An End-to-End Approach for Asserted Named Entity Recognition and Relationship Extraction in Biomedical Text

Nourah M Salem[1], Elizabeth K White[1], William Baumgartner[1], and Lawrence E. Hunter[1]

[1]Computational Bioscience Program, University of Colorado, Anschutz Medical Campus, Aurora, CO, 80045, USA

*Corresponding author: nourah.salem@cuanschutz.edu

Abstract

In this study, we focus on subtask 2 of the BioRED track for extracting and analyzing biomedical entities and their relationships from biomedical literature. We developed an end-to-end framework that uses a series of Large Language Models (LLMS), such as the Flair model, to identify various biomedical entities and pass them to BioBERT for relation extraction. To augment the system's performance, we incorporated coreferencing resolution along with the use of resources like CRAFT and Pubtator to enrich our training data for Named Entity Recognition (NER). Moreover, we applied similarity measures for the linguistic contexts of named entities to match their mentions over longer distances. We used positional data to assess the odds that the relations we found might be novel. Finally, we used PheKnowlator, a graphical knowledge base, to get insights into the contextual environment of the entities and weigh the likelihood of them participating in particular relations. Although our work is preliminary, these techniques show promise for finding relations between entities in biomedical papers, even when the relations are subtle, when they span longer distances, or when they are implied rather than stated directly.

Introduction

Extracting biomedical entities and the relations between them is critical to advancing researchers' understanding of biomedicine (1-4). Publications like the ones in PubMed/PMC provide a rich source of these entities and the relations that connect them, but also pose challenges due to their unstructured natural language: entity mentions can take many forms, and relations between them can be distant, implicit, or constructed by many smaller relations.

Advancements in machine learning and natural language processing are pivotal in biomedical text mining. Large language models like BioBERT and Flair are carving a path through complex tasks like Named Entity Recognition (NER) and Relationship Extraction (RE) (5,6). BioBERT's specialized pre-training on biomedical corpora enables a more nuanced contextual understanding, while Flair's strength lies in contextual string embeddings. We used both models within the BioCreative VIII Track 1 (BioRED) tasks (7). We also examined the effectiveness of additional techniques to better understand the input data, augment sparse training data, and validate certain outputs. Techniques like coreferencing resolution associate different expressions with single entities, enriching the presence of biomedical mentions (8). Knowledge graphs offer structured representations of interconnected biomedical entities and can validate extracted relations (9). Cosine similarity is instrumental in analyzing and comparing contextual similarities between entity pairs potentially in a relationship (10). We are also exploring positional information about where novel relations are presented in the title and abstract to see if this aids in distinguishing novel versus established relations.

Material and Methods

Database

For the NER and ID mapping tasks, 16833 entities spanning the 6 biological entity types in the BioRED training data were used for training NER models in addition to 17847 already harmonized and standardized annotated entities from PubTator (11) that belong to the 6 entity types from BioRED. Only the matching entity types (CHEBI, PR, and CL) were selected from 67 CRAFT fully annotated articles (12) for the NER task only. Also, 6456 mentions of the BioRED entities were identified using the coreferencing technique.

For the RE task, we paired the types of entities that can have relations with each other based on the BioRED task entities matching reference. We allowed the relations to span one, two, three, or four sentences. Later, we filtered them down using cos similarity to 29,120 samples from BioRED and downloaded 3326 single sentences from PubMed for the underrepresented classes (Bind, Cotreatment, Drug Interaction, and Conversion) based on the presence of certain cues from a list we built to represent each class, such as: ("bind", "dual therapy", "converted to", "drug-drug interaction").

Methodology

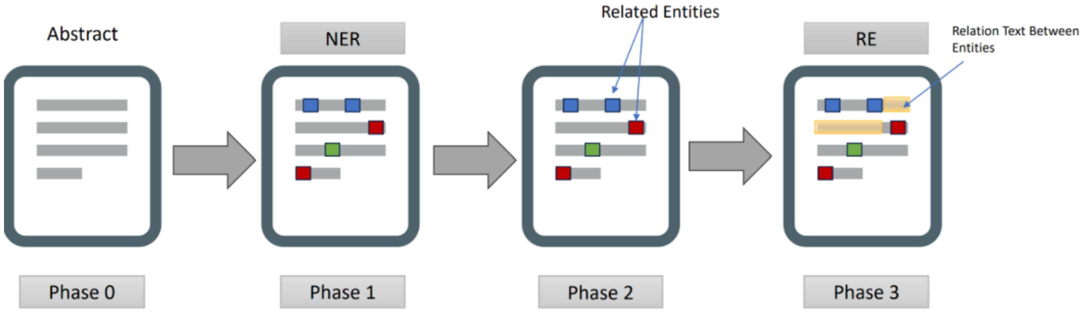

Figure 1. The four-phase process of extracting Biomedical entities and relationships from abstracts: Phase 0 begins with the raw abstract, Phase 1 involves Named Entity Recognition (NER) to identify entities, Phase 2 connects related entities, and Phase 3 highlights the textual relations between these entities.

Phase 1: Our pipeline begins with identifying the BioRED-selected set of entity types. We trained seven Flair models for each entity using the BioRED training set, fine-tuning the embeddings with the PubMed and PMC corpus. Using the SequenceTagger class, the models have a hidden size of 256, stacked embeddings, and a CRF for sequence modeling. Additionally, we developed another integrated Flair model for all six entities using PubTator data and another model based on the CRAFT full-text articles, for the entity types: Chemical and Cell line. The comprehensive coverage of full articles provides more context around the entities, which is useful for tasks that require understanding beyond the abstract.

Phase 2: extracted entities are mapped to relevant identifiers. We gathered 500K mapped entities from PubTator articles, using them as a reference dictionary to link entities with their appropriate IDs. The focus then shifts to pairing related entities. In this pursuit, a couple of approaches were tested. We initially used the brute-force method of pairing sequential entities in text. it faced the issue of lengthy context spans including irrelevant text. To counter this, the strategy was adjusted to extracting relationships within single sentences only to ensure that the text relevance is maintained. Using fine-tuned BioBERT, our second approach is based on a binary classifier that categorizes text into having a relation or not. Here, statements that contain pairs of entities are the positive samples, while the negative examples are the ones that don't have any pairs. For the second approach: instead of only focusing on single sentences, we opted to regulate text segment extraction around entities based on their contextual similarity, assessed using the 'paraphrase-distilroberta-base-v1' Sentence Transformer model that feeds to cosine similarity functions. Only contexts with a similarity score above 0.5 were chosen for the classifier's training. Table 1 provides insights into the optimal window sizes derived from various tests and manual evaluations.

Table 1: representing the suggested window size for context similarity measure based on the average text length and number of sentences in the text.

Number of sentences spanning entities	***Ave. Text Length (Word Count)***	***Window Size***
1	*0 - 30*	*Full text*
2	*30 - 53*	*20*
3	*53 - 75*	*30*
4	*75 - 95*	*40*

The final step in this phase examines the use of knowledge graphs to validate the binary model predictions by verifying existing paths between entities to confirm their relationships. We used PheKnowLator (13), a knowledge graph that integrates ontologies from multiple domains, including biomedical ontologies, literature, and databases, providing a rich and comprehensive representation of biomedical Knowledge. For traversing the graph data, we used the Depth-First Search (DFS) to note already visited nodes, ensuring an efficient and loop-free exploration. Importantly, there is no restriction on the length of the paths explored. This can ensure that identified relationships between entities are not only existent but also biologically or clinically meaningful.

Phase 3: We categorized the extracted sentences based on the types of relationships they represent and determined their novelty. To accomplish this, we trained two fine-tuned BioBERT models. The fine-tuning for all BioBERT models included adding a dropout layer for regularization, two fully connected linear layers for transformation, and a LogSoftmax activation function to ensure that the output values can be interpreted as log probabilities for each class, facilitating effective multi-class classification. The initial model classifies the input text into one of the 8 relationships as outlined by BioRED, while the second one determines if the identified relationship presents novel information—essentially distinguishing between key findings of a manuscript and previously established knowledge.

Results and Discussion

For run1: using subtask 2 test data, which consists of abstracts only, the NER models effectively extracted biomedical entities, primarily utilizing the PubTator-trained model for all entity types extraction. For sequence variants, a distinct model from the BioRED seq-var training dataset was used. The CRAFT model served as a validation mechanism for PubTator extractions. We also used PubTator annotations dataset for ID mapping, yielding a precision (72.36%) surpassing the average in the normalization task (69.02%). Overall, the integration of PubTator annotations, whether for NER model training or ID mapping, demonstrated noteworthy performance compared to the average F1 scores (76.87% and 63.36%) for NER and ID respectively. Run 2 was omitted, as its results closely mirrored Run 1. Table 2 represents each task's results over the different runs. Conversely, there was a significant decline in scores starting from the entity pairing task, which persisted in subsequent tasks. We attribute this low pairing performance to factors: (a) the sub-optimal brute-force pairing approach of only single sentences carrying entity pairs and (b) our entity set in this trial lacked 26.86% of its true positives. We also believe that the (Relation, Novelty) low performance is also caused by the initial incorrect pairings.

Table 2: Performance metrics of different NLP tasks across multiple runs, showcasing Precision (P), Recall (R), and F1-score (F) values.

Task	Run1 (P/ R/ F)	Run3 (P/ R/ F)	Run4 (P/ R/ F)
NER	**0.72/ 0. 73/ 0.72**	-	-
Normalization (ID)	**0.72/ 0.60/ 0.65**	-	-
Entity pair	**0.31**/ 0.03/ 0.06	**0.29**/ **0.10**/ 0.15	0.19/ **0.38**/ 0.25
Entity pair+Relation type	0.16/ 0.01/ 0.03	0.14/ 0.04/ 0.07	0.08/ **0.18**/ 0.11
Entity pair+Novelty	0.18/ 0.01/ 0.03	0.18/ 0.06/ 0.09	0.13/ **0.27/** 0.17
Entity pair+Relation type+Novelty	0.09/ 0.01/ 0.01	0.08/ 0.03/ 0.04	0.05/ **0.12/** 0.07

For Run 3, with the availability of the subtask1 test data, which included the annotated gold standard entities zalong with the abstracts, we had access to the complete list of entities for pairing. We opted for the binary classifier approach to validate possible pairings instead of using the brute-force algorithm, focusing still on single sentences containing relations. The precision remained consistent, indicating a similar true positive to the false positive ratio for the extracted pairs in this sample. Notably, there was an increase in recall to 10%, meaning that the model identified 10% of the correct pairs. This positive outcome prompted us to evaluate larger text segments encompassing the entities using the same binary classifier approach for run 4.

Run 4: we started using cosine similarity in this run as we started increasing the text containing relation spans. We examined contextual proximity for entity pairs across up to four sentences. Of 800K refined segments, 384K were deemed relation-bearing by our binary classifier. This run yielded 38% recall, which is near the subtask 2 entity pairing recall average (40%). However, the precision dropped, indicating increased false positives. Therefore, we anticipate that a higher filtering similarity threshold (> 0.5) is required to extract the very close context surrounding entities, and larger spans (4+ sentences to have relations) can decrease False positives.

For the PheKnowLator trial, we verified 292 relationships amongst entity pairs, from 207 abstracts. On the other hand, when considering a brute-force pairing strategy, which produced 1778 entity pairs from the same number of abstracts, it potentially included a huge number of possible false positives and non-informative relations. This comprehensive verification procedure highlights the accuracy and reliability of the relationships identified using KGs as references.

Future Directions

We conducted a simple analysis to determine the position of novel relations (spanning only single sentences) in the abstracts of the training dataset, hypothesizing that they're typically found at the end of the abstracts. By analyzing their normalized positions, we found around 32% of these novel relations exist in the last 86% section of the abstracts. Further samples can prove a significant location which in turn, can aid in validating the novelty model's predictions by examining the position of the novelty statement in the abstract.

A conjunction technique can improve deciphering complex relationships and entities within scientific texts, providing a simpler and more clear association between related entities. We hope to use it for 2+ entities composing relation.

References

1. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. (2007) Frontiers of biomedical text mining: current progress. Briefings in Bioinformatics., 8(5):358–375.

2. Lars Juhl Jensen, Saric J, Bork P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics., 7(2):119–129.

3. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. (2008) Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics., 9(1):207.

4. Simmons M, Singhal A, Lu Z. (2016) Text Mining for Precision Medicine: Bringing structure to EHRs and biomedical literature to understand genes and health. Advances in experimental medicine and biology.,939:139–166.

5. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining Wren J, editor. Bioinformatics.;36(4).

6. Akbik A, Blythe D, Vollgraf R. (2018) Contextual String Embeddings for Sequence Labeling.,1:1638–1649.

7. Luo L, Lai P-T, Wei C-H, Arighi CN, Lu Z. (2022) BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics.,23(5).

8. Ng V. (2010) Supervised Noun Phrase Coreference Research: The First Fifteen Years. ACLWeb.,1396–1411.

9. Ehrlinger L, Wöß W. (2016) Towards a Definition of Knowledge Graphs.

10. Google A. (2001) Modern Information Retrieval: A Brief Overview.

11. Wei C-H, Allot A, Leaman R, Lu Z. (2019) PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Research.,47(W1):W587–W593.

12. Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, Baumgartner WA, Cohen KB, Verspoor K, Blake JA, et al. (2012) Concept annotation in the CRAFT corpus. BMC Bioinformatics.,13(1).

13. Callahan TJ, Tripodi IJ, Stefanski AL, Cappelletti L, Taneja SB, Wyrwa JM, Casiraghi E, Matentzoglu NA, Reese J, Silverstein JC, et al. (2023) An Open-Source Knowledge Graph Ecosystem for the Life Sciences.