Review of Biomedical Relation Extraction

Stanley Chika ONYE [1], Arif AKKELEŞ [2] and Nazife DIMILILER [3]

[1,2] Eastern Mediterranean University, Department of Mathematics,

Famagusta North Cyprus via Mersin 10, Turkey

[3] Eastern Mediterranean University, School of Computing and Technology,

Famagusta North Cyprus via Mersin 10, Turkey

E-mail: stanley.onye@emu.edu.tr [1], arif.akkeles@emu.edu.tr [2] and nazife.dimililer@emu.edu.tr [3]

Corresponding Author:

Stanley Chika ONYE

CT106, School of Computing and Technology,

Eastern Mediterranean University,

Famagusta North Cyprus via Mersin 10, 99628, Turkey

E-mail: stanley.onye@emu.edu.tr

Abstract

Relation extraction task as a part of information extraction in biomedical domain has been reviewed in this paper. This paper covers an overview on some of the currently available biomedical corpora used for relation extraction, it also presents a review on some of the techniques used in the domain. We discuss evaluation techniques used for relation extraction systems as well as the relation extraction methods under two categories: supervised learning and unsupervised learning. The most prevalently used approaches related to these supervised and unsupervised classification methods are also discussed.

Keywords: Relation extraction, information extraction.

Introduction

The amount of text predominately in an unstructured and consistently changing or varying format is vastly increasing dailyand is also available in different platforms. Therefore, the need to make these data more understandable and useful to humans by adding semantic information becomes paramount. However, (Bach & Badaskar,2007) recorded that the sheer volume and heterogeneity of the data renders manual extraction of valuable information almost impossible. Therefore, the task of extracting useful information from these text data must be automated. One of the most preliminary information extraction tasks, Named Entity Recognition (NER), have achieved an acceptable degree of success and some state-of-the-art Named Entity (NE) recognizers such as (Bikel, Schwartz, & Weischedel, 1999) and (Finkel, Finkel, Grenager, & Manning, 2005), have been able to add NE labels to data automatically with high accuracy. Relation Extraction (RE) is the process that follows the NE identification and it aims at discovering relations between named entities such as organization, location, person, gene, chemical, disease etc. Examples of these relations are chemical-disease, organization-location. In this paper, we will focus on RE an area of Information Extraction (IE) and RE on biomedical domain which deals with obtaining structured and detailed relation between entities from biomedical related text.

1.1. Relation Extraction in Biomedical Domain

Information Extraction (IE) is a field of computational linguistics which plays a significant role in efficient data management and can be defined as a process of getting structural data from unstructured text data. IE consists of many processes/steps in which different types of information are retrieved in each of those steps. RE is a step in IE which normally comes after named entity recognition and co-reference resolution. Relation extraction by Culotta, McCallum, and Betz (2006, p2) is defined as “the task of discovering semantic connections between entities. In text, this usually amounts to examining pairs of entities in a document and determining (from local language cues) whether a relation exists between them.”The major focus of biomedical research currently, shifted from the individual biological entities (e.g. chemicals, diseases, genes or proteins) to the whole biological systems, thereby, making the demand for extracting relationships between biological entities (e.g. chemical-diseases, drug-drug interactions) from biomedical text for knowledge discovery and to produce scientific hypotheses increasingly imperative(Chapman & Cohen, 2009; Zweigenbaum, Demner-Fushman, Yu, & Cohen, 2007). The discovery of automatic relation extraction offers an interesting solution to the problem of manually transforming this information from unstructured text into a structured form as it reduces the time spent by researchers on reviewing the literature and enabling them to scan significant number of scientific articles rapidly (Erhardt, Schneider, & Blaschke, 2006).In recent years, a lot of approaches with differences in their techniques have been proposed (Zhou and He, 2008).Generally, a relation extraction system consists of mainly three modules, namely text preprocessing, parsing, and relation extraction as shown in Fig. 1.

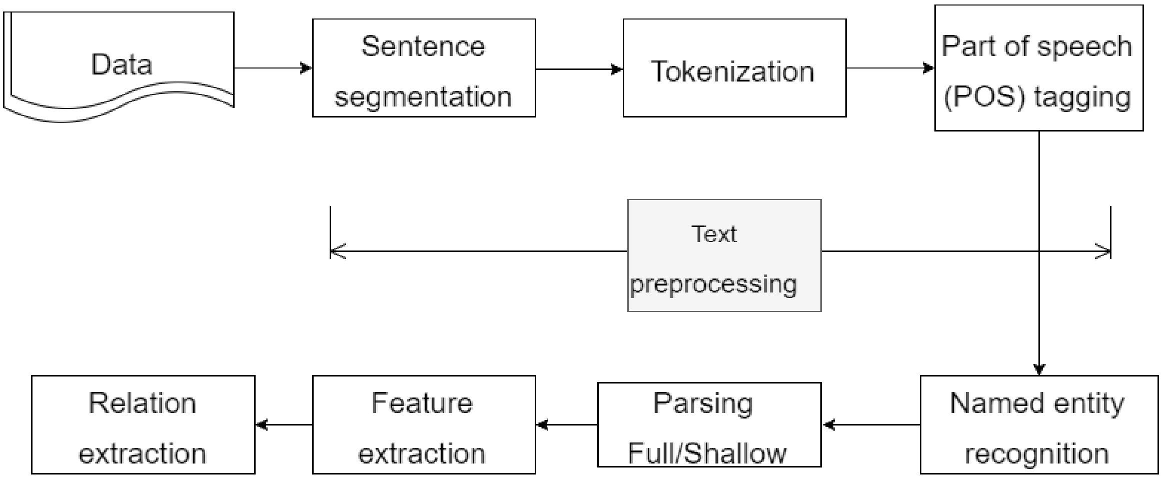

Figure 1: A basic relation extraction system.

1.1.1. Text Preprocessing.

Text preprocessing is divided into several basic text modules for subsequent processing. Presently, relation extraction are sentence-based ensuring that larger block of texts like abstracts, whole documents, etc. are split into single sentences which are further split into tokens which serves as inputs used in the subsequent steps. Sentence splitting and tokenization in biochemical domain has proven to be very complicated due to the extensive and unconventional use of punctuation marks in the terminology of the domain (Onye, Dimililer, & Akkeleş, 2016).Part of Speech (POS) tagging is the process of assigning one of the parts of speech (e.g. include nouns, verbs, adverbs, adjectives, etc.) to the given token. The aim of biomedical NER is to identify the biomedical entities (e.g. names of chemicals, drugs, genes and proteins) that are mentioned in the text. If the text is a sentence, it is called a sentence-based RE. NER is a prerequisite step for any relation extraction systems (Bikel, et al. 1999). Also, due to the complexities of the biomedical terminology in biomedical text arising from the non-standardized names, recognizing named entities is seen as a daunting task (Ohta, Pyysalo, Kim, & Tsujii, 2010).

1.1.2. Parsing. Parsing is the process of analyzing a given sentence by individually determining the structures of every token from their constituent parts. Based on the parsing results, related features pertaining to a sentence can be extracted and further used to enhance the process of extracting a relation between biomedical entities present in the text. A parsing process comprises of two components: these are a parser and a grammar. The grammar is a declarative component while the parser which is just a computer program is a procedural one. Irrespective of the language being used, the parser remains unchanged but the grammar changes based on the language. Therefore, a system can parse different languages just by changing the grammar. But the grammar and parser both depend on the grammar formalism (Robin, 2015).

1.1.3. Relation Extraction. This is the center of a relation extracting system and there presently exists numerous systems for extracting relations from texts and the series of methods available for dealing with this problem. The extraction methods used in biomedical RE can be classified into three: knowledge-based, supervised and self-supervised methods (Etzioni, Banko, Soderland, & Weld, 2008). Also, the techniques used in relation extraction systems can be divided into four groups, namely: rule-based, co-occurrence, pattern-based, and machine learning approaches (Bùi, 2012).These are discussed in section3.

2. Biomedical Corpora

The abundance of textual data available in biomedical domain has promoted the RE research in this field. The biomedical corpora play important roles in the evolution of relation extraction techniques, by providing data to incorporate or retrain available Natural Language Processing (NLP) tools to work with biomedical texts, to train ML-based relation extraction approaches, and to expedite automatic performance evaluation of relation extraction systems. The number of biomedical corpora presently available is small because developing annotated corpora is both a time-consuming and error prone quest. There exists some available corpora in biomedical research and can be accessed here: (Corpora for Biomedical Natural Language Processing, 2010; The WBI corpus repository, 2011).

2.1. BioCreative Corpora

BioCreative is a challenge driven organization providing tasks and inviting teams of biocurators to participate in them. The increasing number of groups working in the research area of text mining is a key motivation for this organization. There are no generally accepted standards or shared evaluation criteria to enable comparison among the different approaches that biocurators implemented, despite the increased activity in this area. The main emphasis of this organization is on the comparison of various implemented methods and the community assessment of scientific progress, rather than on the purely competitive aspects. From 2004 (BioCreative I) when the first challenge was introduced through 2016 (BioCreative 16), different corpora have been created for every challenge The BioCreative V challenge probably has the largest annotated corpus of all the BioCreative challenges with about 1,500 abstracts evenly divided into training, development and testing data sets. These individual challenges were categorized into tasks and possibly subtasks. Some of these tasks are Gene Normalization (GE), Gene Mention (GM) tagging, Gene Ontology, Protein-Protein Interaction (PPI), Chemical-Disease Relation (CDR) to namea few. Therefore, when compared to other corpora, this corpora is the most diversified in terms of the different type of tasks the individual corpus have been created to tackle. For example when compared with the PPI corpora consisting of corpora from five different domains with task mainly discovering relations between interacting proteins pairs in the biomedical text we can see that the corpora in BioCreative handle different types of tasks. Also, aside from BioCreative 2016 challenge, which had no official training nor testing sets, the other BioCreative challenges had datasets designed for their assigned tasks. More detailed information on these challenges and their corpora can be found here (BioCreative, 2008).

2.2. GENIA Corpus

The GENIA version 3.0 has probably one of the largest publicly available annotated corpus in the biomedical domain. It consists of about 2,000 Medline abstracts with almost 100,000 annotations for biological terms which were hand-coded, almost 20,000 sentences and more than 400,000 words (Kim, Yoon, & Yang, 2008). Additionally, this corpus contains annotation for linguistic structures such as part-of-speech and syntactic structures. This corpus has also been used to retrain NLP tools such as POS taggers, tokenizers, and full parsers to work with biomedical text. The GENIA corpus is freely available at (The WBI corpus repository, 2011).

2.3. PPI Corpora

The PPI corpora comprises of five different and independent corpora created for different purposes. They are AIMed, BioInfer, LLL, IEPA, and HPRD50. These corpora contain annotation for named entities and PPI pairs, but they vary in various aspects such as size (e.g. number of abstracts ranging from 50 to 200), biological coverage (i.e. how they are retrieved from PubMed by use of different keywords) and annotation policy (e.g. direction of interactions, interaction type). In (Pyysalo et al. 2008), they transformed the various PPI corpora into an XML-based format and proposed the use of only undirected and un-typed PPI pairs from these corpora for evaluation purposes in order to reduce the differences between these PPI corpora. Performances between PPI extraction methods have become easier with the application of these unified PPI corpora. Airola, et al., (2008) proposed some new evaluation methods called cross-learning and cross-corpora based on these PPI corpora. In cross-learning, when four corpora are used for training and the fifth one is used for testing, while in cross-corpora, one corpus is used for training and the other four corpora are used for testing. These evaluations help to properly reveal the ability of an extraction method adapting to any new text with relatively unknown characteristics.

3. Methods

Some of the various methods and systems deployed in the extraction of desired relations from biomedical texts are presented here.

3.1. Supervised methods

Supervised methods is the most frequently implemented machine learning in practice. Supervised learning relies on having a training set already prepared and tagged to serve as domain-specific examples. This method requires the creation of a suitable tagged corpus for training and an untagged one for testing purposes. Supervised learning is a time consuming task and efficiency hugely depends on a lot of factors such as the irregularity in proteins or genes names(e.g. ActA, E. coli), abbreviations (e.g. Dr., VA), inline citations, punctuations (e.g. L-dopa, 1-[2-(3,4-dichlorophenyl)ethyl]-4-methylpiperazine), token or NE extraction and feature extractions. The retrieval of complex biomedical or chemical entity names demands special attention as the process varies from one corpus to another since there is currently no general consensus regarding NE annotation in the biomedical community. This also makes it very difficult to compare the existing systems due to corpus incompatibilities.

In supervised learning input variables are given and output variables and then an algorithm is applied to learn the mapping function from the input into the output.The aim here is to predict the output variables of a data from the input data based on a well approximated mapping function. Also, supervised learning can be grouped into two problems:

Regression: This is when the output variable has a real value, such as “calories” or “weight” and “height”.

Classification: Here the output variable is a category, such as “white” or “black”, “disease” or “no disease” and “buyer” or “seller”.

Based on the nature of input for classification, supervised learning can be divided into kernel and feature based methods. Also, classification problems uses Machine Learning (ML) techniques which will be discussed later in this section. This is the act of using extracted features to serve as cues to the system in making decisions as to which label a sample can be classed into. Both semantic and synthetic features can be extracted from a given text. Syntactic features extracted from text (abstract and or sentence) can include the (1) tokens (each word in a sentence), (2) Bag of Words (BOW), (3) sentence length, (4) entity types, (5) number of words between the two entities, (6) number of verbs in the sentence, (7) frequency of entity mention etc. The path between the two entities in the dependency parse tree gives semantic cues. It is possible to use only syntactic features for RE as done by Onye et al., (2016) when they designed a model that used only syntactic features extracted from the corpus they worked. These extracted features are presented to the classifier in the form of a feature vector, for training and testing. According to Dimililer (2010), the number of features used in supervised learning increases the search space therefore, it is advisable to select features that will act as good indicators of existing entity relations during RE as not all features are going to be useful; the subset of useful features can be selected empirically based on experiments or an optimization technique can be applied in order to get the subset of available features that gives the optimal performance in terms of a performance metric such as the overall F-score. In order to generate these features from biomedical text, NLP tools can be used. There are available tools applied in NLP for text preprocessing. The most frequently used tools for biomedical text mining applications are listed in Table 1.

Table 1: A list of common NLP tools for biomedical text (Bùi, 2012).

Category	Name	URL
Sentence splitter	Lingpipe	http://alias-i.com/lingpipe
	Enju	http://www.nactem.ac.uk/y-matsu/geniass
Tokenization	OpenNLP	http://opennlp.apache.org
	Stanford	http://nlp.stanford.edu/software/tokenizer.shtml
	Lingpipe	http://alias-i.com/lingpipe
POS tagger	Enju	http://www.nactem.ac.uk/GENIA/tagger/
	OpenNLP	http://opennlp.apache.org/
	Stanford	http://nlp.stanford.edu/software/tokenizer.shtml
Shallow parser	OpenNLP	http://opennlp.apache.org/
	Illinois chunker	http://cogcomp.cs.illinois.edu/page/software
Full parser	Stanford	http://nlp.stanford.edu/software/lex-parser.shtml
	Charniak–Lease reranking	ftp://ftp.cs.brown.edu/pub/nlparser/rerankingparserAug06.tar.gz
	OpenNLP	http://opennlp.apache.org/
	Enju	http://www.nactem.ac.uk/enju/

3.1.1. Kernel Based. In RE, string based kernels are used as described in (Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002), which states that for a pair of strings, a string-kernel computes their similarity based on the on the number of common subsequences they both have. The similarity between the two strings increases with the increasing number of common subsequences between the pair of strings.

***3.1.1.1. Kernel Based Features*.** These features exploits valuable input data representations like shallow parse trees and are designed to remedy the difficulty involved in NLP for the developing of structured input data representations and obtaining an optimal subset of relevant features. In a great number of classes, data sets are not always linearly separable (Kim, Ohta, Tateisi, & Tsujii, 2003). In such cases, training data can be mapped into higher dimensional space through a non-linear mapping. In this new space, the classes may become separable enabling the application of the hyperplane classifier. The kernel function enables such non-linear mapping by taking a representations of two examples and computing their similarities (Kim et al., 2003). There are a lot of proposed kernels ranging from user-defined to general purpose ones. In relation extraction, user-defined kernels are mainly different from each other by their structural representations and how the similarity functions are calculated (Tikk, Thomas, Palaga, Hakenberg, & Leser, 2010). User-defined kernels are designed to handle a specific task for a particular domain while general kernels are applicable to arbitrary tasks. Some common kernels are BOW kernel, Sub tree (ST) kernel, Graph kernel, etc. (Bùi, 2012).

Bag of Word kernel: This uses feature vector as an unordered sets of words, and then calculates the similarity between two compared vectors (Bunescu et al., 2005).

Sub Trees kernel: This uses the syntactic tree representations of sentences. The similarity between the two inputs is calculated by counting the number of common sub-trees (Kim et al., 2008).

Graph kernel: With this kernel, the similarity between two input graphs is calculated by counting weighted shared paths of all possible paths. These paths are obtained from the linear order graph and dependency parse tree (Airola, et al., 2008).

3.1.2. Machine Learning. ML techniques study how to automatically learn to make perfect predictions based on past observations received or information acquired in order to classify a problem or problems into defined or labeled categories. They are generally used as a key component of RE methods and relies on the statistical analysis of the data to deduce a general rule on the data. In RE, annotated texts, phrases or sentences containing entities with relations between or amongst them are needed. Therefore, ML methods are tasked to learn rules from given examples with aim of differentiating characteristics of data from each other in order to predict unknown data from given information or knowledge from known data (Witten and Frank, 2005). In a case where ML technique is used to predict an association (chemical-disease, protein-protein etc.) between a pair of entities from unseen data it corresponds to a supervised machine learning standard in which both a training and testing set of data exists. In the training process, ML method learns about the properties and characteristics (called features) of the text samples in the training data set in order to build a perfect model. Based on these features, ML tries to classify samples into the same class by considering the degree of common features they have than samples of the other class or classes.

3.1.3. Evaluation. The evaluation of entity-relation extraction depends on the method applied and dataset used. To evaluate the performance of a relation extraction system, the following metrics are used: recall, precision, and the F1-score.

$$Recall\ (R) = \frac{TP}{(TP + FN)} \tag{1}$$

$$Precision\ (P) = \frac{TP}{(TP + FP)} \tag{2}$$

$$F1\ Score = \frac{2RP}{P + R} \tag{3}$$

Where:

TP (true positives): is the number of relations that were correctly extracted.

FN (false negatives): is the number of relations that the system failed to extract.

FP (false positives): is the number of relations that were incorrectly extracted.

The F1-score is the harmonic mean of recall and precision.

3.2. Unsupervised methods

In unsupervised learning methods only an input data is given and there are no corresponding output variables. It can also be viewed as a way of finding groups or patterns from an input data, a form of data compression or multi-dimensional reduction (Busa, 2014). The goal here is to model the underlying structure or distribution in the data in order to learn more about the data. They are called unsupervised learning because unlike supervised learning there are no correct answers nor trainer. Algorithms are left to their own abilities to discover and present the interesting structure in the data. Unsupervised learning problems are further grouped into:

Clustering: This tries to discover the presence of inherent groupings in data, such as grouping customers by purchasing behavior or diseases by symptoms.

Association Rule: This tries to discover rules that describe large portions of data. For example, the people that suffer from disease X show the symptoms Y.

These are some of the examples of unsupervised learning algorithms: Apriori algorithm for association rule learning problems and k-means for clustering problems.

3.2.1. Evaluation. In unsupervised learning, there is no prediction target and the selection of a perfect model is almost impossible since there is no accuracy measure to aid such. As explained by Busa (2014), evaluation can be done using some internally or externally defined evaluation methods. In internal evaluation methods, external factors are not considered or replied upon. This method depends on a definition of a set or sets of desired features of the mapping to be considered or that are been considered. This can be done by either creating an error metric or by defining an optimization function (such as minimizing the Stochastic Series Expansion (SSE) in k-means). SSE method plots the sum of squared error for different clusters. In Spectral clustering, the Eigen map is measured or maximized. Penalty Method uses the Bayesian information criterion. Stability based method tries to produce similar clustering on data from similar origin, and then evaluates several models, and selects the model which gives the highest level of stability. In case of external evaluation methods, there exist known class labels, a comparison between the class labels and clustering labels can be made to confirm the clustering accuracy. Furthermore, in external evaluations, the model can be converted into a supervised one by setting up a human/expert panel.

4. Conclusion

In this paper, we have reviewed the aspect of Relation extraction and also Relation extraction in biomedical domain in which we discussed about some of the available biomedical corpora and the tasks they were created to tackle. We also briefly discussed some steps taken to reduce the incompatibility between the PPI corpora and also about the diversity of the BioCreative corpora in terms of the different type of tasks they tackle. Finally, we discussed two methods of relation extraction, their groupings and evaluation techniques.

5. References

Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., & Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC bioinformatics, 9(11), 1.

Bach, N., & Badaskar, S. (2007). A review of relation extraction. Literature review for Language and Statistics II.

Bikel, D. M., Schwartz, R., & Weischedel, R. M. (1999). An algorithm that learns what's in a name. Machine learning, 34(1-3), 211-231.

BioCreative: Critical Assessment of Information Extraction in Biology (2008). [Online] Available: http://www.biocreative.org/ (2008)

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., & Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. Artificial intelligence in medicine, 33(2), 139-155.

Busa, N. (2014), Unsupervised Learning: Model Selection and Evaluation. [Online] Available: http://www.natalinobusa.com/2014/03/unsupervisedlearningmodel.html (March 3, 2014)

Bùi, Q. C. (2012), Relation Extraction Methods for Biomedical Literature.(PhD thesis, Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA)). [Online] Available: http://hdl.handle.net/11245/2.146141.

Chapman, W. W., Cohen, K. B. (2009). Current issues in biomedical text mining and natural language processing. Journal of biomedical informatics, 42:757-9.

Corpora for Biomedical Natural Language Processing (2010). [Online] Available: http://compbio.ucdenver.edu/ccp/corpora/obtaining.shtml. (September 8, 2010)

Culotta, A., McCallum, A., & Betz, J. (2006, June). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 296-303). Association for Computational Linguistics.

Dimililer, N. (2010). Biomedical Named Entity Recognition from Text using a Vote Based Multiple Classifier System (Unpublished doctoral thesis). Eastern Mediterranean University, North Cyprus.

Ekbal, A., Saha, S., & Sikdar, U. K. (2013). Biomedical named entity extraction: some issues of corpus compatibilities. SpringerPlus, 2(1), 1.

Erhardt, R. A. A., Schneider, R.,& Blaschke, C (2006). Status of text-mining techniques applied to biomedical text. Drug discovery today, 11:315-25.

Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2008). Open information extraction from the web. Communications of the ACM, 51(12), 68-74.

Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 363-370). Association for Computational Linguistics.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. I. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl 1), i180-i182.

Kim, S., Yoon, J., & Yang, J. (2008). Kernel approaches for genic interaction extraction. Bioinformatics, 24(1), 118-126.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. Journal of Machine Learning Research, 2(Feb), 419-444.

Nakashole, N., Weikum, G., & Suchanek, F. (2012). Discovering and exploring relations on the web. Proceedings of the VLDB Endowment, 5(12), 1982-1985.

Ohta, T., Pyysalo, S., Kim, J. D., & Tsujii, J. I. (2010). A Re-Evaluation of Biomedical Named Entity–Term Relations. Journal of bioinformatics and computational biology, 8(05), 917-928.

Onye, S. C., Dimililer, N., &Akkeleş, A. (2016). Chemical-disease Relation Extraction with SVM and Enhanced Internal Features. 3rd International Conference on Data Mining, Electronics and Information Technology (DMEIT'16).

Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F., & Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. BMC bioinformatics, 9(3), 1.

Robin, (2015), Parts of Speech Tagging. [Online] Available: http://www.language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.html (March 19, 2015)

The WBI corpus repository (2011),[Online] Available: http://corpora.informatik.hu-berlin.de(November 23, 2011)

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., & Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. PLoS Comput Biol, 6(7), e1000837.

Witten, I.H., & Frank, E. (2005). Data Mining Practical Machine Learning Tools and Techniques. 2nd edition. Morgan Kaufmann Publishers; 525.

Zhou, D., & He, Y. (2008). Extracting interactions between proteins from the literature. Journal of biomedical informatics, 41(2), 393-407.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., & Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. Briefings in bioinformatics, 8(5), 358-375.