lasigeBioTM at SemEval-2023 Task 7: Improving Natural Language Inference Baseline Systems with Domain Ontologies

Sofia I. R. Conceição [†] **, Diana F. Sousa** [†]**, Pedro M. Silvestre** [†]**, Francisco M. Couto** [‡]
LASIGE, Faculdade de Ciências, Universidade de Lisboa,
1749-016 Lisbon, Portugal
[†]{sconceicao,dfsousa,psilvestre}@lasige.di.fc.ul.pt
[‡]fjcouto@edu.ulisboa.pt

Abstract

Clinical Trials Reports (CTRs) contain highly valuable health information from which Natural Language Inference (NLI) techniques determine if a given hypothesis can be inferred from a given premise. CTRs are abundant with domain terminology with particular terms that are difficult to understand without prior knowledge. Thus, we proposed to use domain ontologies as a source of external knowledge that could help with the inference process in the SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT). This document describes our participation in subtask 1: Textual Entailment, where Ontologies, NLP techniques, such as tokenization and named-entity recognition, and rule-based approaches are all combined in our approach. We were able to show that inputting annotations from domain ontologies improved the baseline systems.

1 Introduction

Natural Language Inference (NLI) determines whether a given hypothesis can be deduced from a given premise (Romanov and Shivade, 2018). Particularly in the clinical text, where there is much variation in terminologies, NLI is more challenging (Romanov and Shivade, 2018). The systems have to deal with obstacles that are constantly present in the text, such as homonyms, acronyms, or abbreviations, leading to ambiguity since, without context or background, some can be mapped to diverse expanded forms that are not entirely correct in the context (Krallinger et al., 2008; Couto and Krallinger, 2020). Take, for example, a part of a sentence with this text "*Amyotrophic lateral sclerosis (ALS) patients...*" ALS here is the acronym for the disease Amyotrophic lateral sclerosis, but ALS is also the gene symbol for the human gene SOD1 superoxide dismutase 1. Without context, when only presented with the ALS word, it is not possible to know if it is referring to the disease, the gene, or even another possibility not explored in this example.

Clinical trials play a fundamental role in discovering new health information (National Library of Medicine, National Center for Biotechnology Information, 2023). In these studies, the aim is to perform pre-defined interventions in volunteers. For instance, the goal can be for clinicians to study the effects of different drug concentrations on patients. The outputs of these studies are registered in Clinical Trials Reports (CTR). These reports store all the information about the conditions to be selected to participate in the trial, groups of participants, information about dosage and duration, results, and adverse events. CTRs are rich in domains with specific terms that are not easy to grasp without prior knowledge. Domain ontologies can provide this external knowledge. Ontologies are defined by Gruber (1993) as being the "specification of conceptualization" and provide a common vocabulary with represented shared knowledge (Gruber, 1993). In this scenario, it provides domain-specific semantics to the models that can help make the connection between semantics and information extraction. Because biomedical ontologies are typically represented as directed acyclic graphs, with each node representing an entity and the edges representing known relationships between those entities, ancestors can be used to obtain further information about an entity. Ancestors will contribute with knowledge that cannot be directly assessed in the text (Lamurias et al., 2019). Incorporating this domain knowledge might be significant to grasp all the subtleties and richness of biomedical writing when using Natural language Processing (NLP) approaches to get more accurate predictions.

This paper presents the participation of our team, lasigeBioTM (user *dpavot*), at the SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) subtask 1 (Jullien et al., 2023). This task provides a collection

of breast cancer CTRs and statements about them to infer the relation label of entailment or contradiction. Our approach combines ontologies, NLP techniques such as tokenization and named-entity recognition, and rule-based approaches. All code and steps to reproduce the results regarding this participation are available online [1]. The main goal of our participation was to assess if the introduction of external knowledge provided by domain ontologies would improve baseline systems approaches similar to the one provided in the challenge starter kit.

2 Related Work

There are some instances where authors integrated domain ontologies with the biomedical text. Lamurias et al. (2019) created the BO-LSTM by incorporating biomedical ontologies and ancestry information alongside a deep learning Long Short-Term Memory model. BO-LSTM was developed to extract drug-drug interactions using the ChEBI ontology, and it demonstrated that integrating ontologies enhanced categorization made by the model. Instead of the whole instance, the authors used the Shortest Dependency Path between the target entities. Besides the word embeddings, the BO-LSTM model incorporates WordNet as an external source of information, a generic English language ontology, and domain-specific ontologies. Additionally, each entity was matched to an ontology concept in order to obtain their ancestors.

Using the previous system as a base, Sousa and Couto (2020) created the BiOnt, which expands the BO-LSTM with a multi-ontology integration (four types of domain-specific ontologies) and uses WordNet hypernyms. It uses Gene Ontology (GO), Human Phenotype Ontology (HPO), Human Disease Ontology (DO), and the Chemical Entities of Biological Interest ontologies (ChEBI), which can be combined in order to classify ten distinct types of relations. Using three distinct datasets that represented drug-drug interactions, phenotype-gene relations, and chemical-induced disease relations, BiOnt had an improvement of 4.93%, 4.99%, and 2.21% of the F1-score in each dataset, respectively.

Some studies already incorporate domain knowledge to perform NLI tasks. A combination of biLSTM with attention word embeddings with definitions of medical concepts provided by the Unified Medical Language System (UMLS) was used to perform NLI on clinical texts (Lu et al., 2019). Another study from Sharma et al. (2019) also employs the domain knowledge provided by the UMLS by incorporating it by knowledge graph embeddings and combining it with the BERT-based language model BioELMo.

3 Methodology

This section describes the pipeline used at the SemEval-2023 Task 7 Subtask 1. Fig. 1 provides a representation of our system.

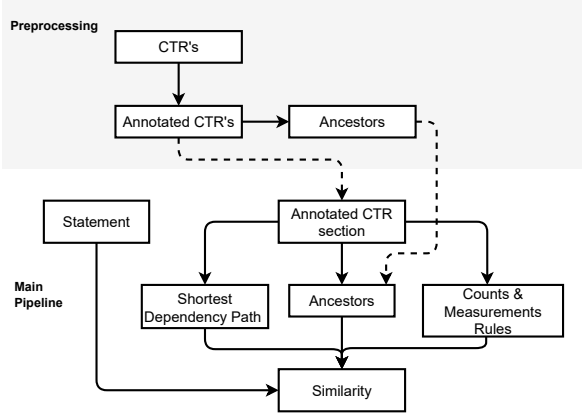

Figure 1: Preprocessing and main pipeline overview. In the preprocessing phase, annotations and ancestors are obtained, and then in the main pipeline, the target section is combined with information from the shortest dependency path, ancestors, and measurement rules. The output is given by the similarity of the statement and the enriched CTR section text.

3.1 Task Description

The SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) (Jullien et al., 2023) consisted in using NLI techniques to narrow the gap regarding the high volume of produced CTRs. In subtask 1: Textual Entailment, the main goal was to determine the inference relation within the statement and the CTR, if it was entailment or contradiction.

Experts in the domain created statements. They could have two different compositions, one only making statements about a single CTR or another where there was a comparison between two CTRs.

The domain experts also produced the CTRs divided into four main sections: eligibility criteria, intervention, results, and adverse events.

[1] https://github.com/lasigeBioTM/SemEval2023_Task-7

3.2 System Overview

Our system uses a generic integrated rule-based NLP system employing external information from ontologies. Scispacy (Neumann et al., 2019), a python package containing models for processing biomedical, scientific, or clinical text, was used for the NLP functions. The tokenization and dependency parsing model was *en_core_sci_lg*, which contains 785k vocabulary and 600k word vectors for biomedical data.

First, we performed a preprocessing step to extract annotations of entities of interest in all CTRs. Next, we obtained the ancestors of all unique entities. In the main pipeline, using the files containing the statements, given the CTRs IDs, the denoted text section is selected from the annotated CTR file. We used this section text to extract the shortest dependency path (SDP) between pairs with the ancestors of the respective annotations. If a numerical value is presented in the statement, rules for simplifications of the numerical comparison are also performed. Finally, we add the SDP, annotations, ancestors, and rules to the text section. We compare the combined information from the CTR and the hypothesis using the similarity function from the scispacy tokenization. The similarity function provides a scalar similarity score between 0 and 1, with the highest score indicating more similarity.

3.2.1 Annotations

To enrich the dataset provided by the task organizers, we annotated scientific/medical entities of interest by linking them to eight different ontologies identified as relevant. These ontologies where BioAssay Ontology (BAO) (Abeyruwan et al., 2014), clinical LABoratory Ontology (LABO) (Barton et al., 2019), Ontology of Adverse Events (OAE) (He et al., 2014), Chemical Entities of Biological Interest (ChEBI) (Degtyarenko et al., 2007), Human Disease Ontology (DO) (Schriml et al., 2022), Gene Ontology (GO) (Ashburner et al., 2000), Human Phenotype Ontology (HPO) (Robinson et al., 2008), and Clinical measurement ontology (CMO) (Shimoyama et al., 2012). Respectively, these ontologies target drug screening data (e.g., *is substrate of*, BAO:0000117), laboratory test specifications (e.g., *has maximal value*, LABO:0000119), adverse advents (e.g., *induces*, OAE:0000186), chemical entities (e.g., *iron*, CHEBI:82664), human diseases (e.g., *meningioma*, DOID:3565), gene products (e.g., *single strand break repair*, GO:0000012), human phenotypes (e.g., *arachnodactyly*, HP:0001166), and clinical measurements (e.g., *R wave duration*, CMO:0000271). The distribution of recognized entities per ontology and per type of document is presented in Table 1.

Ontology	CTR	Train	Dev	Test
BAO	12417	200	15	42
LABO	1662	17	0	2
OAE	4486	328	36	91
ChEBI	17434	548	43	191
DO	16941	415	71	122
GO	1212	14	1	3
HPO	14906	512	51	146
CMO	874	29	2	4
Total	69932	2063	219	601

Table 1: The distribution of entities per ontology and per type of document that was provided.

Since each entity is matched to an ontology concept with a unique ID, it was possible to obtain the ancestors. For each unique ID, a list of all ancestors and most common labels were obtained. These ancestors were then used to enrich the CTRs sentence with annotations.

3.2.2 Counts and Measurements Rules

Quantitative evidence in the form of counts and measurements plays a crucial role in scientific discourse, providing support for findings (Harper et al., 2021). This type of data is essential for obtaining accurate and precise measurements to ensure reliable data and valid conclusions. Extracting numerical data from the stext, along with its associated entities and time scales, improves the structure and analysis of information. Comparing numbers in the text provides context and enhances its meaning, leading to better-informed decisions and analysis. This process can lead to a better comparison between the given Statements and the CTR information available.

Considering the following example:

Clinical Trial ID: NCT02953860
Statement: "Patients in NCT02953860 receive more mg of Enzalutamide than Fulvestrant over the course of the study."
Label: "Entailment"
CTR line: "500mg of Fulvestrant will be given IM on days 1, 15, 28, then every 4 weeks as per

standard of care (SOC) and 160mg of Enzalutamide will be given, in conjunction with Fulvestrant, PO daily."

From this CTR text, it is important to extract:

- count: 500, measure: mg, of: Fulvestrant, days: [1,15,18], weeks: every 4;
- count: 160, measure: mg, of: Enzalutamide, days: daily.

With this information is possible to conclude that more milligrams of Enzalutamide were given when compared to Fulvestrant, making the label of the statement Entailment.

3.2.3 Shortest Dependency Path

One method to decrease information noise is using the Shortest Dependency Path (SDP). The SDP allows filtering the minimal necessary information between two identified entities in the text (Xu et al., 2015). For each sentence with annotated entities, we obtain the Shortest Dependency Path (SDP) between each pair of entities. Given the sentence "known untreated or active central_nervous_system (cns) metastases" the resulting SDP is 'active' - 'untreated' - 'metastases' - 'central_nervous_system'.

4 Results and Discussion

Our team (user *dpavot*) reached the overall ranking in the 21st position, obtaining a 0.661 (18) F1-score, 0.511 (23) precision, and 0.936 (5) recall. Although this shows some improvement, the obtained scores in a balance binary task are the result of the model mainly predicting the entailment class. The best-performing combination consisted in having ancestors on single CTRs while removing them from the CTRs comparison.

CTRs comparison statements were the ones that provided a significant challenge to our pipeline. Comparison statements consisted of 30% (60) of the development set. For these occurrences, our approach consisted of joining the section text from each CTR as a unique text. This method resulted in noise since the text's origin was lost, and it was impossible to distinguish which part of the text belonged to which CTR. The approach also resulted in a massive comparison text, with much non-relevant information to compare with the statement.

Regarding count rules, we explored in which situations it fails. One of the reasons count rules may fail is if the measurement of the number is not included in the list of created measurements. This is also true for timespan measurements ("day", "week", "month", *et cetera*).

Spacy is used to verify the entity "of" to which the number refers. However, it's important to note that Spacy may not always accurately identify the children or the right entity of the number, which can lead to inaccurate conclusions. One practical example of this occurrence is demonstrated in the following example:

Considering the CTR text: "Documented menopausal status premenopausal (having menstrual periods or FSH <35) or postmenopausal (12 months since last menstrual period with intact uterus and at least one ovary or FSH 35 or previous bilateral oophorectomy." Our rules, wrongly extract:

- count: 35, of: postmenopausal, months: 12.

4.1 Ablation Studies

In order to understand if the introduction of the annotations produced a positive effect on the baseline systems, we performed ablation studies using the development set. Although we did not insert annotations in the starter kit, it was used as the primary baseline of comparison. Our ablation studies consisted in running the pipeline without annotated CTRs, without the SDP, and without the Counts and Measurements Rules. The results are present in Table 2.

Method	F1	P	R
NLI4CT Starter Kit	0.502	0.486	0.520
No Annotated CTRs	0.522	0.515	0.530
No SDP	0.66	0.500	0.970
No Count Rules	0.662	0.500	0.980
Full Pipeline	0.667	0.500	0.980

Table 2: Scores from the development set.

These results showed that each one of the inserted methods improved the baseline results. Moreover, it showed that the combination of these three achieves the highest scores.

4.2 Complementary Experiments

In developing our pipeline, several ensembles were tested to find which combination would result in better performance. The tried approaches are described in this section.

Leaving all the full text from the section produced worse results than using only the annotated sentences. This outcome may be because the relevant entities regarding the statement are identified in the annotation process discarding sentences that do not contribute meaningfully to the inference. Annotations with SDP were also tested on the statements, but the ensemble got worse results than just using the statement text without additions. These results could be due to the short size of the statements, making the annotations so close that the produced path did not hold enough relevant information.

5 Conclusion

Although the state-of-the-art employs deep learning techniques, these are very "hungry" for data and sometimes do not have the pervasiveness ability when encountering different conditions from those on the training set (Romanov and Shivade, 2018). Since the provided dataset was small, we explored the combination of rule-based NLP enriched with external information. As previously stated, our goal was to evaluate if inputting annotations from domain ontologies could improve the baseline systems. Our results showed that the addition of the entities annotations improved these systems.

As for future work, we wish to explore incorporating thresholds regarding non-specific annotations and ancestors to generate more significant annotations.

Acknowledgements

This work was supported by FCT through funding of LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020); and FCT and FSE through funding of PhD Scholarship (ref. SFRH/BD/145221/2019) attributed to DFS; FCT through funding of PhD Scholarship (ref. UI/BD/153730/2022) attributed to SIRC and LASIGE (ref. UIDP/00408/2020), funded by FCT, credited to PMS.

References

Saminda Abeyruwan, Uma D Vempati, Hande Küçük-McGinty, Ubbo Visser, Amar Koleti, Ahsan Mir, Kunie Sakurai, Caty Chung, Joshua A Bittker, Paul A Clemons, et al. 2014. Evolving bioassay ontology (BAO): modularization, integration and applications. In *Journal of biomedical semantics*, volume 5, pages 1–22. BioMed Central.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Adrien Barton, Paul Fabry, Luc Lavoie, and Jean-François Ethier. 2019. LABO: An ontology for laboratory test prescription and reporting. In *JOWO*.

Francisco M. Couto and Martin Krallinger. 2020. Proposal of the first international workshop on semantic indexing and information retrieval for health from heterogeneous content types and languages (SIIRH). In *Advances in Information Retrieval*, volume 12036, pages 654–659, Cham. Springer International Publishing.

Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350.

Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021. SemEval-2021 task 8: MeasEval – extracting counts and measurements and their related contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316, Online. Association for Computational Linguistics.

Yongqun He, Sirarat Sarntivijai, Yu Lin, Zuoshuang Xiang, Abra Guo, Shelley Zhang, Desikan Jagannathan, Luca Toldo, Cui Tao, and Barry Smith. 2014. OAE: the ontology of adverse events. *Journal of biomedical semantics*, 5(1):1–13.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*.

Martin Krallinger, Alfonso Valencia, and Lynette Hirschman. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome biology*, 9(2):1–14.

Andre Lamurias, Diana Sousa, Luka A Clarke, and Francisco M Couto. 2019. BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. *BMC Bioinformatics*, 20:10.

Mingming Lu, Yu Fang, Fengqi Yan, and Maozhen Li. 2019. Incorporating domain knowledge into natural language inference on clinical texts. *IEEE Access*, 7:57623–57632.

National Library of Medicine, National Center for Biotechnology Information. 2023. About clinicaltrials.gov. https://beta.clinicaltrials.gov/about, Last accessed on 2023-02-09.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Lynn M Schriml, James B Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J Allen Baron, Rebecca Jackson, Susan M Bello, Cynthia Bearer, et al. 2022. The human disease ontology 2022 update. *Nucleic acids research*, 50(D1):D1255–D1261.

Soumya Sharma, Bishal Santra, Abhik Jana, Santosh Tokala, Niloy Ganguly, and Pawan Goyal. 2019. Incorporating domain knowledge into medical NLI using knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6092–6097, Hong Kong, China. Association for Computational Linguistics.

Mary Shimoyama, Rajni Nigam, Leslie Sanders McIntosh, Rakesh Nagarajan, Treva Rice, DC Rao, and Melinda R Dwinell. 2012. Three ontologies to define phenotype measurement data. *Frontiers in genetics*, 3:87.

Diana Sousa and Francisco M. Couto. 2020. BiOnt: Deep learning using multiple biomedical ontologies for relation extraction. In *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science*, volume 12036, pages 367–374, Cham. Springer International Publishing.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1785–1794, Lisbon, Portugal. Association for Computational Linguistics.