Deep Learning System for Biomedical Relation Extraction Combining External Sources of Knowledge*

Diana Sousa[0000−0003−0597−9273]

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
dfsousa@lasige.di.fc.ul.pt

Abstract. Successful biomedical relation extraction can provide evidence to researchers about possible unknown associations between entities, advancing our current knowledge about those entities and their inherent processes. Multiple relation extraction approaches have been proposed to identify relations between concepts in literature, namely using neural networks algorithms. However, the incorporation of semantics is still scarce. This project proposes that using external semantic sources of knowledge along with the latest state-of-the-art language representations can improve the current performance of biomedical relation extraction both in English and non-English languages. The goal is to build a relation extraction system using state-of-the-art language representations, such as BERT and ELMo, with semantics retrieved from external sources of knowledge, such as domain-specific ontologies, graph attention mechanisms, and semantic similarity measures.

Keywords: Biomedical Relation Extraction · Deep Learning · Semantics.

1 Motivation

The volume of unstructured textual information currently available widely surpasses the ability of analysis by a researcher, even if restring it to a domain-specific topic. Biomedical literature is the standard method that researchers use to share their findings mainly in the form of articles, patents and other types of written reports [8]. Thus, scientific articles are the primary source of knowledge for biomedical relations, including human phenotypes and other biomedical entities, such as genes and diseases. Processing the amount of information available is only feasible using text mining techniques.

Deep learning is widely used to solve problems such as speech recognition, visual object recognition, and object detection. However, deep learning methods

* This work was supported by FCT through funding of DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017 (http://dest.rd.ciencias.ulisboa.pt/), LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020, and PhD Scholarship, ref. SFRH/BD/145221/2019.

that effectively identify and extract relations between biomedical entities in the text are still scarce [12]. Lately, efforts regarding new pre-trained language representation models have been proposed with BERT [23,6], and applied to the biomedical domain with BioBERT [11], achieving promising results. These pre-trained models can act as information layers for a biomedical RE deep learning model that uses not only the training data but also external sources of knowledge like domain-specific ontologies combined with graph attention mechanisms or semantic similarity measures. External sources of knowledge, such as the Gene Ontology (GO) [2] and the Human Phenotype Ontology (HPO) [15], can provide highly valuable information for the detection of relations between entities in the text [10], each containing several thousands of terms and annotations.

To the best of our knowledge, there is no deep learning RE system that includes in their data representations the information encoded in ontologies combined with other types of semantics to identify and extract relations between biomedical entities in articles.

2 Background and Related Work

Using different sources of information to support automated extracting of relations between biomedical concepts contributes to the development of our understanding of biological systems [22]. Researchers have proposed several RE approaches to identify relations between concepts in biomedical literature, namely, using neural network algorithms. The use of multichannel architectures composed of multiple data representations, as in deep neural networks, leads to state-of-the-art results. The right combination of data representations can eventually lead us to even higher evaluation scores in RE tasks.

Semantic resources such as knowledge bases and graphs can contain highly structured background data, particularly for the biomedical domain [13]. These resources play a fundamental role in the way we store, organize and retrieve information. Biological knowledge bases are commonplace for researchers and clinicians to access all types of biomedical data retrieved from biomedical literature [1]. Researchers can explore these resources regarding information retrieval systems, so one can rely on more than the literature itself to train a RE model. By integrating semantic resources, we feed the training process with extra, highly relevant information about each entity in the relation and the connections that that entity establishes within the known semantic universe. Using heterogeneous graphs attention mechanisms to represent indirect relations between different type entities, such as genes and diseases in the biomedical domain, can be a viable additional external source of knowledge to preexisting deep learning RE systems [24]. Thus, enabling us to find representations of an indirect relation between two entities using knowledge graphs. The knowledge graphs to implement heterogeneous graphs attention mechanisms could be ontologies representing the entities of interest and their semantic relationships in a given domain. An ontology is a structured way of providing a common vocabulary in which shared knowledge is represented [7]. Word embeddings can learn how to detect rela-

tions between entities but manifest difficulties in grasping each entity's semantics and their specific domain. Domain-specific ontologies provide and formalize this knowledge. Biomedical ontologies are usually structured as a directed acyclic graph, where each node corresponds to an entity and the edges correspond to known relations between those entities. Thus, a structured representation of the semantics between entities and their relations, an ontology, allows us to use it as an added feature to a machine learning classifier.

3 Research Questions and Methodology

This doctoral proposal can be divided into three main research questions (RQ):

- **RQ1:** Can the latest advances in language representations be used to create a state-of-the-art RE deep learning system? (Sub-section 3.1)
- **RQ2:** Can we use biomedical semantics as an add-on for RE systems? (Sub-section 3.2)
- **RQ3:** How can we evaluate RE systems regarding the biomedical domain in English and non-English languages? (Sub-section 3.3)

The RE systems will go through ongoing evaluation as new information is added, using different benchmark datasets: the semantic relations between pairs of nominals corpus SemEval-2010 Task 8 [9], the drug-drug interactions corpus SemEval-2013 task-9 [17], and the Phenotype-Gene Relations corpus [19]. This project will use three distinct state-of-the-art evaluation metrics: recall, precision, and F-measure to compare the results obtained with different datasets and approaches.

3.1 Deep Learning System

Each set of biomedical entities has distinct textual characteristics, inherent to unique contexts. Each entity will be identified with a domain-specific Named-Entity Recognition (NER) system [25]. Regarding Named-Entity Linking (NEL), entities such as genes, chemicals, diseases, and proteins, will be matched to an identifier through the corresponding ontology. These tasks need to be optimized to perform RE.

The RE system between the linked identified entities is going to be built using bidirectional Long Short-Term Memory (LSTM) networks, a deep learning method that deals with long sentences of words, with a similar architecture to Reccurrent Neural Networks (RNN), based on the work of Lamurias et al. [10] (BO-LSTM system). These models use different types of information, known as channels, such as word embeddings, part-of-speech tags, grammatical relations, and WordNet hypernyms [4] to maximize performance. Each of these channels has different types of input information and is responsible for one of the model layers. All of these layers can be connected to a softmax layer outputting the probabilities of each class.

3.2 Semantics as an Add-on for RE Systems

Taking advantage of semantics can provide supplementary information that may not be present in the training data. Ontologies formalize existing knowledge about entities such as genes [2], and diseases [16]. By representing each entity as the sequence of its ancestors, it is possible to detect new relations between entities that were not evident by only using the training data. Also, a new word embedding layer is going to be built, taking advantage of semantics/attention mechanisms. Word embeddings usually represent a variable-length sentence into a fixed-length vector, where each element of the vector encodes some semantics of the original sentence. The innovation resides in adding the ontology semantics of the identified entity to each vector, as well a graph attention mechanism, and test the use of semantic similarity measures. This work will explore some avenues, such as creating an annotation vector, along with the pre-existing entity vector, that expresses ascendants, descendants, and their connections to be fed to the model, including cross-domain relationships already established as for different types of biomedical entities.

3.3 Evaluation Tactics of RE Systems to the Biomedical Domain

Apart from the standard evaluation tactic reported, some paths can facilitate the evaluation of different approaches, including the development of an improved automated corpus creation based on the PGR corpus for system assessment [19]. Improving automating corpus creation is of interest to create training data for the developed systems since some biomedical relations do not have gold standard corpus available to use to test the quality of these systems. Leveraging on previous work [19] it is possible to generate multiple silver standard corpus for different entities with good enough results. These results have been demonstrated to be sufficient for training deep learning-based systems [20], and constitute a solid contribution to the Information Retrieval (IR) field. Also, apply domain-specific ontologies of non-biomedical topics, for example, the Planteome, a plant ontology [5], using benchmark datasets. Finally, making use of the translation of some ontologies like the HPO, and the DECS ontology [3] (i.e., Health Sciences Descriptors in Portuguese and Spanish) linked to English mesh terms [14], will allow us to study the effect of different languages in the system.

This thesis's early contributions consist on four publications, including a book chapter about neural networks [22], a conference paper describing the integration of multiple ontologies into a deep learning system [18], a journal paper describing improving accessibility and distinction between negative results in biomedical RE using the PGR dataset [21], and a journal paper on an approach to create biomedical training corpora using distant supervision and crowdsourcing [20].

4 Research Issues for Discussion

I seek suggestions and comments on how to improve this proposal. I am specifically interested in discussing how to integrate external knowledge into a relation extraction system effectively in a seamless and generalizable way.

References

1. Arnaboldi, V., Raciti, D., Van Auken, K., Chan, J.N., Müller, H.M., Sternberg, P.W.: Text mining meets community curation: a newly designed curation platform to improve author experience and participation at wormbase. Database **2020** (2020)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)
3. Campanatti-Ostiz, H., Andrade, C.: Health sciences descriptors in the brazilian speech-language and hearing science. Pro-fono: revista de atualizacao cientifica **22**(4), 397 (2010)
4. Ciaramita, M., Altun, Y.: Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 594–602. Association for Computational Linguistics (2006)
5. Cooper, L., Meier, A., Laporte, M.A., Elser, J.L., Mungall, C., Sinn, B.T., Cavaliere, D., Carbon, S., Dunn, N.A., Smith, B., et al.: The planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic acids research **46**(D1), D1168–D1180 (2018)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
7. Gruber, T.R., et al.: A translation approach to portable ontology specifications. Knowledge acquisition **5**(2), 199–221 (1993)
8. Hearst, M.A.: Untangling text data mining. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 3–10. Association for Computational Linguistics (1999)
9. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Séaghdha, D.O., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38. Association for Computational Linguistics (2010)
10. 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**(1), 1–12 (2019)
11. 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)
12. Li, F., Zhang, M., Fu, G., Ji, D.: A neural joint model for entity and relation extraction from biomedical text. BMC bioinformatics **18**(1), 198 (2017)
13. Li, Z., Lian, Y., Ma, X., Zhang, X., Li, C.: Bio-semantic relation extraction with attention-based external knowledge reinforcement. BMC Bioinformatics **21**, 1–18 (2020)
14. Papagiannopoulou, E., Papanikolaou, Y., Dimitriadis, D., Lagopoulos, S., Tsoumakas, G., Laliotis, M., Markantonatos, N., Vlahavas, I.: Large-scale semantic indexing and question answering in biomedicine. In: Proceedings of the Fourth BioASQ workshop. pp. 50–54 (2016)

15. Robinson, P.N., Mundlos, S.: The human phenotype ontology. Clinical genetics **77**(6), 525–534 (2010)
16. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease ontology: a backbone for disease semantic integration. Nucleic acids research **40**(D1), D940–D946 (2012)
17. Segura-Bedmar, I., Martínez, P., Herrero-Zazo, M.: Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 341–350 (2013)
18. Sousa, D., Couto, F.M.: BiOnt: deep learning using multiple biomedical ontologies for relation extraction. In: European Conference on Information Retrieval. pp. 367–374. Springer (2020)
19. Sousa, D., Lamurias, A., Couto, F.M.: A silver standard corpus of human phenotype-gene relations. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1487–1492 (2019)
20. Sousa, D., Lamurias, A., Couto, F.M.: A hybrid approach toward biomedical relation extraction training corpora: combining distant supervision with crowdsourcing. Database **2020** (2020)
21. Sousa, D., Lamurias, A., Couto, F.M.: Improving accessibility and distinction between negative results in biomedical relation extraction. Genomics & Informatics **18**(2) (2020)
22. Sousa, D., Lamurias, A., Couto, F.M.: Using neural networks for relation extraction from biomedical literature. In: Cartwright, H. (ed.) Artificial Neural Networks, pp. 289–305. Springer US, New York, NY (2021)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
24. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems (2020)
25. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 2145–2158. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018)