Extended dependency graph for BioC-compatible protein-protein interaction (PPI) passage detection in full-text articles

Yifan Peng[1], Cecilia Arighi[1,2], Cathy H. Wu[1,2], and K.Vijay-Shanker[1]

[1]Computer & Information Sciences
[2]Center for Bioinformatics & Computational Biology
University of Delaware, Newark, DE 19716, USA
{yfpeng,arighi,wuc,vijay}@udel.edu

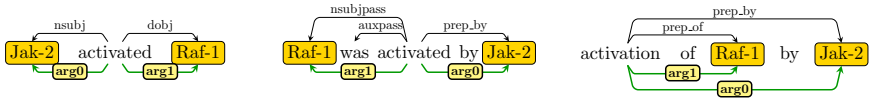
Abstract. Protein-protein interaction (PPI) is important in the field of experimental biology as well as bioinformatics. In BioCreative V, we participated in the BioC task and developed a PPI system to detect passages with PPIs in the full-text articles. By adopting the BioC format, the output of the system could be seamlessly added to the biocuration tool with little effort required for the system integration. Our PPI system utilizes Extended Dependency Graph as an intermediate level of representation to abstract away syntactic variations in the sentence. As a result, we only use three basic predicate-argument rules to extract PPI pairs in the sentences, and two additional rules to detect additional passages with PPI pairs. Experiments on 20 in-house full-text articles show that we are able to obtain a recall of 77.8. By using only three basic rules, experiments on AIMed further confirm that we can achieve a precision of 91.5 of sentence selection and an F-value of 62.8 of instance selection.

Key words: protein-protein interaction, relation extraction, text mining, BioC

1 Introduction

Protein-protein interaction (PPI) extraction detects statements of physical interactions between proteins from biomedical literature. Many efforts have contributed to different aspects of PPI in the bio-text mining community; from PPI document classification, to PPI or PPI method detection [1, 6, 9, 14, 15]. In particular, the BioCreative V BioC task (Track 1) proposes to build a framework that allows different text mining tools to be seamlessly integrated into a pipeline for literature curation of protein interactions (both genetic and physical interactions) to be evaluated by BioGRID database curators [5]. Our team participated in this task by contributing in detecting passages with PPIs over full-text articles.

Full-text articles by nature use various syntactic constructions. These textual variations can be problematic for PPI systems to account for. A central theme to this work is the hypothesis that the varied forms of PPI mentions are essentially



(a) (b) (c)

Fig. 1: Sample EDG with an active (a), passive (b), or normalized (c) verb.

due to certain syntactic structural complexities. By capturing regularities underlying these complexities, we can build a system where the extraction patterns are kept simple.

We have recently proposed a novel text representation, the Extended Dependency Graph (EDG) that abstracts away certain text variations [11]. EDG not only considers syntactic dependencies between words in a sentence, but also utilizes information beyond syntax to capture different dependencies. In particular, EDG adds numbered arguments in the dependency graph to provide consistent argument labels across different textual forms. For example, Fig. 1 shows EDGs of three text fragments with syntactic edges above the words and numbered argument edges below. The numbered argument edges, *arg0* and *arg1*, unify the realization of active, passive and nominalized forms of the verb "activate" for purposes of PPI detection.

In the BioCreative V BioC task, one contribution of our project is using EDG to extend the framework for fast development of pattern-based biomedical relation extraction. This intuition is partially based on our previous work that leverages syntactic variations in a language to achieve high precision [12], as well as the work that applies sentence simplification to improve the coverage of extracted relations [13]. EDG allows us to use only three sets of basic rules to detect PPI pairs.

Another contribution is proposing a set of task-specific rules to boost recall of PPI passage detection. In PPI mention detection, the system needs to identify whether a given protein pair in a sentence has PPI relationship or not. But in the BioC task, we aim to identify passages mentioning PPIs, and do not need to extract specific PPI instances. By exploiting this difference, we propose a set of task-specific rules based on the human PPI annotation samples. After detecting PPI instances using basic rules, the additional rules extend the boundary of passages containing PPI, and detect more passages over the full-text article by utilizing the detected PPI mentions.

We conducted two experiments to test the system. First, one of the authors (CA) annotated 20 full-text articles. The annotations mark (1) the passages in abstracts and result sections with PPIs that are newly discovered in the article, and (2) all unique PPIs in the full-text based on BioGRID. Experiments on these 20 in-house full-text articles show that we are able to obtain a good PPI extraction system with a recall of 77.4. Second, we evaluated the system on AIMed [3]. We obtained an F-value of 75.4 at a precision 91.5 for the detection of sentences with PPI.

Summarizing our participation in the BioCreative V BioC task, we have (1) developed a module to find passages with physical PPIs using BioC format [7],

(2) applied the module for full-text articles, and (3) integrated the module into the collaborative framework for the BioGRID annotation system. The input of our module are BioC documents with gene/protein named entity marked and the output is to add BioC annotation indicating which sentence or block of sentences contains PPIs. The final whole system proves that our system can be seamlessly added to the biocuration tool with little effort required for the integration.

2 Methods

2.1 Extended dependency graph

In this paper, we use EDG to represent the structure of the sentence [12]. The vertices in an EDG are labelled with information such as the text, part-of-speech, and lemma. If an entity mention spans multiple tokens in a sentence, we merge their corresponding vertices into one vertex.

EDG has two types of dependencies. The syntactic dependencies are obtained from CCProcessed dependencies output by applying Stanford dependencies converter [8] on a parse tree obtained by Bllip parser [4]. The other type of dependencies is the numbered arguments, whose idea is based on the guidelines of PropBank [2]. For the PPI detection task, we use only *arg0* and *arg1* in EDG.

To create *arg0* and *arg1* in EDG, we use different syntactic dependencies obtained from the Stanford typed dependencies. We also detect non-syntactic relations such as *is-a*, *member-collection*, and *part-whole* and propagate *arg0* and *arg1* using these relations. More details of EDG construction can be found at [12].

2.2 Basic predicate-argument rules

EDG abstracts away the syntactic variations in the sentence. Thus by using EDG, the number of rules to extract PPIs are greatly reduced. In our system, only three sets of rules are used. The list of different trigger words can be found at `http://www.eecis.udel.edu/~ypeng/bc5bioc/support_materials.html`.

1. Protein $\xleftarrow{arg0}$ PPI predicate trigger $\xrightarrow{arg1}$ Protein
2. Protein $\xleftarrow{arg0}$ PPI noun trigger $\xrightarrow{arg0}$ Protein
3. Protein $\xleftarrow{prep\_of}$ process trigger $\xleftarrow{arg0}$ indirect trigger $\xrightarrow{arg1}$ Protein

Rule 1 is a set of most basic and strict rules. We use PPI triggers (e.g., associate" and "bind") and post-transcriptional modification triggers (e.g., "acetylate" and "methylate") in the system. Because EDG has unified different forms of predicates in the vertices, we only need to list lemmas of triggers in the rules.

Rule 2 accounts for triggers that are not derived from verbs (e.g., "complex"). This rule matches the noun phrase such as "$[\text{X}]_{\text{protein}}$-$[\text{X}]_{\text{protein}}$ complex".

Rule 3 accounts for indirect PPI triggers such as "block" and "mediate". Indirect trigger often indicates a biological process whose arguments are not proteins but some activities of proteins. In our system, the process triggers include "activity" and nominalization of PPI triggers whose suffixes are "-ion".

To match the basic rules to EDG, we use the subgraph-matching algorithm [10]. For each rule, a subgraph is constructed. Both nodes and edges in the subgraph are predicates of EDG nodes and edges. The worse-case complexity of the subgraph-matching algorithm is $O(n^2k^n)$ where n is the number of vertices in EDG and k is the vertex degree.

2.3 Non predicate-argument rules to increase the recall

So far, we have discussed EDG with basic rules to detect PPI interacting partners, and then select the corresponding sentences. To conform to the BioCreative V BioC task, we felt that other sentences that contain the detected protein pairs might also be the interest of biocurators. For example, if we know that "CP" and "Rpt1" interact elsewhere in PMID 19412160, then we would like to pick the sentences such as "In contrast, another group of Rpts, including Rpt1, Rpt2, and Rpt5, is proposed to initially assemble free of the CP in the BP1 complex" in the same document. However, the basic rules in previous section are insufficient in this case. To pick such sentences, the following two rules are applied when we find two proteins are known to interact somewhere in the document. It is noteworthy these rules only boost the recall of passages detection.

Experimental techniques with 2 proteins. To identify new PPIs, experiments are conducted and described in the paper. Such description will be captured by our system when both the experimental technique and two proteins are mentioned in the same sentence. In our system, we used 5 technique keywords: "2-hybrid", "BIFC", "cosedimentiation", "ITC", and "pulldown".

Extension with PPI triggers and 2 proteins. In some complicated sentences, the PPI triggers and two proteins are mentioned but there is no direct edge between trigger nodes and proteins in EDG. This is especially true when this sentence ($S1$) is followed by another sentence ($S2$) where the interaction between these two proteins has already been detected. The hypothesis is that the block of sentences is a continuation of the same topic. In such case, we combined $S1$ and $S1$ into one passage.

3 Evaluation and analysis

In the BioC task, 120 articles are provided for the annotation. Since there is no gold annotation, we randomly chose 20 articles as the test set. One of the authors (CA), who is an experienced biocurator, annotated the abstract and result sections. We chose result sections because they describe experimental results about PPI events. We also got the PPI information for these 20 articles from the BioGRID database. Note, BioGRID only marks PPIs in the articles, but does not identify the passages. Table 1 shows the recall of PPI passage detection. We did not report the precision because the curator only annotated part of the document and only identified the PPIs that are new in the document. Over the total 120 full-text articles, our system extracted 15,529 passages with PPI,

Table 1: Recall on 20 in-house documents (only abstract and result sections).

Section	Passages with PPI	Unique PPI by BioGRID
Abstract	80.0	–
Results	77.6	–
Total	77.8	74.1

among which 78.7% are from abstracts and result sections. The counts confirmed that both sections are most important to describe PPIs.

In the second experiment, we applied the relation extraction system on the AIMed corpus [3], which is commonly used in PPI extraction task and has been suggested by the task organizers as a training set (for machine learning systems). Table 2 reports two sets of performance metrics based on how we compare the system annotation with the gold standard. Both results were obtained by using only the basic rules (Section 2.2). The first row shows the performance of selecting sentences with PPI. We conducted this experiment because the BioC task is for PPI passage detection rather than the PPI instance detection. Likewise, we modified the AIMed annotations to indicate which sentence has the PPI. For the sentence selection task, we achieved a high precision of 91.5. The second row shows the performance of detecting PPI pairs, which is the traditional PPI extraction task, thus our results are comparable with other works. We obtained F-value of 62.8. It is noteworthy that we achieved these results by using just three basic rules, and the results are comparable with, although slightly lower than, those of the start-of-the-art systems. We believe that this suggests the advantages brought out by using EDG. In future, we will include more rules to improve the performance.

Table 2: Evaluation results on AIMed.

	TP	FP	FN	Prec.	Recall	F-value
Sentence	367	34	206	91.5	64.1	75.4
Instance	552	205	448	72.9	55.2	62.8

4 Conclusion

In BioCreative V BioC task, we developed a PPI system to detect passages with PPIs in the full-text articles. By using the BioC format, the output of the system could be seamlessly added to the biocuration tool with little effort. In addition, we evaluated the PPI system on a set of 20 documents, for passages with PPIs, as well as the widely used AIMed corpus. Both experiments confirm that the system is able to achieve high recall for PPI passage detection ($\sim$82 on 20 documents) and high precision ($\sim$90 on AIMed).

The development of our system is based on the semantic dependencies between entities that is critical for either pattern-based or machine learning systems. We believe this information is not task-dependent and an enhanced understanding will contribute to developing systems for various relation extraction tasks, such as genetic interactions defined in BioGRID in this track.

Acknowledgments. Research reported in this manuscript is supported by the National Science Foundation under Grant No. DBI-1062520.

References

1. Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., Salakoski, T.: All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC bioinformatics 9(Suppl 11), S2 (2008)
2. Bonial, C., Hwang, J., Bonn, J., Conger, K., Babko-Malaya, O., Palmer, M.: English PropBank Annotation Guidelines. Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder (2012)
3. Bunescu, R.C., Ge, R., Kate, R.J., Marcotte, E.M., Mooney, R.J., Ramani, A.K., Wong, Y.W.: Comparative experiments on learning information extractors for proteins and their interactions. Artificial Intelligence in Medicine 33(2), 139–155 (2005)
4. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: ACL. pp. 173–180 (2005)
5. Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., ODonnell, L., et al.: The biogrid interaction database: 2013 update. Nucleic acids research 41(D1), D816–D823 (2013)
6. Chowdhury, F.M., Lavelli, A., Moschitti, A.: A study on dependency tree kernels for automatic extraction of protein-protein interaction. In: BioNLP workshop. pp. 124–133 (2011)
7. Comeau, D.C., Doğan, R.I., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Fabio Rinaldi, M.T., Valencia, A., Verspoor, K., Wiegers, T.C., Wu, C.H., Wilbur, W.J.: BioC: A minimalist approach to interoperability for biomedical text processing. Database 2013, 1–15 (2013)
8. De Marneffe, M.C., Manning, C.D.: Stanford typed dependencies manual. Stanford University (April 2015)
9. Erkan, G., Özgür, A., Radev, D.R.: Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In: EMNLP-CoNLL. vol. 7, pp. 228–237 (2007)
10. Liu, H., Keselj, V., Blouin, C., Verspoor, K.: Subgraph matching-based literature mining for biomedical relations and events. In: AAAI Fall Symposium Series (2012)
11. Peng, Y., Gupta, S., Wu, C.H., Vijay-Shanker, K.: An extended dependency graph for relation extraction in biomedical texts. In: BioNLP workshop. p. 21 (2015)
12. Peng, Y., Torii, M., Wu, C.H., Vijay-Shanker, K.: An NLP framework for fast development of pattern-based biomedical relation extraction systems. BMC bioinformatics 15, 285 (2014)
13. Peng, Y., Tudor, C.O., Torii, M., Wu, C.H., Vijay-Shanker, K.: iSimp: A sentence simplification system for biomedical text. In: BIBM. pp. 211–216 (2012)
14. Zhang, M., Zhang, J., Su, J., Zhou, G.: A composite kernel to extract relations between entities with both flat and structured features. In: CICLing-ACL. pp. 825–832 (2006)
15. Zhou, G., Zhang, M., Hong, D., Zhu, J.Q.: Tree kernel-based relation extraction with context-sensitive structured parse tree information. In: EMNLP-CoNLL. pp. 728–736 (2007)