External knowledge in distant supervision for n-ary relation extraction

Sofia I. R. Conceição[1,*]

[1]LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

Abstract

There is scarcity in the development of high quality and large annotated corpora containing n-ary biomedical relations extracted from literature, that allow the direct application of state-of-the-art relation extraction systems. Currently, the majority of the datasets and systems cover mostly relations with two entities and make poor use of external biomedical knowledge resources, despite the existence of knowledge bases such as ontologies and knowledge graphs. N-ary relation extraction in the biomedical field could provide a more reliable notion of the existing system such for example which driver genes cause a certain phenotype if mutated in a specific tissue, i.e. a ternary relation of gene-phenotype-tissue. The hypothesis is that the introduction of external knowledge would help to improve the existing state-of-the-art models. Additionally, distant supervision is one of the methods that can be used to surpass the lack of annotated corpus in the biomedical domain also using ontologies to inject external knowledge. This combination of approaches will help to unravel new organizational principles of biological systems and enhance the n-relation extraction systems in the biomedical field.

Keywords

biomedical relation extraction, distant supervision, n-ary relations, text mining

1. Introduction

The noteworthy pace of information generation in the biomedical field creates a demanding challenge in keeping our data and knowledge up to date. Relation Extraction (RE) is a central task in Natural Language Processing [1], but there is a scarcity of annotated corpora in the biomedical domain. A route to surpass this is to create a silver standard corpus which is annotated by an automatic method and partially validated [2]. Distant supervision is one of the methods to create this type of corpus, it assumes that if any two entities have a known relation in a knowledge base, these entities will express that relation when both are present in the same sentence [1, 3, 4]. This approach, although reducing the expense of the creation of a dataset, creates plenty of noise such as the two entities can be in the same sentence without expressing a relation; the knowledge base can be incomplete, thus not identifying a relation; or there can be various instances for the relation of the entities, although not all of them are identified.

Numerous studies tried to minimise the noise present on these corpora, such as treating

10th edition of the PhD Symposium on Future Directions in Information Access (FDIA) at ESSIR 2022 : The 13th European Summer School in Information Retrieval, July 18 – 22, 2022, Lisbon, Portugal

*Corresponding author.

siconceicao@fc.ul.pt (S. I. R. Conceição)

0000-0002-8891-3546 (S. I. R. Conceição)

distant supervision as a multi-instance problem [3, 4]. Multi-instance assigns a label to a bag of sentences with the same pair of entities instead of generating sentence-level labels. In binary classification if at least one sentence is true then the bag is labelled true, while contrarily is labelled negative. Recently Li et al.[4] developed a framework that obtains state-of-the-art (SOTA) values and uses an entity-aware embedding module and a self-attention enhanced selective gate mechanism to integrate task-specific entity information into word embedding and then generates a complementary context-enriched representation for PCNN. Although, there are some problems, such as each sentence in a bag can be regarded as an independent individual and do not have any relationship with other sentences and lacks external knowledge, that would help to improve the prediction quality.

Moreover, most of the corpus and relation extraction approaches only focus on binary relations. However, sometimes to fully characterise the relation other relevant entities may be necessary. Many studies have explored this theme, such as in [5, 6, 7].

For example, current studies have shown that cancer driver genes are mutated depending on the tissue, which is not explained by gene expression patterns across tissues [8]. So, it is necessary to extract ternary relations of gene, phenotype/ disease and tissue to have a valid relation.

The novelty of this project is the identification of the context (in this case, the tissue) in relation extraction by enhancing the recent frameworks to extract n-ary relation with external knowledge.

2. Objectives

This project aims at developing a new tool to enhance relation extraction of n-ary relations from biomedical literature by having as main hypothesis that using domain knowledge improves relation extraction for n-ary biomedical relations.

The goals of this project can be summarised as:

1. Creation of silver standard corpus of ternary gene-phenotype/disease-tissue relations using distant supervision: The ternary relation will provide a reliable source of biological interactions more accurately describing the real system. This corpus will be used to expand relations using current literature and to train relation extraction models;
2. Develop prototype for 3-ary relations by using the existing relation extraction systems to identify solutions and challenges.
3. Develop a new model for distantly supervised n-ary relation extraction using ontologies: this will be used to extract n-ary relations in single sentences in biomedical context using enriched semantic representations, capable of dealing with automated generated corpus with labelling noise. The use of ontologies will provide context to improve prediction quality and the fusion of multiple sources of ontologies will contribute for a more robust background for the relations;

2.1. Project Workflow

2.2. Creation of a 3-ary corpus

This corpus will provide an initial prototype for biomedical 3-ary relations. Others n-ary biomedicals datasets are already available but are very scarce. There are distantly supervised datasets such as drug-gene-mutation [9] which is used for cross sentence n-ary RE and, the chemical-disease-gene, annotated at document-level with multi-label entity pairs (pairs can have more than one relation) [10]. Few gold standard also exist such as a drug-drug combination dataset with n-ary relations [11] and the PGxCorpus of drug-gene-phenotype, which extracts relations at sentence level [12]. Many of the existing interactome data do not consider the spacial context of the interactions such as tissue or cell type. So, the dataset that will be created in this task will be based on 3-ary relations of cancer gene-phenotype/disease-tissue in single sentences.

2.2.1. Implement SOTA Named Entity Recognition (NER) and Named Entity Linking (NEL) to detect gene, diseases/phenotypes and tissue concepts.

In this step ontologies that encode domains of knowledge such as diseases (Disease Ontology), phenotypes (Human Phenotype Ontology) and anatomical concepts (Foundational Models of Anatomy) will be used to provide more precise annotations. Abstracts related to cancer from the PubMed database based on the DisGeNET database which provides the PubMed ID of articles with relations of genes and variants associated with various cancers will be used to extract sentences with the three target entities. For distant supervision co-occurrence pairs of gene-phenotype, will be compared with the HPO knowledge base [13] annotations that link HPO terms to genes and pairs of gene-tissue will be compared with the with the NCG7.0 Network of Cancer Genes Healthy Drivers [14]. Since a ternary relation is being considered, the final label will be true only if both labels in distant supervision are true. The expected output for this task is a silver standard corpus, in which a subset will be validated to be used as a training input for the RE systems.

2.3. Develop first prototype of 3-ary relations

In this task, the dataset developed in task 1 will be used on existing biomedical RE methods to identify solutions and challenges. Solutions might be existing models or approaches to deal with biomedical relation extraction namely deep learning RE models such as BiOnt [15] and BERT-based models [16, 17].

All these systems have proven their utility with SOTA results, but they are still limited to extract binary relations. For the identification of the possible challenges, besides the limitation of the RE systems explained previously, an error analysis will be performed to understand critical errors present on the dataset elaboration pipeline. These errors can be on the named-entity-recognition phase, such as for example the tool mixing gene symbols with diseases acronyms and labeling the wrong entity. In the distant supervision many well-known problems can be introduced such as incomplete knowledge bases (KB) giving wrong labels [4], as well two entities co-occurring on the same sentence not always reflect a relation [3] and additionally

the long-tailed relations problems were many relations only have a few facts represented on the KB's [18].

All these problems contribute to the noise on the dataset and its origin must be understood in order to develop new tools capable of dealing and/or minimizing this noise.

3. Replicate SOTA solutions

SOTA models out of the biomedical scope will be used applying the created 3-ary corpus and others available that were previously mentioned. Replicating SOTA solutions will provide an insight of what is lacking on the SOTA's systems to achieve top performance when applied to the biomedical field. The work from Jia et al. [19] is one example of what can be used. It uses multi-scale modeling which increases precision, even in the presence of noisy labels resulting from distant supervision. This method provides SOTA results for n-ary relations in diverse document scale mention level such as whole document, paragraph and sentence, moreover it has already been proven its usefulness on the biomedical field, extracting drug-gene-mutation relation.

4. Development of an N-ary model with external knowledge

Some models have already proved that the incorporation of domain knowledge contributes to a better performance of the model [15, 20]. Thus, incorporating domain knowledge on a SOTA model might have a significant impact on the model performance regarding RE in biomedical field.

Incorporation of external knowledge can be achieved by adapting the models to integrate, besides the traditional word embeddings, the embeddings of the given external knowledge, and that will be acting as a buffer to the model relation extraction. The works of [4, 7] are some examples of these types of systems that uses stack or parallel modules that focus on different extraction of features to improve the model performance.

BiOnt, which is a model based on bidirectional Long Short-Term Memory systems that incorporates the Word2Vec word embeddings with multi-ontology integration (four types of domain specific ontologies) [15], will be used as a base system. This system takes advantage of domain-specific ontologies to provide standard vocabulary and domain integration knowledge. All the incorporated ontologies are from the biomedical field, and the system has already proven to benefit from this incorporation when performing on three distinct biomedical datasets: drug-drug, phenotype-gene and chemical-induced disease relations. The ontology embeddings used on this system allows the representation of the relations between the ancestors ontology that corresponds to the entity, and in case of tie, it is chosen the most specific term. The new system will be the expansion of BiOnt to extract ternary gene-phenotype-tissue relations using single sentences. Although using cross-sentence level for n-ary relation extraction is the most popular method, there is still a lot of information at sentence-level that can be used for n-ary extraction, such as in the example : "The previously described Met53Ile [GENE **CDKN2A**] mutation located in exon 2 was detected in a female patient with [PHENOTYPE **melanoma**] metastatic to the

[TISSUE **regional lymph nodes**], multiple primary cutaneous lesions, atypical naevi and a first-degree relative with melanoma".

5. Conclusions

The gigantic task of interpreting biomedical sentences makes it difficult to automatically extract relations. This is more difficult when more than a relation between two entities is the target and also the arising problems of using distant supervision. However in the recent years efforts have been made to incorporate n-ary relation and also reduce the noise caused by distant supervision. Thus, there is still a space to integrate and explore n-ary relations in the traditional challenges and where this project can have an important role.

6. Acknowledgments

This work was supported by FCT through project DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017, and the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020. To professor Francisco Couto and professor Francisco Pinto for advising my project.

References

[1] X. Li, J. Yang, H. Liu, Z. Liang, L. Qu, Y. Zhang, K. Shao, D. Zhao, Overview of distant supervised relation extraction, in: 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), volume 4, IEEE, 2021, pp. 1287–1292.

[2] D. Sousa, A. Lamúrias, F. M. Couto, A silver standard corpus of human phenotype-gene relations, arXiv preprint arXiv:1903.10728 (2019).

[3] D. Zeng, K. Liu, Y. Chen, J. Zhao, Distant supervision for relation extraction via piecewise convolutional neural networks, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1753–1762.

[4] Y. Li, G. Long, T. Shen, T. Zhou, L. Yao, H. Huo, J. Jiang, Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 8269–8276.

[5] L. Song, Y. Zhang, Z. Wang, D. Gildea, N-ary relation extraction using graph state lstm, arXiv preprint arXiv:1808.09101 (2018).

[6] H. Wang, M. Tan, M. Yu, S. Chang, D. Wang, K. Xu, X. Guo, S. Potdar, Extracting multiple-relations in one-pass with pre-trained transformers, arXiv preprint arXiv:1902.01030 (2019).

[7] D. Zhao, J. Wang, H. Lin, X. Wang, Z. Yang, Y. Zhang, Biomedical cross-sentence relation extraction via multihead attention and graph convolutional networks, Applied Soft Computing 104 (2021) 107230.

[8] J. J. Bianchi, X. Zhao, J. C. Mays, T. Davoli, Not all cancers are created equal: Tissue specificity in cancer genes and pathways, Current opinion in cell biology 63 (2020) 135–143.

[9] N. Peng, H. Poon, C. Quirk, K. Toutanova, W.-t. Yih, Cross-sentence n-ary relation extraction with graph lstms, Transactions of the Association for Computational Linguistics 5 (2017) 101–115.

[10] D. Zhang, S. Mohan, M. Torkar, A. McCallum, A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes, arXiv preprint arXiv:2204.06584 (2022).

[11] A. Tiktinsky, V. Viswanathan, D. Niezni, D. M. Azagury, Y. Shamay, H. Taub-Tabib, T. Hope, Y. Goldberg, A dataset for n-ary relation extraction of drug combinations, arXiv preprint arXiv:2205.02289 (2022).

[12] J. Legrand, R. Gogdemir, C. Bousquet, K. Dalleau, M.-D. Devignes, W. Digan, C.-J. Lee, N.-C. Ndiaye, N. Petitpain, P. Ringot, et al., Pgxcorpus, a manually annotated corpus for pharmacogenomics, Scientific data 7 (2020) 1–13.

[13] S. Köhler, L. Carmody, N. Vasilevsky, J. O. B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry, et al., Expansion of the human phenotype ontology (hpo) knowledge base and resources, Nucleic acids research 47 (2019) D1018–D1027.

[14] L. Dressler, M. Bortolomeazzi, M. R. Keddar, H. Misetic, G. Sartini, A. Acha-Sagredo, L. Montorsi, N. Wijewardhane, D. Repana, J. Nulsen, et al., Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the network of cancer genes (ncg) resource, Genome biology 23 (2022) 1–22.

[15] D. Sousa, F. M. Couto, Biont: deep learning using multiple biomedical ontologies for relation extraction, in: European Conference on Information Retrieval, Springer, 2020, pp. 367–374.

[16] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

[17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. URL: https://doi.org/10.1093/bioinformatics/btz682. doi:10.1093/bioinformatics/btz682.

[18] T. Liang, Y. Liu, X. Liu, H. Zhang, G. Sharma, M. Guo, Distantly-supervised long-tailed relation extraction using constraint graphs, IEEE Transactions on Knowledge and Data Engineering (2022).

[19] R. Jia, C. Wong, H. Poon, Document-level n-ary relation extraction with multiscale representation learning, arXiv preprint arXiv:1904.02347 (2019).

[20] A. Lamurias, D. Sousa, L. A. Clarke, F. M. Couto, Bo-lstm: classifying relations via long short-term memory networks along biomedical ontologies, BMC bioinformatics 20 (2019) 1–12.