events, the latter besides these could be used to populate and form knowledge bases (KBs).

As the first step, entities and events in a text should be recognized and extracted. This is usually addressed as named entity recognition task. There are 2 types of methods to solve this kind of tasks: rule or pattern-based and machine learning methods. The most popular rule-based clinical NLP systems include MetaMap[1], MedLEE[2], and KnowledgeMap[3]. However, these systems do not require labeled data and simple to use, they would not perform poor on unseen data or data with the different structure. In this way, machine learning methods provide the ability for the system to recognize more sophisticated patterns,

Among conventional machine learning methods the most popular and well performing are Conditional Random Fields (CRFs) [2]. Typically to perform on par with more advanced methods these systems require heavy feature engineering from different linguistic levels: orthographic (e.g. length of word, capitalized letters), syntactic (e.g. POS tags), semantic (e.g. dependency graph information, UMLS concept unique identifier) information, word n-grams and word clusters, gazetteers, etc. However, being quite transparent these systems are mostly task-specific and lack of robust feature representations (e.g. using bag-of-words feature representations these synonyms "high blood pressure" and "hypertension" would have nothing in common). However, recently

To overcome this issue, the concept of distributed word representation or word embedding was introduced. The most popular system representing this concept called word2vec[4]. This highly increased the generability of machine learning methods. For example, in paper [], by introducing a joint inference framework to a CRF model were able to get advanced results on ACE2005 corpus[5].

Partly due to introducing word embeddings, other tools of machine learning, called deep learning, gained its popularity. To the moment, the most used neural architectures include Convolutional Neural Networks (CNNs) [3] and Recurrent Neural Networks (RNNs) [4], [5], [6].

For the task of NER, the state-of-the-art systems usually combine different architectures like variant of RNNs - LSTM with the CRF model showed in numerous tasks to be effective [7], [8], [9], [10], [11]. Other deep learning methods achieving state-of-the-art results include Gated Recurrent Unit Networks with top performance on BioNLP'16 Shared Task [12]. The goal of this task was extracting biomedical events between biotope and bacteria from biomedical literature. To overcome the limited human annotated data in the biomedical domain, different techniques are used such as distant supervision to generate additional samples to train NER [13], [14] and transfer learning methods [15].

In this work, we present LSTM-CRF architecture to extract mentions of substances, drugs and triggers in shared task on DDI extraction of TAC track.

The outline of the paper is as follows. Section 2 describes the used architecture as well as pre- and postprocessing. Section 3 discusses the experiments and inspects the learnt feature detectors. Section 4 concludes the paper.

II. METHODS

In this section, we describe representation schema used in experiments following by methods and word embedding construction and network description. We conclude with system postprocessing description.

[1],[2] Institute of Technology Tallaght, Dublin, Ireland
Corresponding author:
liliya.akhtyamova@postgrad.ittdublin.ie

[1]https://metamap.nlm.nih.gov/

[2]http://www.medlingmap.org/taxonomy/term/80

[3]https://www.mindjet.com/features/knowledge-map/

[4]https://radimrehurek.com/gensim/models/word2vec.html

[5]http://www.itl.nist.gov/iad/mig/tests/ace/2005/

A. Representations

For testing hypothesis, we employed 2 commonly used tagging schemes. The first one is based on beginning-inside-outside (BIO) mapping. The second schema is an extension of the first one with adding the end of an entity and single entity mapping (BIOES). In our experiments, we test both schemes.

B. LSTM-CRF

For approaching Task 1 of DDI extraction track, we used the state-of-the-art bidirectional LSTM-CRF system with word and characters embeddings [8].

C. Embeddings and Network Training

Word Embeddings. For our experiments we used pre-trained word embeddings presented in [16]. It was shown for a set of sequence tagging tasks these embeddings are most efficient [17]. Representing each word as a node in the dependency graph, word embeddings are optimized to learn in a way that they maximize the probability of other words within distance one and two. They are trained with window size 300 using skip-gram variants on English Wikipedia August 2015 dump of 2 billion words.

Character-Based Embeddings. Character-level information, especially pre- and suffixes of words, can contain valuable information for information retrieval tasks, especially in a biomedical domain. For this task, we test different approaches. First approach derived from the paper [8] and uses convolutional neural networks (CNNs) [18] to generate task-specific character level representations. The second approach [4] is implemented by using bidirectional LSTMs. Both approaches learn representations while training. In our experiments, we used the default parameters mentioned in that papers. For more details on system realizations, refer to the respective papers.

Other Features. In addition to word and character-based embedding features, we incorporated POS tags and casing information. For text tokenization and POS-tagging we used SpaCy package[6].

Network Training. We implemented the BiLSTM network using Keras[7] with TensorFlow backend. The default model parameters from [17] were used.

D. Post-processing

To handle discontinuous and overlapping entities some manual rules were written. Moreover, some parts of entities could belong to different categories of interactions, i.e. "reduce effect" defines *Trigger* and "reduce antiviral effect" defines *SpecificInteraction*. To deal with that we learn two separate systems to predict [*Trigger, Precipitant*] and [*Specific Interactions, Precipitant*]. We then merge both prediction outputs.

[6]https://spacy.io/api/tokenizer

[7]http://keras.io

III. EXPERIMENTS

A. Corpus

For the DDI track of TAC challenge, the data was provided in XML format with all sentences grouped by drugs. The training data were semi-automatically annotated with following manual correction of extracted entities. Overall, 20 SPLs with 22 drug mentions were provided for training, 50 SPLs with over 50 drug labels – for testing. Additional training data was provided in a slightly different format with 180 SPLs in total.

Entities in training data were tagged into three categories:

1) *Precipitant*, which defines a substance interacting with the Labeled Drug, which could be another drug, a drug class or a non-drug substance (e.g., alcohol, grapefruit juice).
2) *Trigger*, which defines word or phrase for an interaction event (e.g., increase/decrease in blood level).
3) *SpecificInteraction*, which defines the results of interactions (e.g., severe hyperkalemia).

In additional training data provided by organizers of challenge there are three types of annotations representing *Trigger* or *Specific Interaction* in main training set – [*Pharmacokinetic, Pharmacodynamic, Unspecified*] with *Pharmacokinetic* and *Unspecified* interactions related to Trigger in main test set and *Pharmacokinetic* had been mix of *Triggers* and *SpecificInteractions* from main guideline. In our experiments, we made an assumption to be *Pharmacokinetic* assigned as *SpecificInteraction*.

Not all the sentences were tagged. Moreover, sentences that did not have interaction events were not tagged with the *Precipitant* tag at all. In our experiments, we consider all sentences with removing odd tags in the postprocessing step. It is counting for around 5000 sentences in a training file. We further split training data on stratified training and validation sets in the proportion of 80:20. Test data included 2 files with 8205 sentences in Test 1 file and 4256 sentences in Test 2 file.

B. Evaluation Metrics

The evaluation metrics include Precision (P), Recall (R) and micro-averaged F1-score (F-score). For more details, please refer to the task website[8] or corresponding paper.

C. Experimental Results and Discussion

In Table 1 the results of the experiment using the official script provided by organizers of competition are listed. Here, *type* denotes the use of partial or exact matching. The primary metric is based on exact matching.

As it could be seen, the system performance based on partial and exact matching is differing insignificantly which means that within found entities they are almost full match with gold standard entities.

However, low recall is related to the overall poor system ability to recognize health-related entities in texts. This is due to fact that we used a general system with default settings not

[8]https://bionlp.nlm.nih.gov/tac2018druginteractions/

Test Data		Precision	Recall	F-score (%)
Test1	+type	34.57	18.90	24.44*
	-type	34.75	19.00	24.57
Test2	+type	37.76	24.07	29.40*
	-type	37.85	24.13	29.47

TABLE I
THE PERFORMANCE SCORES ON TWO TEST SETS IN TASK 1 (* DENOTES PRIMARY METRIC).

tuned to solve health-related NER tasks with discontinuous and overlapping entities. Indeed, word embeddings learned on a dump of Wikipedia should be reconsidered to other ones, adjusted to health domain. Moreover, while dealing well with continuous entities the system performs poorly recognizing long or discontinuous entities. It was shown that gated RNNs could be advantageous over LSTM and CRF over some NER tasks which should be checked on this dataset [5]. Additionally, integrating knowledge base information in deep learning system architecture and attention mechanism could further facilitate performance [19].

Regarding system architecture, we tested the system performance based only on LSTM-based character word embeddings, but it is needed to check other variants, e.g. CNNs character embeddings and try to combine outputs based on them with a model trained on other word embeddings which from the literature has shown to be effective [20]. This could be done through ensembles or bagging techniques.

In the context of post-processing, in our experiments we did not deal with a case of discontinuous entities where two different entities have the same initial word, i.e. "effects antagonized" and "effects of adenosine are antagonized" where the first defines *Trigger* and the second – *SpecificInteractions*.

It should also be mentioned that the text annotation based on BIOES schema is quite vague, and not the best solution for dealing with overlapping and discontinuous entities. Another approach could be the use of multilabel training which showed to be superior to unique labeling schemes on a task of adverse events mentions extraction on social media [10].

IV. CONCLUSION

In this paper, we implement the system to extract entities and drug-drug interaction events on data from structured product labeling. The main bottleneck of this task became the recognition of discontinuous and overlapping entities and events.

In our experiments we use state-of-the-art architecture based on LSTM-CRF network with word and LSTM-based character word embeddings. Some rule-based methods are constructed for postprocessing step to deal with aforementioned problems of separated entities and events. The initial results are promising, depicting area for future improvements.

REFERENCES

[1] J. Goldstein, I. Jaradeh, P. Jhawar, and T. Stair, "ED Drug-Drug Interactions: Frequency & Type, Potential & Actual, Triage & Discharge," *The Internet Journal of Emergency and Intensive Care Medicine*, vol. 8, no. 2, 2004.

[2] C. Fócil-Arias, G. Sidorov, A. Gelbukh, and F. Arce, "Extracting medical events from clinical records using conditional random fields and parameter tuning for hidden Markov models," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5, pp. 2935–2947, 5 2018.

[3] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A Convolutional Neural Network for Modelling Sentences," Tech. Rep.

[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 260–270.

[5] A. N. Jagannatha and H. Yu, "Bidirectional RNN for Medical Event Detection in Electronic Health Records." *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2016, pp. 473–482, 6 2016.

[6] Y. Wu, M. Jiang, J. Xu, D. Zhi, and H. Xu, "Clinical Named Entity Recognition Using Deep Learning Models." *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2017, pp. 1812–1819, 2017.

[7] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," 8 2015.

[8] X. Ma and E. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 1064–1074.

[9] A. Magge, G. Gonzalez-Hernandez, F. Liu, A. Jagannatha, and H. Yu, "Medication and Adverse Drug Event Detection Workshop," Tech. Rep., 2018.

[10] B. Tang, J. Hu, X. Wang, and Q. Chen, "Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF," *Wireless Communications and Mobile Computing*, vol. 2018, pp. 1–8, 4 2018.

[11] S. Misawa, M. Taniguchi, Y. Miura, and T. Ohkuma, "Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition," Tech. Rep.

[12] L. Li, J. Wan, J. Zheng, and J. Wang, "Biomedical event extraction based on GRU integrating attention mechanism," *BMC Bioinformatics*, vol. 19, no. S9, p. 285, 8 2018.

[13] A. Magge, D. Weissenbacher, A. Sarker, M. Scotch, and G. Gonzalez-Hernandez, "Deep neural networks and distant supervision for geographic location mention extraction," *Bioinformatics*, vol. 34, no. 13, pp. i565–i573, 7 2018.

[14] " :: ."

[15] Z. Wang, Y. Qu, L. Chen, J. Shen, W. Zhang, S. Zhang, Y. Gao, G. Gu, K. Chen, and Y. Yu, "Label-aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition," Tech. Rep.

[16] A. Komninos and S. Manandhar, "Dependency Based Embeddings for Sentence Classification Tasks," in *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*, San Diego, California. Association for Computational Linguistics, 2016, pp. 1490–1500.

[17] N. Reimers and I. Gurevych, "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 2017, pp. 338–348.

[18] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 12 1989.

[19] E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, and V. Malykh, "Medical concept normalization in social media posts with recurrent neural networks," *Journal of Biomedical Informatics*, vol. 84, pp. 93–102, 8 2018.

[20] R. Kavuluru, A. Rios, and T. Tran, "Extracting Drug-Drug Interactions with Word and Character-Level Recurrent Neural Networks," *IEEE Int Conf Healthc Inform.*, pp. 5–12, 2017.