Probability model with Ensemble learning and Data augmentation for named entity recognition (NER) and relation extraction (RE) tasks

Cong-Phuoc Phan, Gia-Han Ngo, Ben Phan, and Jung-Hsien Chiang*

Department of Computer Science and Information Engineering, National Cheng Kung University, No.1, University Road, Tainan City 701, Taiwan

*Corresponding author: E-mail: jchiang@mail.ncku.edu.tw

Abstract

Biomedical Natural Language Processing (NLP) is an innovative field that uses advanced computational techniques to extract and utilize information from biomedical literature. It enables researchers and healthcare professionals to access, analyze, and apply textual data for various purposes, including clinical decision support, drug discovery, and knowledge discovery. In this paper, we introduce a multi-techniques approach to biomedical relation extraction and named entity recognition, demonstrating competitive performance when evaluated using Precision, Recall, and F1 Score in comparison to existing methods.

Introduction

Biomedical Named Entity Recognition (NER) and Relation Extraction (RE) play a critical role in drug discovery, clinical decision support, and life science research by identifying and categorizing entities like genes, proteins, diseases, and drugs. These entities are key for information retrieval, literature curation, and knowledge extraction from vast unstructured biomedical data. Different ontologies, such as MeSH for chemicals, dbsnp for variants, NCBI Taxonomy for Species, and OMIM for diseases, are used for NER to annotate entities with their identifiers, Figure 1 shows an example of NER annotation with two different ontologies. An identifier is a unique code or name for multiple entities sharing the same concept. For instance, both "Hepatocyte nuclear factor-6" and "HNF-6" are annotated with the identifier 3175 in NCBI Gene Ontology, and "End-stage renal disease" and "ESRD" share the identifier D007676 in *MeSH*. There are several ways to categorize these identifiers, such as *Pubtator* (1) classify them into *Gene*, *Chemical*, *Species*, *Diseases*, *Mutation*, and *CellLine*, while BioRED (2) modified the *Variant* to *Mutation*. There are the other classes such as *Chromosome*, *Protein* and more. NER results are used to uncover connections through RE, helping researchers understand complex biological and medical phenomena. These connections are categorized into types like

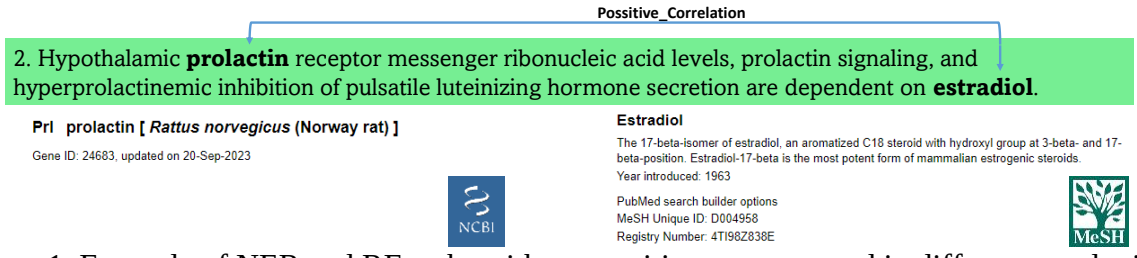

Figure 1. Example of NER and RE tasks with two entities are annotated in different ontologies.

Positive Correlation, Negative Correlation, Association, Binding, Drug Interaction, Co-treatment, Comparison, and Conversion, as defined by BioRED, based on the document's context. Notably, an additional annotation factor known as *Novelty* has gained attention and been introduced as a property of relations. It serves to indicate whether a relation is a well-established fact or a novel discovery, thus enhancing the utility of new knowledge extraction and mitigating redundancy in information. In order to facilitate the aforementioned tasks and fulfill various NLP requirements in the field of biomedicine, numerous text corpora have been developed, accompanied by extensive research efforts. (3-7). Nevertheless, the majority of these methods focus on the relationship between two entities within a single sentence, even though numerous documents demonstrate that the actual relations can extend across multiple sentences. In addition, these methods pay attention to only a specific domain such as drug-drug interaction (4), and protein-protein interaction (5), so that lack of useful knowledge that the other domain may include. To the best of our knowledge, the most recent advancements in this field are BioRED and BioREX methods. BioRED (2) operates by taking document abstracts with annotated entities and identifiers as input data, subsequently leveraging PubmedBERT for fine-tuning and predicting relations, relation types, and the novelty property. While BioRED excels in relation extraction tasks, it does not encompass named entity recognition (NER) tasks. BioREX(8) offers enhanced relation extraction performance, but it lacks NER task integration and does not include novelty prediction. On the NER front, AIONER(9) is a high-performing model for extracting biomedical entities, but it doesn't cover biomedical identifier identification, a key element for relation extraction tasks.

In an effort to combine the strengths of the aforementioned methods and address their respective functional limitations, such as AIONER's absence of identifier mapping, BioRED's omission of NER tasks, and BioREX's lack of novelty prediction, we introduce our data-centric AI solution. This solution is designed to comprehensively meet all the previously mentioned requirements.

Material and Methods

Database and evaluation metrics: To develop and evaluate our system, we use data from BioCreative VIII challenge, which contains 600 abstracts with full annotated entities and relations for system development and 10,000 documents for final evaluation. Three main metrics are used to evaluate the performance of each team are Precision, Recall, and F1 score.

Method description: Our method covers a full process from naming entity recognition to relation extraction and novelty prediction. This full process can be split into three activities: NER, Relation extraction, Novelty prediction.

1) NER: The role of this activity is to extract the biomedical words from a given text and then categorize these entities into the entity types and entity identifier. There are three approaches we used to apply:

Approach 1: Use SciSpacy with the en_ner_bionlp13cg_md model for entity extraction, followed by a Transformer model for classifying entity types and identifiers using training data and data from MeSH, OMIM, Mutation2Pubtator, Cellosaurus, and Species2Pubtator.
Approach 2: Utilize AIONER with fine-tuning to extract entities from input text, then apply the Approach 1 model for identifier information extraction.

Approach 3: Use regular expressions to identify pre-existing entities in a corpus of 600 annotated abstracts, eliminate duplicates, and store the results as a dictionary file for filtering incoming text.

Combining any of the two options can lead to duplicated entities with distinct types. For instance, the entity "ArsB" can occur as both a Chemical (C581941) and a Genes (OMIM: 611542) entity. To address this, we created a probability model based on entity type co-existing within documents. This model comprises a *probability score table* and a *fulfillment method*.

Probability table: This is the set of probability scores of co-occurrences of entity types of pairs which calculated by the mean number of entities of each type existing in the same document. Generally, the number of pairs in table P is calculated by the number of permutations of the number of entity types which is depicted in ***Equation 1.*** Here, we set k = 2, and open for further

$$P(n,k) = \frac{n_t!}{(n_t - k)!}$$

where:

n_t : the number of entity type in our dataset.

k : the number of entity types existing in the same document.

Equation 1: Number of co-existing pairs of entity types

$$P = [p(et_i, et_j)] = [(\frac{\sum e_i}{n_{ij}}, \frac{\sum e_j}{n_{ij}})],$$

where:

et_i, et_j	: two entity types which co-exist in the same document.
e_i, e_j	: the corresponding number of entities in et_i and et_j
p(et_i, et_j)	: the probability of co-existence pair between et_i, et_j
n_{ij}	: the number of documents which contain both et_i and et_j

Equation 2: Probability scores of each co-existing pairs of entity types

$$fulfilScore(a, docA) = \frac{\sum_{et_i \, \epsilon \, docA} \left[\frac{|et_i| * P(et_i, et(a))[1]}{P(et_i, et(a))[0]}\right]}{|docA(et_i)|}$$

where:

a, et(a)	*:* the entity and its type need to add into the entity list of docA
docA	*:* the specific document needs to add entity a
et_i,	: the entity type which is existing in docA.
P(- , -)[0], P(- , -)[1]	: the left and the right values of the table P at the pair (-,-)
$\|docA(et_i)\|$	: the number of entity type in docA

Equation 3: Fulfilment score calculation

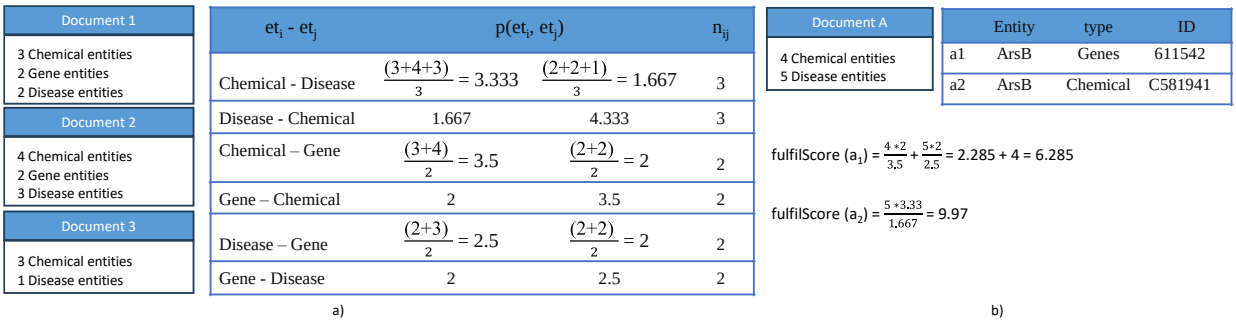

Figure 2: a) Construct probability table from 3 sample documents, b) Calculate fulfilment scores for two sample entities a_1 and a_2, the result show that the a_2 is higher prefer than a_1

improvement, and so, only two entity types co-existing are concerned in our system. The score of each probability in P is constructed by Equation ***2***.

In Figure 2 show example for 3 sample documents with 3 entity types, 6 possible pairs of entity types. 3 documents contain Chemical – Disease co-existence, while 2 documents contain only Chemical – Gene or Gene – Disease.

Fulfilment method: This is an equation (Equation 3) used to calculates the *fulfilment score* based on Probability table. Each document has a finite set of entity list from merging between any NER approaches. An entity would be added to the entity list of a specific document must have the greater fulfillment scores.

2) Relation extraction: Extract all possible identifier pairs from the NER and evaluate them for relation extraction using three methods: Data Augmentation, Pretrained Model Fine-tuning, and Ensemble Learning.

Data Augmentation: To enhance the limited data for training, we built a module to retrieve responsed data from Claude and Bingchat through third-party APIs. Specifically, the module sent several requests to paraphrase the input text to create similar sentences in the same context.
Finetune pretrained models: Apply 2 pretrained models, AutoModelForSequenceClassification and BertForSequenceClassification, to extract relations. These models take input sentences with tagged entity types and return SoftMax outputs, return the relation type with the highest score.
Ensemble Learning: Merge the two pretrained models output and give the best predictions.

3) Novelty detection: Use same models with Relation Extraction with different training approaches:

None-No-Novelty: Based on the relation detection results, remove all the "None" relation, and predict No, Novel for each relation prediction.
No-Novelty: Not depending on the detected relations set but based on entity pairs. For each entity pair's relation prediction, predicts if its novelty is none/no/novel by using rules:

(Novelty = "None" & Relation != "None") => Relation = "None"

(Novelty = Novelty) => Relation = predicted relation

Results and Discussion

The result of each of following runs is a combination of approaches in each mentioned activity.

Run 1: Entity pairs are extracted from PubmedBERT with enhanced by probability model.
Run 2: Entity pairs from Run 1 without enhanced by applying probability model & fulfilment.
Run 3: Entity pairs extracted by AIONER.
(3 runs use the data augmentation and pretrained models without Ensemble learning.)

In the unofficial submission for RE tasks, we apply 5 runs as described:

Run 1: Ensemble Learning with the fixed weights are the scores from official submissions. The existing relations are removed and replaced by voting results.

Table 1: Evaluation results returned by BioCreative Organizers (in the order: Precision/Recall/F1Score)

Official runs	NER			ID			Entity Pair			EntityPair|RelationType			Entity Pair|Novelty			Entity Pair|RelationType|Novelty		
Subtask2-Run1	0.83	0.74	0.78	0.78	0.74	0.76	0.47	0.33	0.39	0.35	0.24	0.29	0.37	0.26	0.31	0.27	0.19	0.22
Subtask2-Run2	0.60	0.80	0.69	0.53	0.78	0.63	0.30	0.40	0.34	0.22	0.30	0.25	0.23	0.32	0.27	0.17	0.23	0.20
Subtask2-Run3	0.61	0.90	0.73	0.49	0.70	0.58	0.25	0.12	0.16	0.18	0.09	0.12	0.19	0.09	0.13	0.14	0.07	0.09
Subtask1-Run1							0.77	0.69	0.73	0.56	0.50	0.53	0.60	0.54	0.57	0.44	0.40	0.42
Subtask1-Run2							0.76	0.71	0.73	0.57	0.53	0.55	0.59	0.56	0.57	0.45	0.42	0.43
Subtask1-Run3							0.71	0.75	0.73	0.52	0.55	0.53	0.54	0.58	0.56	0.39	0.42	0.41
Subtask1-Run4							0.69	0.78	0.73	0.52	0.59	0.55	0.53	0.60	0.56	0.40	0.46	0.43
Unoffical-Run1							0.77	0.68	0.72	0.57	0.52	0.55	0.60	0.53	0.56	0.44	0.41	0.43
Unoffical-Run2							0.76	0.69	0.72	0.56	0.53	0.54	0.59	0.54	0.56	0.44	0.41	0.42
Unoffical-Run3							0.73	0.78	0.75	0.54	0.58	0.56	0.51	0.55	0.53	0.38	0.41	0.40
Unoffical-Run4							0.45	0.03	0.06	0.27	0.02	0.03	0.31	0.02	0.04	0.20	0.01	0.03
Unoffical-Run5							0.29	0.03	0.06	0.18	0.02	0.04	0.20	0.03	0.04	0.12	0.02	0.03

Run 2: The same as Run 1 + The best run is kept the same and add new voting results.
Run 3: Reuse the BioREX model for relation extraction and BioRED to predict Novelty.
Run 4: Run 3 + Data augmentation + Finetuning on train/dev/test datasets. The novelty prediction uses *No-Novelty* approach for novelty prediction.
Run 5: The same as Run 4 but use the None-No-Novelty approach for novelty prediction.

References

1. C. H. Wei, H. Y. Kao, and Z. Lu, “PubTator: a web-based text mining tool for assisting biocuration.,” Nucleic Acids Res, vol. 41, no. Web Server issue, 2013, doi: 10.1093/nar/gkt441.

2. L. Luo, P. T. Lai, C. H. Wei, C. N. Arighi, and Z. Lu, “BioRED: A rich biomedical relation extraction dataset,” *Briefings in Bioinformatics*, vol. 23, no. 5. 2022. doi: 10.1093/bib/bbac282.

3. J. Li *et al.*, “BioCreative V CDR task corpus: a resource for chemical disease relation extraction,” *Database (Oxford)*, vol. 2016, 2016, doi: 10.1093/database/baw068.

4. M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, and T. Declerck, “The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions,” *J Biomed Inform*, vol. 46, no. 5, 2013, doi: 10.1016/j.jbi.2013.07.011.

5. M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, “Overview of the protein-protein interaction annotation extraction task of BioCreative II,” *Genome Biology*, vol. 9, no. SUPPL. 2. 2008. doi: 10.1186/gb-2008-9-s2-s4.

6. S. Pyysalo *et al.*, “BioInfer: A corpus for information extraction in the biomedical domain,” *BMC Bioinformatics*, vol. 8, 2007, doi: 10.1186/1471-2105-8-50.

7. J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, “GENIA corpus - A semantically annotated corpus for bio-textmining,” in *Bioinformatics*, 2003. doi: 10.1093/bioinformatics/btg1023.
8. P.-T. Lai, C.-H. Wei, L. Luo, Q. Chen, and Z. Lu, “BioREx: Improving biomedical relation extraction by leveraging heterogeneous datasets,” *J Biomed Inform*, vol. 146, 2023, doi: 10.1016/j.jbi.2023.104487.
9. L. Luo, C. H. Wei, P. T. Lai, R. Leaman, Q. Chen, and Z. Lu, “AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning,” Bioinformatics, vol. 39, no. 5, 2023, doi: 10.1093/bioinformatics/btad310.