Semantic Information Retrieval: Exploring Dependency and Word Embedding Features in Biomedical Information Retrieval

Majid Rastegar-Mojarad[1,2], Ravikumar Komandur Elayavilli[1], Yanshan Wang[1], Sijia Liu[1], Feichen Shen[1], Hongfang Liu[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[2]University of Wisconsin-Milwaukee, Milwaukee, WI, USA

***Abstract*—We present an information retrieval system developed to retrieve evidence sentences for a given Biological Expression Language (BEL) statement. Previously, as a participant of the BEL challenge in BioCreative V, we proposed and developed a system called BELTracker, which mainly focused on lexical features. In the BEL challenge of BioCreative VI, we explored usage of syntactic and semantic features in identifying and ranking evidence sentences. Similar to BELTracker, the new system has 3 main components: indexing, retrieval, and ranking. In this system, we trained several classifiers for the ranking process. These classifiers have lexical, syntactic, and semantic features such as unigrams, bigrams, word embedding, and dependency-word embedding. Our evaluation showed that the new system obtained higher mean average precision on BioCreative V's test data when compared to BELTracker under full, relaxed, and context criteria. The challenge organizers provided 100 BEL statements as test data and we submitted 2 set of results 1) baseline (using ElasticSearch results), and 2) using the classifiers to re-rank the ElasticSearch results. The baseline system achieved 30.6% and 45.8% precision under full and partial criteria, respectively. Using various classifiers for ranking, the precision of the system increased to 31.6% and 50.2%, respectively, for full and partial criteria.**

Keywords*—*Biological Expression Language; Dependency-word embedding; Information Retrieval; Semantic Information Retrieval; Sentence retrieval; Word embedding

I. Introduction

The aim of information retrieval (IR) systems is to identify relevant resources to users' queries. Most IR systems consider scientific literature as resource and return abstracts or snippets of relevant literature as results. For some users, like physicians at point-of-care, reviewing or going through these results is time consuming and not practical. The ideal results for these users should be brief and pointed, especially if the users are looking for any evidence of relation between entities. For example, if a physician is interested in finding any evidence of interaction between 2 drugs, the evidence would most likely be in a single sentence (or 2 consecutive sentences). There have been several attempts in the biomedical domain to implement such IR systems which allow users to enter entities and type of relation between them and retrieve evidence sentences (1–6). The main limitation of these systems is focusing on unary or binary relations. They are unable to retrieve relevant evidence for biological observations which contain more than 2 entities and relations. The organizers of BioCreative V introduced a new track, called Biological Expression Language (BEL), which addresses this limitation. The track had 2 main tasks: 1) extracting BEL statements from a given sentence, and 2) identifying relevant sentences for a given BEL statement (7,8). BEL is 1 of the main representatives of biologic networks. A BEL statement has several components: normalized entities, functions, relations, namespace, and sequence positions. Table I illustrated 2 sentences and 3 extracted BEL statements.

TABLE I. Examples of BEL statements

Sentence	We showed that HSF1 is phosphorylated by the protein kinase RSK2 in vitro. We demonstrate that RSK2 slightly represses activation of HSF1 in vivo
Extracted BEL	**1:** kin (p (HGNC: RPS6KA3)) increases p (HGNC: HSF1, pmod (P)) **2:** kin (p (HGNC: RPS6KA3)) decreases tscript (p (HGNC: HSF1))
Sentence	Exposure of neutrophils to LPS or TNF-α resulted in increased levels of the transcriptionally active serine 133-phosphorylated form of CREB
Extracted BEL	p (MGI: TNF) increases p (MGI: CREB1, pmod (P, S, 133))
BEL Elements: Relationship, Function, Entity, Namespace, Sequence position	

BioCreative VI includes BEL track as well, and our team participates in the track and tries to improve our previous system, called BELTracker (9), and this paper describes our new system. One of the limitations of BELTracker is relying upon lexical features and heuristic approaches to rank returned results. In the new system, we try to address these issues by training several classifiers for ranking process and using semantic features along the way. Similar to BELTracker, the system has 3 main components: indexing, retrieval, and ranking components. First, we index all informative sentences in MEDLINE abstracts. For a given BEL statement, the system constructs a query using all the elements in the BEL statement and retrieves the most relevant sentence from the index based on occurrence of the elements. Finally, using lexical and semantic features, the system ranks the results and returns the top 10 relevant sentences.

Herein we will discuss our method. Next, we will present the results of comparison of the systems and the performance of the new system on the test set, which was evaluated manually by the organizers. Lastly, we discuss the results and limitations of the system.

II. Methods

Our system has 3 main components: indexing, retrieval, and ranking.

A. Indexing Component

As the system aims to retrieve evidence sentences for BEL statements, we only store and index *informative sentences* from MEDLINE abstracts. We call a sentence *informative* if it contains at least 2 biomedical entities and a relationship between the entities. To identify these sentences, the system relies on Semantic Medline Database (SemMedDB (10)). SemMedDB is a relational database and stores all informative sentences, from MEDLINE abstracts, which are extracted by a rule-based system, as described by Kilicoglu et al (10). The indexing component retrieves all the sentences from SemMedDB and indexes them in a text search engine, called ElasticSearch. Unlike the other 2 components that run for each query (BEL statement), this component only runs once and prepares the index for the system.

In the previous system (BELTracker), we indexed abstracts from PubMed and full text articles available in PubMed Central. In the current index, we do not have full-text article sentences because SemMedDB does not cover full-text articles.

B. Retrieval Component

For a given BEL statement, the system first retrieves the most relevant sentences from the index, which is the responsibility of the second component, retrieval. The retrieval component identifies all the elements (Table I) in the given BEL and uses external and expert-generated resources to find their synonyms and then generates an appropriate ElasticSearch query. The retrieval component in the system is as the same as BELTracker (9), and the only difference is that we utilize a newer version of the resources.

C. Ranking Component

The retrieval component returns at most 1,000 relevant sentences to the given BEL statement. ElasticSearch retrieves these sentences based on appearance of the BEL's elements somewhere in the sentence, and it does not consider any semantic feature in the retrieval process. It is obvious that simply based on co-occurrence of the elements in a sentence, we can not conclude existence of relation between the elements. The third component of the system, the ranking component, investigates existence of relation between the elements and ranks the sentences based on their relevancy to the BEL statement. In order to rank the evidence sentences, 3 classifiers classify them based on existence of any relation between the elements. Here we describe these 3 classifiers and how the ranking component uses the classification results for ranking the sentences.

1. First Classifier: Entity-Entity Classifier

Instead of using co-occurrence of entities (of the BEL statement) in a sentence as the indicator of the existence of a relation between them, we propose to train a binary classifier, called *Entity-Entity* (EE) *classifier*. Each BEL statement (in the training and test data) has at least 2 entities and EE classifier is calculating the likelihood of relation (regardless of type of relation) between the BEL entities in the retrieved sentences by the retrieval component. The instances of EE classifier are sentences containing at least 2 entities; the positive instances showing relation between the entities and the negative instances otherwise. One of the challenges to train the EE classifier is the training data. The training data provided by the organizers contains only positive instances (6,000 sentences and 11,000 BEL statements extracted from them). In order to generate negative instances, we employ distant supervision technique and SemMedDB. There are sentences in SemMedDB, which contain 2 biomedical entities with *co-exist* relation type. This relation type indicates that the rule-based system was not able to identify any specific type of relation between the entities and they only co-occurred in the sentence. We utilized these sentences as negative instances for the EE classifier. In the training process, the name of entities in the sentences are masked and replaced with a general term such as *Entity*. As features for the EE classifier, unigrams, bigrams, and word embedding of terms between entities are used. Since word embedding contains semantic relationship information, adding word embedding to the feature list allows us to move beyond lexical features. We trained embedding on PubMed abstracts.

2. Second Classifier: Function-Entity Classifier

The second classifier aims to calculate the probability of relation between functions and entities in the retrieved sentences. For example, if the given BEL statement is:

cat(HGNC:XIAP) decreases cat(HGNC:CASP9)

there are 2 function-entity relations: *cat-XIAP* and *cat-CASP9*. This classifier, called *Function-Entity* (FE) *classifier*, examines existence of both pairs in the retrieved sentences and calculates the probability of each relation in the sentences. For each function that appeared more than 100 times in the training data, we build a binary classifier.

In order to train FE classifiers, both positive and negative instances are available in the training data. Unigrams and bigrams of surrounding words of the entity (window 3-5) are utilized as the features for FE classifiers. Beside lexical features, we evaluated using word embedding, dependency-based word embedding (11), and abstract meaning representation (AMR) embedding (12) as other features for FE classifiers. In the following sentence, we illustrate how we generate dependency contexts to train dependency-based word embedding. A similar approach was used to generate AMR

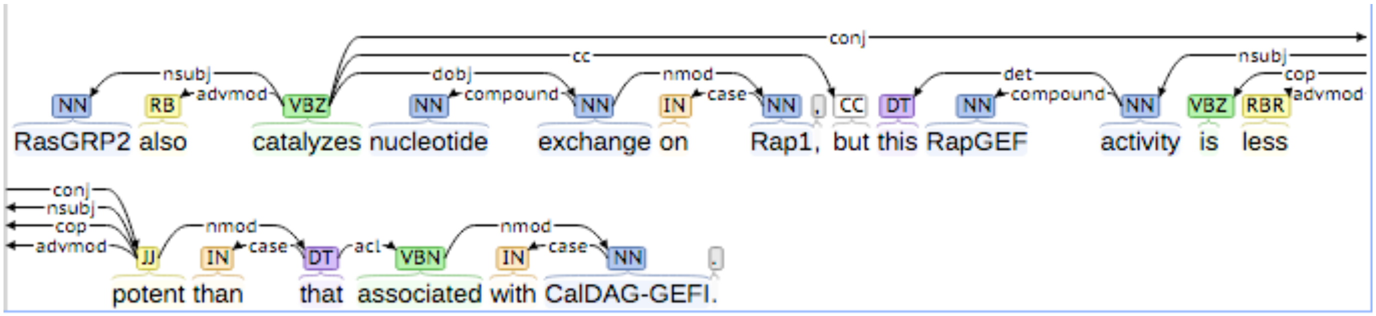
conj
nsubj
advmod
cc
dobj
compound
nmod
case
det
compound
nsubj
cop
advmod
NN RB VBZ NN NN IN NN , CC DT NN NN VBZ RBR
RasGRP2 also catalyzes nucleotide exchange on Rap1, but this RapGEF activity is less
conj
nsubj
cop
advmod
nmod
case
acl
nmod
case
JJ IN DT VBN IN NN .
potent than that associated with CalDAG-GEFI.

Figure I: Dependency tree for this sentence:
"RasGRP2 also catalyzes nucleotide exchange on Rap1, but this RapGEF activity is less potent than that associated with CalDAG-GEFI" PMID:10918068

embedding from AMR graphs shown in Wang et al (12). Consider this sentence:

"RapGEF activity is less potent than that associated with CalDAG-GEFI."

We generate dependency tree for the sentence(Figure I). Then for each term, we produce dependency contexts for all incoming and outgoing edges. Dependency context for each edge is a combination of *governor of edge*, *direction of edge (inv for incoming edges)*, and *the term on the other side of edge*. Table II contains dependency contexts for first 3 terms in the sentence.

TABLE II. DEPENDENCY CONTEXTS

	Terms		
	RasGRP2	***Also***	***Catalyzes***
Dependency Contexts	1)nsubj_inv_catalyzes	1)advmod_inv_catalyzes	1)nsubj_RasGPR2 2)advmod_also 3)conj_potent 4)cc_but 5)dobj_exchange

3. *Third Classifier: Relation Classifier*

There are 2 types of relationship in the training data: *increase* or *decrease*. We train another binary classifier to categorize the retrieved sentences based on type of relationship. The training data contains both positive and negative instances for this classifier. More details about these classifiers, such as feature list, are explained in our previous work (9).

4. *Final score calculation*

After obtaining probabilities (from the classifiers) for each retrieved sentence, we aggregate the results and calculate a score for each sentence. This score is used for ranking the sentences and selecting top relevant evidence sentences. The score is calculated as follows:

$$\text{Score}_{\text{sentence}} = W_{EE} * P_{EE} + W_{FE} * P_{FE} + W_{\text{relation}} * P_{\text{relation}}$$

P represent the probability produced by the classifiers and *W* indicate the weights assigned to each classifier. As we do not have appropriate data to learn each classifier's weight, we assign weights to each classifier based on importance of each element and training data of the classifiers. The weights are: $W_{EE} = 0.4$, $W_{FE} = 0.5$, and $W_{\text{relation}} = 0.1$. The FE classifier has the highest weight because the data used to train this classifier has less noise compared to the EE classifier. The relation classifier has a low weight, because we observe that rarely 2 sentences contain all elements of a BEL statement but convey 2 different relationship types. Meaning, if entities and functions of a given BEL statement appear in a sentence, most likely the sentence has the relationship type mentioned in the BEL statement.

III. RESULTS AND DISCUSSION

Using the test set of BioCreative V, we compared mean average precision (MAP) of new and previous systems. Table III shows MAP of the systems and 3 different scenarios (worst, random, best) (scenarios and criteria are described in (9)).

TABLE III. COMPARING MAP OF BOTH SYSTEMS AND 3 SCENARIOS

Criteria	**Systems and scenarios**				
	Worst	***Random***	***Previous System***	***New System***	***Best***
Full	31.7	46.5	49.0	56.96	74.2
Relaxed	45.9	58.4	62.1	65.05	80.4
Context	55.2	65.7	68.9	73.15	83.5

The results in Table III show the new system obtains higher MAP compared to the previous system; however, there is still a lot of room for improvement (comparing to the best possible MAP for BioCreative V test set). Unfortunately, we are not able to compare precision of the systems, because the evaluation should be done by domain expert, who we do not have access to.

TABLE IV. PRECISION OF EACH RUN

Systems	**Criteria**	
	Full	***Partial***
Baseline	30.6	45.8
Using the classifiers for ranking	31.6	50.2

For BioCreative VI, the organizers provided 100 BEL statements as a test set and asked the participants to return up to 5 relevant evidence sentences for each BEL statement. We submitted 2 sets of results. In the first run (baseline), we ranked the sentences based on ElasticSearch score (in fact, we did not engage the ranking component). In the second run, the sentences are ranked using the ranking component. Table IV shows precision of each run for *full* and *partial* criteria. The results show that semantic features can improve the performance of the system. Using semantic features, the system gained 1% and 5% in precision for full and partial criteria, respectively.

There are several limitations in this work. The main limitation is the absence of manually generated training data for the classifiers. The classifiers are training by the dataset which the entities are not annotated manually (keyword search is used to detect and annotate the entities). Using the classifiers is the novelty of this work, but lack of having comprehensive and clean training data forced us to simplify the relations in BEL statements. For example, EE classifiers do not consider relations between more than 2 entities (which is possible in nested BEL statements). As mentioned before, the system is only search sentences in PubMed abstracts, because SemMedDB does not cover full text articles. This leads to another limitation that the system misses evidence sentences which are not in the abstracts.

In the future, we will focus our work on generating cleaner training data for the classifiers and adding sentences from full-text articles available in PubMed Central to the index.

IV. Conclusion

As a participant in BEL track of BioCreative VI, we implemented an information retrieval system to retrieve evidence sentences for BEL statements. Compared to our previous system, BELTracker, we tried to use semantic features such as: word embedding, dependency-word embedding, and AMR embedding in the ranking process. The results showed that these semantic features can improve performance.

References

1. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics. 2005 Jan 1;21(suppl 2):ii252-ii258.
2. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics. 2004 Oct 8;5:147.
3. Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio—the analysis and navigation of molecular networks. Bioinformatics. 2003 Nov 1;19(16):2155–7.
4. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W783-786.
5. Ohta T, Matsuzaki T, Okazaki N, Miwa M, Sætre R, Pyysalo S, et al. Medie and Info-pubmed: 2010 update. BMC Bioinformatics. 2010 Oct 6;11(Suppl 5):P7.
6. Ravikumar KE, Wagholikar KB, Li D, Kocher J-P, Liu H. Text mining facilitates database curation - extraction of mutation-disease associations from Bio-medical literature. BMC Bioinformatics. 2015.
7. Fluck J, Madan S, Ellendorff TR, Mevissen T, Clematide S, Lek A van der, et al. Track 4 Overview: Extraction of Causal Network Informationin Biological Expression Language (BEL). In 2015.
8. Fluck J, Madan S, Ansari S, Kodamullil AT, Karki R, Rastegar-Mojarad M, et al. Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL). Database (Oxford) 2016 Aug 20 [cited 2017 Sep 13];2016.
9. Rastegar-Mojarad M, Komandur Elayavilli R, Liu H. BELTracker: evidence sentence retrieval for BEL statements. Database (Oxford). 2016 May 12.
10. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012 Dec 1;28(23):3158–60.
11. Levy O, Goldberg Y. Dependencybased word embeddings. In: In ACL. 2014.
12. Wang Y, Liu S, Rastegar-Mojarad M, Wang L, Shen F, Liu F, et al. Dependency and AMR Embeddings for Drug-Drug Interaction Extraction from Biomedical Literature. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics. ACM; 2017. p. 36–43.