IDENTIFYING PASSAGES DESCRIBING PROTEIN-PROTEIN INTERACTION DETECTION METHODS IN BIOMEDICAL FULL TEXT ARTICLES USING INFORMATION RETRIEVAL METHODS

by

Ferhat Aydın

B.S., Computer Engineering, Boğaziçi University, 2011

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2016

IDENTIFYING PASSAGES DESCRIBING PROTEIN-PROTEIN INTERACTION DETECTION METHODS IN BIOMEDICAL FULL TEXT ARTICLES USING INFORMATION RETRIEVAL METHODS

APPROVED BY:

Assist. Prof. Arzucan Özgür .....
(Thesis Supervisor)

Assist. Prof. Albert Ali Salah .....

Assoc. Prof. Deniz Yuret .....

DATE OF APPROVAL: 12.07.2016

ACKNOWLEDGEMENTS

I am truly grateful to my advisor Arzucan Özgür. Her guidance and support have been essential in the development of this thesis. I would also like to thank the other members of my thesis committee, Albert Ali Salah and Deniz Yuret for their careful criticism and insightful feedback that improved the quality of this work.

I am very thankful to my undergraduate advisor Fatih Alagöz. Also, with his encouragement, I could start to this master program. I would like to thank all the past and current members of the Computer Engineering Department of Boğaziçi University to give insight and enthusiasm of Computer Science.

This thesis would have not been possible without the support of my family. Words cannot express my gratitude to my parents, my sisters and my brother for unconditionally providing their love, guidance, and support, and making it possible for me to pursue my goals in life.

Special thanks go to my girlfriend for her great support and patience, and for loving me so much.

I would like to thank Zehra Melce Hüsünbeyi and Hacer Karataş for their help in annotating the data set.

I would like to thank the BioC Task organizers for organizing the shared task and for their help with the questions.

This research was partially supported by Marie Curie Career Integration Grant (CIG) within the 7th European Community Framework Programme (Project acronym BioLitContextMining).

ABSTRACT

IDENTIFYING PASSAGES DESCRIBING PROTEIN-PROTEIN INTERACTION DETECTION METHODS IN BIOMEDICAL FULL TEXT ARTICLES USING INFORMATION RETRIEVAL METHODS

Information regarding the physical interactions among proteins is crucial, since protein-protein interactions (PPIs) are central for many biological processes. The experimental techniques used to verify PPIs are also vital for characterizing and assessing the reliability of the identified PPIs. A lot of information about PPIs and the experimental methods are only available in the text of the scientific publications that report them. In this thesis, we approach the problem of identifying passages with experimental methods for physical interactions between proteins as an information retrieval search task. The baseline system is based on query matching, where the queries are generated by utilizing the names (including synonyms) of the experimental methods in the Proteomics Standard Initiative - Molecular Interactions (PSI-MI) ontology. We propose two methods, where the baseline queries are expanded by including additional relevant terms. The first method is a supervised approach, where the most salient terms for each experimental method are obtained by using the term frequency-relevance frequency (tf.rf) metric over 13 articles from our manually annotated data set of 30 full text articles, which is made publicly available as an additional contribution of this study. The first method is evaluated on the test set consisting of the remaining 17 articles and achieves better recall score compared to the baseline. On the other hand, the second method is an unsupervised approach, where the queries for each experimental method are expanded by using the word embeddings of the names of the experimental methods in the PSI-MI ontology. The second method achieves better recall and F-measure scores over the test set compared to the baseline.

ÖZET

BİYOMEDİKAL TAM METİN MAKALELERDE PROTEİN-PROTEİN ETKİLEŞİMİ TESPİT YÖNTEMLERİNİN BETİMLENDİĞİ PASAJLARI BİLGİ ERİŞİM YÖNTEMLERİ İLE BELİRLEME

Proteinler arası etkileşimlerin birçok biyolojik süreçte aktif rol oynaması, aralarındaki fiziksel etkileşimlere dair bilgileri de fazlasıyla önemli kılıyor. Proteinler arası etkileşimleri doğrulamak için kullanılan deneysel teknikler, etkileşimlerin güvenilirliğini değerlendirebilmek için oldukça önemli. Proteinler arası etkileşimlere ve bu etkileşimleri tespit etmek için kullanılan deneysel yöntemlere dair verilerin önemli bir kısmı sadece bilimsel yayınlarda gömülü durumda bulunuyor. Bu çalışmada, tam metin makalelerde, proteinler arası fiziksel etkileşimleri tespit etmek için kullanılan deneysel yöntemlerin anlatıldığı pasajları belirleme problemine bilgiye erişim alanının arama problemi olarak yaklaşılıyor. Temel sistem, Proteomics Standard Initiative - Molecular Interactions (PSI-MI) ontolojisinde yer alan deneysel yöntem isimleri derlenerek yaratılan sorguların metin içerisinde eşleştirilmesine dayanıyor. Bu temel sistem sorgularının, alakalı terimler ile genişletilmesine dayalı iki yeni yöntem daha geliştirildi. İlk yöntem gözetimli bir yaklaşım olup, terim sıklığı-ilgililik sıklığı (tf.rf) metriğinin, el ile etiketleyip bu çalışmanın ilave çıktısı olarak yayınladığımız 30 makalelik veri kümesinin 13 makalelik eğitim alt kümesi üzerinde uygulanması ile en belirgin terimlerin sorgulara eklenmesine dayanıyor. Bu yöntemin 17 makalelik test kümesi üzerinde değerlendirilmesi sonucu temel sisteme kıyasla bulma skorunda daha başarılı oldugu gözlenmiştir. İkinci yöntem ise gözetimsiz bir yaklaşım olup deneysel yöntem isimleri için öğrenilmiş kelime temsilleri kullanılarak sorgular genişletilmiştir. Bu yöntemin 17 makalelik test kümesi üzerinde değerlendirilmesi sonucu temel sisteme kıyasla bulma ve F-ölçütü skorlarında daha başarılı olduğu gözlenmiştir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . vii
LIST OF TABLES . . . ix
LIST OF ACRONYMS/ABBREVIATIONS . . . x
1. INTRODUCTION . . . 1
1.1. Contributions of the Thesis . . . 7
1.2. Organization of the Thesis . . . 8
2. BACKGROUND . . . 9
2.1. PSI-MI ontology . . . 9
2.2. BioCreative and BioC . . . 11
2.3. Word Embeddings and Word2vec . . . 13
3. RELATED WORK . . . 16
4. DATA SET . . . 21
5. METHODOLOGY . . . 25
5.1. Baseline for Query Matching . . . 26
5.2. Term frequency - Relevance Frequency based Query Generation . . . 28
5.3. Word Embeddings based Query Generation . . . 31
6. EXPERIMENTAL RESULTS . . . 44
6.1. Evaluation . . . 44
6.1.1. Jaccard Index for Passage Similarity . . . 44
6.1.2. Evaluation Measures . . . 45
6.2. Results . . . 48
6.3. Discussion . . . 52
7. CONCLUSION . . . 55
REFERENCES . . . 57

LIST OF FIGURES

Figure 1.1. The growth of Pubmed 2

Figure 1.2. BioGRID example for experimental method 3

Figure 1.3. IntAct example for experimental method 4

Figure 1.4. Sample passage with PPI and experimental method 5

Figure 1.5. Sample text with multiple PPIs and experimental methods 6

Figure 2.1. Two hybrid experimental method in the PSI-MI ontology 10

Figure 2.2. The position of the 'two hybrid' in the hierarchy graph 10

Figure 2.3. Continuous Bag-of-Words and Skip-Gram models of Word2vec . . 14

Figure 4.1. A sample annotation from a paragraph of an article. 22

Figure 5.1. Overall system workflow 25

Figure 5.2. The query expansion algorithm of wor2vec approach. 33

Figure 5.3. The combining algorithm of word2vec based approach 34

Figure 5.4. The cleaning algorithm of word2vec based approach 35

Figure 5.5. The filtering algorithm of word2vec based approach. 36

Figure 6.1. An example paragraph for evaluation logic 47

LIST OF TABLES

Table 4.1. List of experimental methods in data set 23

Table 5.1. Initial query examples for the baseline algorithm 27

Table 5.2. Expanded query with tf.rf (manual) for the 'pull down' 29

Table 5.3. Expanded query with tf.rf (automatic-7) for the 'pull down' 30

Table 5.4. Expanded query with tf.rf (automatic-10) for the 'pull down' . . . 30

Table 5.5. Expanded query using word embeddings for 'two hybrid' 36

Table 5.6. Expanded query using word embeddings for 'pull down' 41

Table 6.1. Performance of the methods on the test set. 50

Table 6.2. Additional results on the test set. 53

LIST OF ACRONYMS/ABBREVIATIONS

a	Number of passages in positive category
c	Number of passages in negative category
BIND	Biomolecular Interaction Network Database
BioC	A simple XML format to share text data and annotations
BioCreative	Critical Assessment of Information Extraction in Biology
BioGRID	Biological General Repository for Interaction Datasets
DIP	Database of Interacting Proteins
IntAct	IntAct Molecular Interaction Database
MI	Molecular Interactions
MINT	Molecular INTeraction Database
PMC	PubMed Central
PPI	Protein-Protein Interaction
PSI-MI	Proteomics Standard Initiative - Molecular Interactions
tf	Term Frequency
tf.rf	Term Frequency - Relevance Frequency

1. INTRODUCTION

Proteins are the core elements of biological structures and the functions of proteins are often modulated through interactions with other proteins. The interactions between proteins are named as Protein-Protein Interactions (PPIs) in the literature. PPIs play important roles in many biological processes including cell cycle control, DNA replication, translation, transcription, and metabolic and signaling pathways [1]. One way to advance the understanding of biological processes is exploring more PPIs, besides understanding the currently known PPIs deeper. Many studies are being conducted in this domain, which are mostly reported in the format of scientific publications. The advancements in the technology increase the capability of producing and consuming huge amounts of biological data, which leads to the publication of more scientific studies. For example, the number of articles indexed in Pubmed [2], which is the main system to index the citations and abstracts of the articles for biomedical literature, has grown exponentially over the past years, as shown in Figure 1.1 [3]. Today, PubMed contains citations of more than 26 million articles and continues growing. Moreover, a subset of the articles in PubMed can be accessible as full text in PubMed Central (PMC) [4], which is an electronic archive of full text journal articles, offering free access to its contents and contains nearly 4 million articles. While the rapid growth of the number of published articles means new findings and knowledge are issued for other researchers, extracting or finding the relevant pieces of knowledge from this huge data pool is getting much more difficult. And yet, the approaches, which aim to solve the difficulty of accessing the right knowledge in the massive amount of data, gain too much importance in the domain.

A lot of information about PPIs are only available in the text of the scientific publications that report them. A number of databases such as Biological General Repository for Interaction Datasets (BioGRID) [5], IntAct Molecular Interaction Database (IntAct) [6], Database of Interacting Proteins (DIP) [7], Molecular INTeraction (MINT) [8], and Biomolecular Interaction Network Database (BIND) [9] have been developed to store PPI information in well structured format in order to facilitate

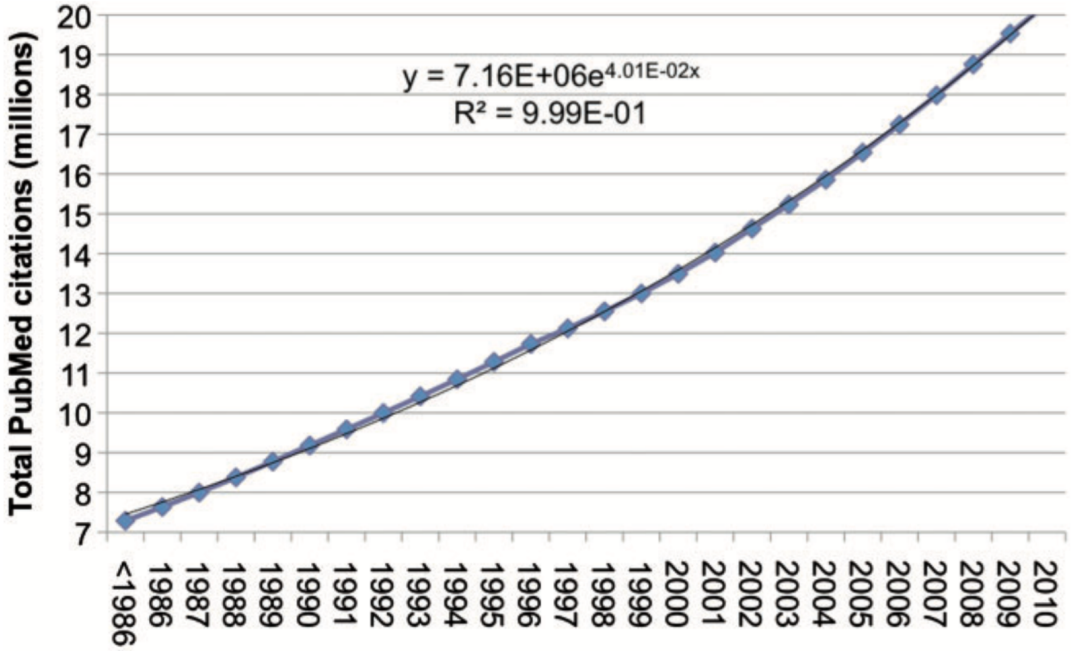


Figure 1.1. The total number of articles in Pubmed from 1986 to 2010 is shown. The growth rate is nearly 4%. The figure is included with the permission of the authors [3].

data retrieval and systematic analysis. The PPI information in these databases is extracted manually by human curators from the published literature. However, manual curation is a laborious and time consuming task. Therefore, it is only able to handle a small fraction of the rapidly growing biomedical literature [10]. In order to address this challenge, several text mining studies have been conducted for automatically extracting information from the published articles. The community-wide shared tasks such as BioCreative [11–13] and BioNLP [14–16] have played important roles for promoting research in this area. Being one of the main tasks in these community-wide efforts, extracting interactions among proteins has gained significant attention from the researchers. Although improvements have been obtained in extracting PPIs from text in the recent years [17, 18], enriching PPIs with context information including the experimental methods used to detect the PPIs has not been well studied yet [19]. Various experimental methods such as 'two hybrid' and 'affinity chromatography' are available for detecting physical interactions between proteins [1]. Two hybrid system

uses transcriptional activity of a reporter gene by using two fusions ('hybrids'), which are constructed between each protein of interest [20]. Affinity chromatography based approaches purify the molecules of interest (baits) and their binding partners by using affinity resins from a biochemical mixture. Experimental methods have different degrees of resolution, confidence, and reliability. Therefore, besides the existence of an interaction between a pair of proteins, the experimental conditions in which this interaction was observed are also very important for the interpretation and assessment of the interaction [19]. Some manually curated databases such as BioGRID and IntAct provide the name and definition of the experimental method, which was used to detect the interaction between proteins, besides the details of the interacting proteins and the corresponding publication, as shown in Figures 1.2 and 1.3, respectively. In such resources, the passages which explain the details and conditions of the conducted experiments to detect the interacting proteins are not given with their exact locations in the articles.

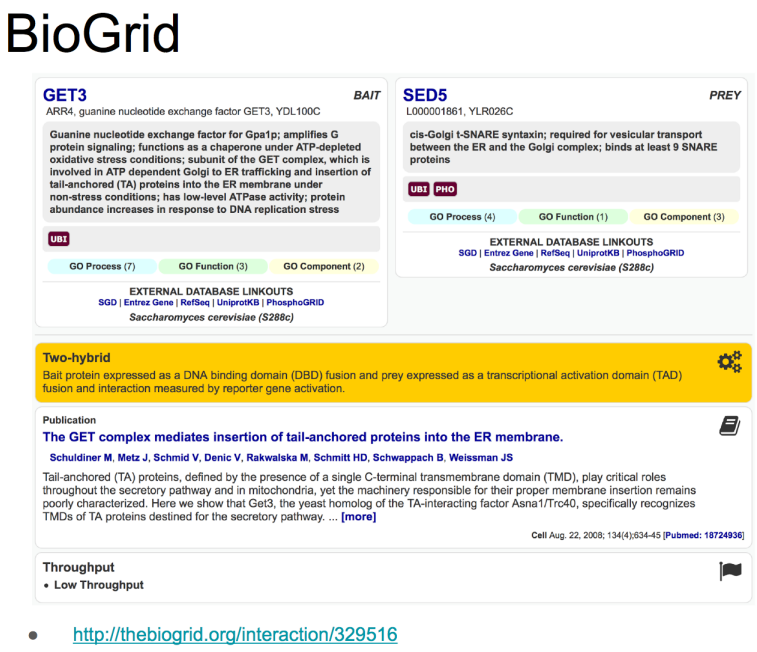


Figure 1.2. A screenshot from the BioGRID database showing information about the interaction between Get3 and Sed5 proteins [21]. The details of the interaction such as which experimental method was used to determine the interaction and in which article are provided.

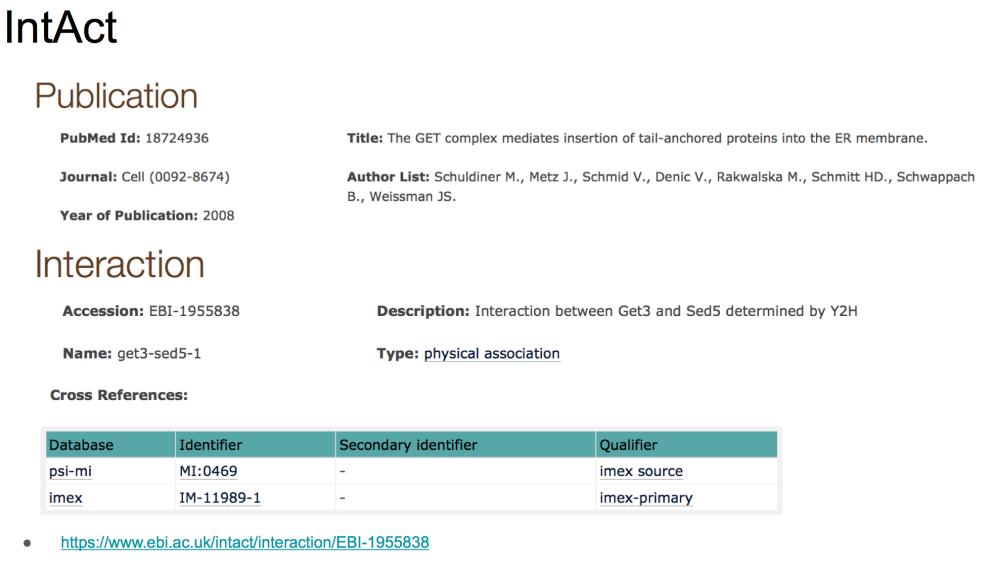
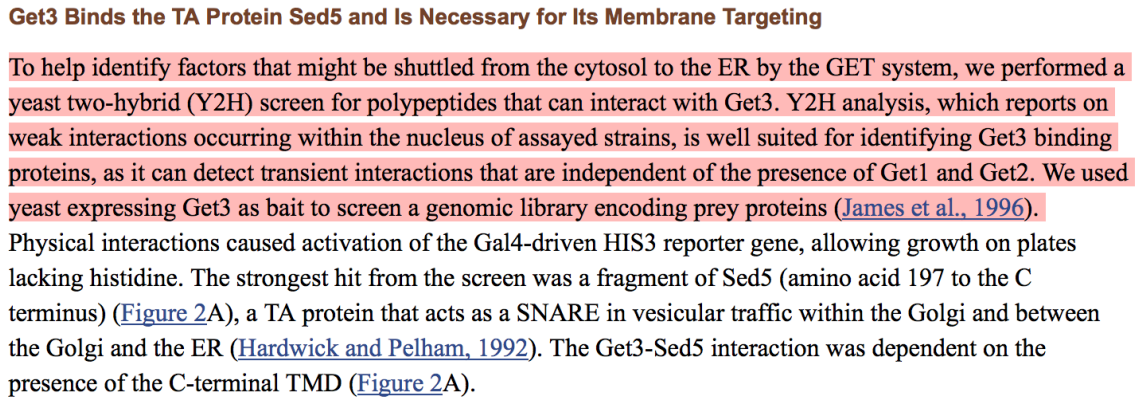


Figure 1.3. A screenshot from the IntAct database showing information about the interaction between Get3 and Sed5 proteins [22]. The details of the interaction such as which experimental method was used to determine the interaction and in which article are provided.

In this work, we approach the problem of experimental method detection as a passage retrieval task. We target identifying passages (i.e., sequences of sentences) where certain experimental methods are described. In many cases, experimental method descriptions span multiple sentences. For example, the interaction between Get3 and Sed5 proteins is detected with the 'two hybrid' experimental method [23] and the passage which explains the details in the article is highlighted in Figure 1.4. The detection of passages describing experimental methods is important for providing evidence for the protein-protein interactions.

Passage-level retrieval is especially crucial for articles in which multiple PPIs and experimental methods are mentioned. Passage-level retrieval can help mapping PPIs to their corresponding experimental methods. For instance, consider the sample text from [24] shown in Figure 1.5. The text (which is a single paragraph in the article) describes three experimental methods used to identify the proteins interacting with the 'TANK-binding kinase 1' (TBK1) protein. Different PPIs were observed



Get3 Binds the TA Protein Sed5 and Is Necessary for Its Membrane Targeting

To help identify factors that might be shuttled from the cytosol to the ER by the GET system, we performed a yeast two-hybrid (Y2H) screen for polypeptides that can interact with Get3. Y2H analysis, which reports on weak interactions occurring within the nucleus of assayed strains, is well suited for identifying Get3 binding proteins, as it can detect transient interactions that are independent of the presence of Get1 and Get2. We used yeast expressing Get3 as bait to screen a genomic library encoding prey proteins (James et al., 1996). Physical interactions caused activation of the Gal4-driven HIS3 reporter gene, allowing growth on plates lacking histidine. The strongest hit from the screen was a fragment of Sed5 (amino acid 197 to the C terminus) (Figure 2A), a TA protein that acts as a SNARE in vesicular traffic within the Golgi and between the Golgi and the ER (Hardwick and Pelham, 1992). The Get3-Sed5 interaction was dependent on the presence of the C-terminal TMD (Figure 2A).

Figure 1.4. Sample paragraph from [23] which contains a passage explaining the interaction between Get3 and Sed5 proteins which is detected with the 'two hybrid' experimental method.

by using the 'tandem affinity purification', 'mass spectrometry studies of complexes', and 'coimmunoprecipitation' experimental methods. Tandem affinity purification is a specific affinity chromatography method to allow purification of different protein complexes from biochemical mixtures by using fusion of multiple tags to the protein of interest [25]. Mass Spectrometry Studies of Complexes is used for the identification of subunit stoichiometry, transient associations, unknown compounds, and structural and chemical properties of macromolecular complexes. Coimmunoprecipitation is a method to separate the protein of interest (bait) from a biochemical mixture by using an antibody specific to the bait [1]. Consider the passage about the 'coimmunoprecipitation' experiment (shown in green). This passage states that no interactions were observed between the protein pairs TBK1-DDX3X, TBK1-IRF3 and DDX3X-IRF3 by using the 'coimmunoprecipitation' experimental technique. This example illustrates that experimental method descriptions may span multiple sentences. In addition, it demonstrates that identifying the passages (i.e., sequences of sentences in a paragraph) describing the experimental methods is important for resolving which method detected which of the PPIs described in text.

To identify novel interactors of TBK1, we generated stable RAW264.7 cell lines that express a GS-TAP-tagged version of TBK1 and purified the TBK1 protein complex by tandem affinity purification (Figure 1A and B) (Burckstummer et al, 2006). Mass spectrometry analysis identified TBK1 along with the core complex components TRF family member-associated NF-kappa-B activator (TANK; 17 peptides), TBK-binding protein 1 (TBKBP1, also referred to as SINTBAD; 10 peptides) and TBKBP2 (also referred to as NAP1 or AZI2; 15 peptides), indicating that this native purification was efficient and in agreement with previously published data on the TBK1 core complex (Bouwmeester et al, 2004). In addition, we identified the DEAD-box helicase DDX3X (RefSeq NP_034158) with five peptides. Immunoprecipitation experiments using tagged TBK1 suggested that its interaction with DDX3X and the transcription factor IRF3 are significantly weaker than the interaction between TBK1 and TANK and therefore not detected by coimmunoprecipitation under stringent conditions (Supplementary Figure 1A). Likewise, we could not detect any association of DDX3X with IRF3 (Supplementary Figure 1B).

Figure 1.5. Sample text with multiple PPIs and experimental methods taken from the Results section of [24]. The passages describing the experimental methods 'tandem affinity purification', 'mass spectrometry studies of complexes', and 'coimmunoprecipitation' are highlighted with yellow, purple, and green, respectively.

We present two query matching based approaches for retrieving passages related to physical PPI detection methods from articles. The first approach is based on generating queries using the term frequency - relevance frequency (tf.rf) metric. The second approach is based on generating queries by using the word embeddings of the experimental method names (i.e., the canonical name and synonyms) in the PSI-MI ontology. We obtained the word embeddings by using the *word2vec* [26] tool, which is an efficient implementation of neural networks based learning techniques for constructing word vectors from large unlabeled data sets with billions of words [27, 28].

1.1. Contributions of the Thesis

We addressed the problem of detecting passages in full text articles that describe specific experimental methods for physical protein-protein interactions as an information retrieval search task. The contributions of this thesis can be summarized as follows:

(i) A supervised query expansion approach based on the term frequency - relevance frequency (tf.rf) term weighting metric is developed.
(ii) An unsupervised query expansion approach that utilizes the word embeddings learned from a large unlabeled biomedical corpus is introduced.
(iii) A data set consisting of 30 full text articles is manually annotated for passages describing experimental methods and made publicly available.

The initial phase of this thesis, which includes only the tf.rf based approach, was published [29] at the BioCreative V Challenge workshop [30]. The paper extracted from this thesis has been submitted [31] (now in the review phase) to the BioCreative Virtual Issue [32] of the Database journal (The Journal of Biological Databases and Curation) [33]. Moreover, the overview paper of the BioC track of the BioCreative V Challenge workshop was published [34].

1.2. Organization of the Thesis

The organization of the thesis is as follows.

Chapter 2 aims to brighten the background of this study by giving some information about the PSI-MI ontology, BioCreative Challenge, BioC format, and finally Word embeddings and the Word2vec tool.

Chapter 3 details the related work on experimental method detection (in the BioCreative Challenge Evaluations), passage segmentation, query expansion based passage retrieval, word embeddings and word2vec based studies.

Chapter 4 introduces the features of the data set, as well as how it is prepared and used in this study.

Chapter 5 details the proposed algorithms and approaches to detect the passages related to experimental methods. After an ontology based query matching algorithm is defined as a baseline, tf.rf and word embeddings based query generation methods are constructed on top of the baseline.

Chapter 6 introduces the Jaccard Index based passage similarity approach to evaluate the performance of the developed approaches. Then, the results of each algorithm are presented and compared with each other.

Chapter 7 concludes the thesis with a summary of the study. The possible future topics which can be investigated based on the results and approaches of this study are given. The position of this thesis with respect to the related work in the literature is explained.

2. BACKGROUND

2.1. PSI-MI ontology

Molecular Interaction (MI) data is very important in the biomedical domain to understand the functions and roles of molecules, which are determined by the interactions among molecules, in the biological metabolisms. It is crucial to have a common and community format to exchange the molecular interaction data effectively in biomedical studies and systems. In 2004, Human Proteome Organisation Proteomics Standards Initiative Molecular Interactions (HUPO PSI-MI) [35] developed and published a community format, which has been named as PSI-MI XML1.0, for protein interaction data with the help of major interaction data producers and providers such as BIND, DIP, and IntAct [36]. In 2007, the HUPO PSI-MI work group has extended the original PSI-MI format and published the current version (PSI-MI XML2.5) by adding more interactor types, extending the descriptions of interacting molecules, and adding kinetic and modeled interaction parameters [37].

Controlled vocabularies (CVs) are used for the standardization of the PSI-MI schema by providing an ID with the namespace prefix 'MI', a name, and a definition for each data object. CVs also have a hierarchical structure, which organize the data objects according to relations among them [37].

The target data objects of this study are under the 'experimental interaction detection' (MI:0045) according to CVs hierarchy [38]. 103 experimental methods are extracted from the PSI-MI ontology [39] with their names, definitions and synonyms. 'Two hybrid' is one of the target experimental methods and its ontology record is shown in Figure 2.1 as an example. On the left side of Figure 2.1, the definition and the synonyms of the two hybrid experimental method can be seen. On the right side of Figure 2.1, the hierarchical position of the 'two hybrid' experimental method is shown. The 'two hybrid' is a specific kind of 'transcriptional complementation assay' (MI:0232) and has siblings like 'one hybrid' and 'three hybrid' as shown in Figure 2.2 [40].

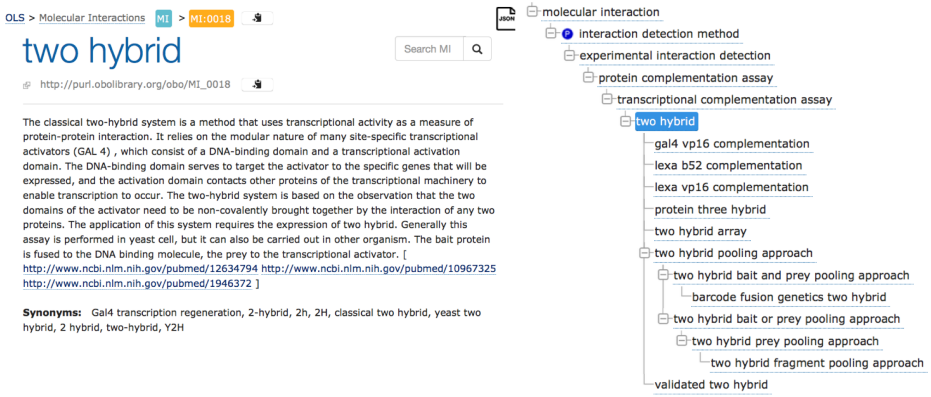


Figure 2.1. The 'two hybrid' experimental method in the PSI-MI ontology is shown.

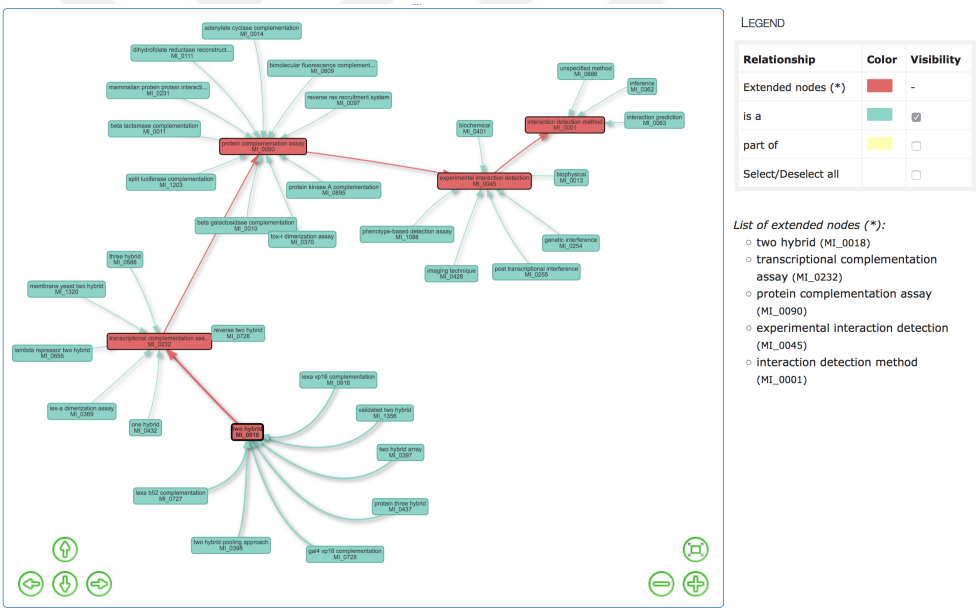


Figure 2.2. The position of the 'two hybrid' in the hierarchy graph.

2.2. BioCreative and BioC

The increasing number of text mining and information extraction based studies which were addressing different problems and using private data sets in the biological domain without common standards or shared evaluation criteria brought a community-wide organization out. The organization of BioCreative (Critical Assessment of Information Extraction in Biology) [41] was motivated to solve the above problems by organizing community challenges with predetermined evaluation criteria to enable comparison of developed systems in terms of performance and measure the capability of adaption to the real world. BioCreative has been organizing series of challenges, ranging from gene name recognition and protein-protein interaction extraction to chemical compound and drug name recognition, to achieve the following goals [42]:

- Perform an assessment of the state of the art for specific tasks.
- Monitor improvements of different approaches by promoting some tasks continuously.
- Define biologically meaningful tasks, which have practical relevance to facilitate efficient information access for the biology community.
- Develop gold standard datasets for training and test purposes with domain experts.
- Explore the evaluation standards for common problems or tasks.
- Reinforce the community by gathering people from different disciplines to ease the exchange of experiences.

In the initial phase of this work, we attended the Collaborative Biocurator Assistant Task (BioC) of the BioCreative V Challenge [43]. The purpose of this task was to create BioC [44] compatible modules which complement each other and integrate into a system that assists the BioGRID database curators during the manual curation process of scientific publications. BioC [45] a simple XML based format to share and represent text data and annotations. Its goals are simplicity, interoperability and broad use and reuse. The BioC track had some goals such as; defining a collaborative task for molecular interaction (MI) information extraction, so different BioC-compatible mod-

ules can be developed independently by separate teams and practical MI tools (i.e., an annotation assistant tool for biocurators) can be obtained by combining these modules. Moreover, a full text benchmark data set is aimed to be defined while evaluating the biocurator assistant tool [46]. The BioC track comprised eight subtasks;

(i) gene/protein named entity recognition (NER),
(ii) species/organism NER,
(iii) normalization of gene/protein names,
(iv) retrieval of passages with PPIs,
(v) retrieval of passages with genetic interactions,
(vi) retrieval of passages with experimental methods for physical interactions,
(vii) retrieval of passages with genetic interaction types,
(viii) development of a visual tool for displaying the annotation results from all the tasks above.

The subtask that we focused is finding passages describing experimental methods used for identifying physical interactions in BioC formatted articles. Our work for this subtask has been published in the BioCreative V Workshop [29]. A complete overview of the BioC track of BioCreative V is accessible at [46].

2.3. Word Embeddings and Word2vec

The term 'word embeddings' has been very popular in the recent years, especially with the prevalent use of neural language modeling in natural language processing studies. The term actually refers to dense representations of words in a low-dimensional vector space in which each word is represented as a vector (i.e., embedded into a vector) and each vector is a point in continuous vector space. In computational linguistics, word embeddings correspond to distributional semantic models (word space models) in which the semantic similarities between words are based on their distributional properties in a large corpus [47]. The distributional semantic models and word embeddings take their roots from the distributional hypothesis in linguistics. According to the distributional hypothesis, words that occur in similar contexts (the context of a word means the words around it) tend to have similar meanings [48]. That means, vectors of similar words are mapped to nearby points in vector space [49]. Therefore, the vectors of semantically similar words should be close to each other in vector space.

There are two main approaches for distributional semantic models; count-based methods and predictive methods. In count-based methods, unsupervised learning techniques are applied to a sparse word-context frequency matrix, which holds the co-occurrence counts of words in a context, to generate dense word vectors. For example, in information retrieval, the term-document matrix which holds the occurrence count of each term in the documents is used to generate embedding of each term [50] by applying Latent Semantic Analysis (LSA), which is a particular application of Singular Value Decomposition (SVD) to reduce large sparse vector spaces for word meaning. Again, by applying truncated SVD to the co-occurrence count based word-context matrix [51], like LSA on term-document matrix, dense vector representations of words can be generated. On the other hand, in predictive models, initial (dense) vectors are generated with random weights for each word in the vocabulary, then the weights are predicted (fine-tuned through back-propagation) according to an objective for maximizing the probability of a word's context in a large corpus [27, 52, 53]. Since semantically similar words occur near each other (same contexts), in neural language models, a randomly initialized word vector is moved close to the vectors of its context words and moved

away from unrelated context vectors at each iteration. Therefore, similar words are represented with similar vectors and the vectors of similar words are located close to each other in the vector space. For example, Word2vec is a feed-forward neural network with a single hidden layer used as a projection layer shared for all words in the architecture [27]. Word2vec proposes two predictive models and a toolkit that is an efficient implementation of these models, which are the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. The CBOW model predicts a word given the context, while the Skip-Gram model does the inverse and predicts context-words from a given word [27] as shown in Figure 2.3.

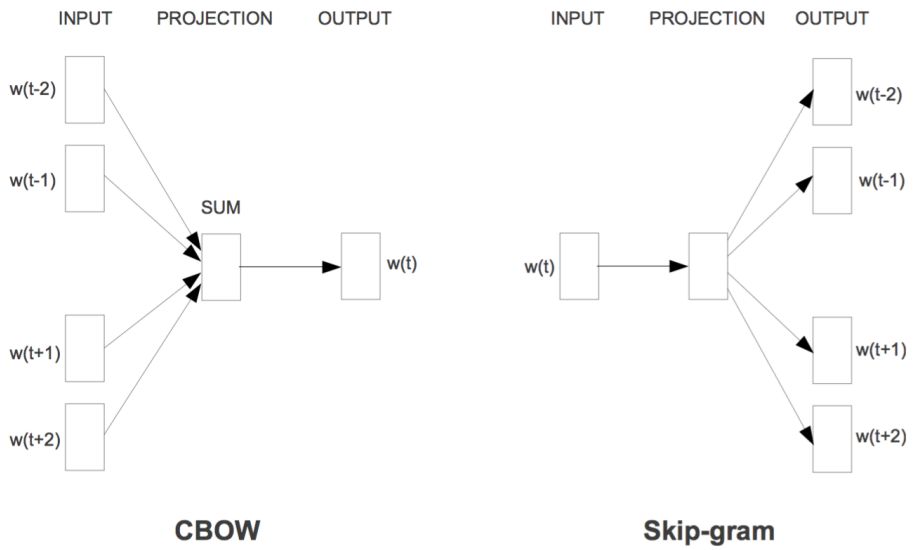


Figure 2.3. Word2vec model architectures; Continuous Bag-of-Words (CBOW) and Skip-Gram model (taken from [27]). The CBOW model predicts a word given the context, while the Skip-Gram model predicts context-words from a given word.

In the Skip-Gram architecture, words are initialized with random vectors of predefined dimension (e.g., 200). The vector weights of each word in a given sequence of words, $w_1, w_2, ..., w_T$, are optimized according to the objective, which is to maximize the average log probabilities of the surrounding words in the context as shown in Equation 2.1, where k is range of context [27, 54].

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{j=-k \\ j\neq 0}}^{k} \log p(w_{t+j}|w_t) \tag{2.1}$$

Since the Skip-Gram architecture is suitable to learn context words for a target word from a large corpus, it can be used to expand the initial queries taken from an ontology to enrich or diversify the queries by learning similar context words for each item in the query.

3. RELATED WORK

Most previous studies on experimental method detection, including the ones in the BioCreative challenges, used pattern matching and/or machine learning based approaches. In the pattern matching based approach the experimental method names in text are matched against the terms in a lexicon or ontology such as the PSI-MI ontology using usually hand-crafted patterns [55–58]. Pattern matching based methods are able to identify the positions of the experimental method mentions in the articles. However, they fail to identify the experimental methods when they occur in forms that do not match the designed patterns. In order to handle approximate matches, Matos et al. developed an Information Retrieval based system, where the test documents are indexed and searched for experimental methods using the Lucene search library [59]. In the machine learning based approach, the task of experimental method detection is in general formulated as a text classification task, where the classes are defined as the experimental methods and the goal is to classify the articles into zero or more of these classes. Machine learning based methods obtained promising results in the BioCreative III challenge, where different classification algorithms such as Naive Bayes [60], Random Forest [61], Support Vector Machines [62], and Logistic Regression [62] were utilized. Machine learning based methods classify articles as containing a certain experimental method or not. Experimental methods can be detected, even if they don't occur with their standard names or synonyms. However, the positions of the experimental methods in the articles mainly are not identified.

The BioCreative II PPI task was organized to develop text-mining tools for a complete manual annotation process for PPIs and evaluate the performance of these tools. Four different subtasks were determined for participants; (i) article classification for PPI; (ii) extraction and normalization of PPIs; (iii) retrieval of the interaction detection methods; and (iv) retrieval of evidence passages for PPIs. A total of 26 teams attended at least one of the subtasks. The problem of identifying the experimental methods used to detect a given PPI in an article was tackled by the third subtask which was Interaction Method Subtask (IMS) [17]. Two teams participated in the

sub-task [63, 64]. Rinaldi et al. obtained promising results by using mostly manually crafted patterns for matching the experimental method terms in the provided ontology against the full text article including the PPI [63]. Ehrler et al. used a pattern matching and vector space retrieval based model [64]. The fourth subtask, namely the Interaction Sentences Subtask (ISS) of the BioCreative II PPI task [17], addressed the problem of retrieving evidence passages that describe interactions between proteins in full text articles. Since multiple evidence passages for an interacting protein pair can be in full text articles, for each protein interaction pair, a ranked list of evidence passages was aimed to be retrieved by the participant teams of the subtask. Each passage had to contain at most three consecutive sentences. The retrieved passages from the participants were compared to the gold passages from biocurators for evaluation using a string similarity method. Team 4 achieved the highest score with 19% of the passages being mapped correctly to the gold passages by using various helper terms such as interaction-indicating keywords, names of experimental methods, and summary-indicating cue words.

A different approach for the same problem, which is the automatic extraction of the interaction detection methods from the biological literature, was proposed by Wang et al. [65]. They developed a generative topic model, the Correlated Method-Word model (CMW), in which the correlation between interaction detection methods and the terms related to methods was formulated in a probabilistic framework to infer interaction detection methods from the words in text. They collected a data set of 5319 full text articles from PubMed with method annotations from MINT and IntAct, which was used for evaluation of their system. Evaluation was done based on the classification of the articles into the experimental methods. The CMW model achieved competitive results against classifiers like SVM, KNN and Naive Bayes on their prepared data set. In addition, the performance of their system outperformed the best result reported in the BioCreative II IMS task. Moreover, they showed that the CMW model is able to capture the correlation between the interaction detection methods and words related to them.

A similar task, namely the Interaction Method Task (IMT), was also addressed at the BioCreative III challenge [11, 66]. The goal in the IMT task at BioCreative III was to identify the experiential methods in a given full text article and map them to the interaction detection method terms in the PSI-MI ontology [36, 37]. The positions of the experimental methods in the articles were not required to be identified. 42 runs were submitted by 8 teams. The main approaches used were matching of manually designed patterns and classification using machine learning methods. Team 90 achieved promising results by defining the problem as a classification problem and then applying Support Vector Machines (SVM) and Logistic Regression (LR) algorithms. Their results are in; (i) macro-averaged case: AUC iP/R (interpolated area under the precision/recall curve) score of 52.974% with an F-measure of 54.616; (ii) micro-average case: precision of 52.30%, recall of 58.25%, F-measure of 55.117, MCC (Matthew's Correlation Coefficient) of 0.5420, and AUC iP/R of 0.3542 [62].

A related task, namely the task of extracting evidence sentences for Drug-Drug Interactions (DDIs) from abstracts, was addressed by Kolchinsky et al. [67]. Supervised machine learning algorithms including Logistic Regression, Support Vector Machines (SVM), Binomial Naive Bayes, and Linear Discriminant Analysis (LDA) were used to classify sentences as evidence sentences for DDI or not. Their results showed that word bigram features were important for good performance. In addition, they observed that domain dictionaries and named entity recognition tools led to improved performance. In our study, we also investigate using a domain-specific named entity recognition tool, as well as domain-specific ontologies such as PSI-MI and INO (Interaction Network Ontology) [68].

Query expansion based approaches are widely used for Information Retrieval (IR) tasks. A search query to a search engine or a question asked to a question answering system can be the initial query and by expanding that query the performance of the designed system is aimed to be improved in general. A good example illustrating the importance of query expansion based approaches in IR tasks is the Clinical Decision Support (CDS) Track of the Text Retrieval Conference (TREC 2014) [69] which was organized to address the task of the retrieval of biomedical articles relevant for an-

swering clinical questions about medical records. Many participants (19 of 23) used some form of query expansion in their work in the CDS track [70]. In terms of query expansion methods, two categories of methods were mainly used by participants. The first category was the usage of lexicons and other ontological resources and used by 13 participants in their work. The second category was the usage of pseudo-relevance feedback (PRF), which is used to extract relevant terms by feeding the initial query with its top results. PRF based query expansion was used by 8 participants.

In the recent years, with the increasing popularity of distributional semantic models in the natural language processing domain, word embeddings have being used also in information retrieval tasks, especially for expanding queries with terms that are semantically similar to the terms in the query. Chen et al. [71] described an experiment on the answer passage retrieval task which targets non-factoid answer passages from the web for non-factoid questions. Non-factoid questions like 'How does Word2vec work?' are generally posted on community question answering sites and the answers are detailed with personal explanations and opinions. They showed that expanding a feature-based framework with new distributional semantics features, which are extracted with tools like Word2vec on external web resources, improves the performance of retrieving more relevant answers that may not contain lexically similar but semantically similar terms to the given question. In addition, learning new distributional semantics features or word embeddings on large topically-unconstrained web resources may not give the desired results for domain specific studies. For example, since the problem definition and proposed solution of this study are specific to the biomedical domain and biomedical literature, we did not train Word2vec on Wikipedia dump corpus, a corpus of biomedical full text articles is used instead. Diaz et al. [72] showed that locally-trained word embeddings overperform globally-trained word embeddings for capturing nuances in a topic-specific information retrieval or natural language processing task. Their study was also based on a query expansion approach by using word embeddings which are learned on topically-constrained corpora, instead of large topically-unconstrained corpora.

There are also some interesting applications of word embeddings and query expansion based approaches apart from classical information retrieval tasks such as passage retrieval or information extraction. One of them is 'query2vec' which is proposed by Grbovic et al. [73] and is a two-level word embeddings based architecture for expanding search queries to improve the search retargeting results. While the temporal context of the search queries in a search session was modeled by using the Skip-Gram model of word2vec at the upper layer, the word sequences within a query were modeled by using the Continuous Bag of Words (CBOW) model of word2vec at the bottom layer of the query2vec architecture. Moreover, by inserting ad clicks data in the query sequences in the upper layer, they got promising results for predicting future ad clicks. Initial queries do not always need to be text based. For example, the query expansion based automatic image captioning method proposed by Yagcioglu et al. [74] uses a given image as the initial query. They used the Word2vec toolkit to generate vectors for captions of similar images to the query image which were retrieved from a data set of captioned images. The weighted average of these caption vectors were used as a new textual query to expand the initial query. Then the closest caption to the final query caption is transferred to the input image. The method obtained promising results compared to the the state-of-the-art methods.

In order to improve retrieval results, expanding the initial query with latent and relevant features may be the most intuitive approach. Due to the success and efficiency of word2vec to learn word embeddings from a corpus in the domain of interest, using word embeddings has become a promising approach for query expansion. The problem of extracting passages related to interacting proteins and the experimental method used to determine the interaction as evidence in biomedical full text articles is addressed as a passage retrieval task in this study. The names (with synonyms) of experimental methods which are taken from an ontology (PSI-MI) are used as initial queries. The semantically similar terms for the names of experimental methods learned with word2vec are used to expand the initial queries. As described in this chapter, query expansion approaches by using word embeddings have been proposed for different problems, but to the best of our knowledge, our study is unique in terms of the problem definition and the proposed approaches for the solution of the problem.

4. DATA SET

To the best of our knowledge, there does not exist a data set annotated for experimental interaction detection methods (with MI ontology identifiers) at the passage level (with exact location in the article). The available data sets for experimental methods are annotated at the article level. In other words, only the list of experimental methods for each article are provided. The BioCreative III IMT data set associates interaction detection methods to articles and comprises 2003, 587 and 223 full text articles for training, development and test purposes, respectively [66]. For BioCreative II Interaction Method Subtask (IMS), 740 and 358 full text articles, which were gathered from the IntAct and MINT databases and curated manually, were provided as training and test sets, respectively [17]. Therefore, we needed to create our own data set and selected a subset of the BioCreative III IMT task data set for annotation. The subset of articles is selected according to the availability of the articles in the *PMC* [4] (subset of PMC), as full text, as well as their availability in BioC format. 30 articles from this subset are randomly selected and annotated for passages (i.e., sequences of sentences) that describe an experimental method as an evidence for a physical PPI and for the specific method that each passage describes by two annotators who have natural language processing and information retrieval background. The disagreements between the two annotators were resolved collaboratively. Then, the annotations of 21 articles are controlled and fixed by a domain expert [75]. The data set of 30 articles is split into two parts, where the first part comprises 13 articles and is used as training set in the tf.rf based approach and as validation set in the word embeddings based approach. The remaining 17 articles, which are checked by the domain expert, are used as test set for all methods developed in this work. The total number of annotated passages, the total number of paragraphs which have at least one annotated passage, and the total number of paragraphs which do not contain any annotated passages in the data set of 30 articles are 370, 292 and 1194, respectively.

A sample annotation from a paragraph of an article in the data set is shown in Figure 4.1. Each annotation has an identifier that is incremented by one throughout the article and two infons, which store key-value pairs with any required information in the context [44]. The value of the 'type' infon is set to 'ExperimentalMethod' for all annotations and the value of the 'PSIMI' infon is set to the PSI-MI identifier of the interaction detection method. The 'text' tag holds the annotated sentence(s). The 'location' tag holds the position of the annotated portion in the article with the 'offset' and 'length' attributes. As illustrated in Figure 4.1, different passages (sequences of sentences) in a paragraph can be annotated with different experimental methods. It is also possible that multiple experimental methods are explained in the same passage of a paragraph. In this case, the corresponding passage of the paragraph is annotated with each experimental method separately. If a paragraph comprises a continuous and coherent explanation of one experimental method, then the whole paragraph is annotated with that method only.

```
<passage>
    <infon key="type">paragraph</infon>
    <offset>17753</offset>
    <text>Infection of eukaryotic cells with large DNA viruses often results in extensive interactions of viral
    gene products with macromolecular pathways of the host cell. By using the yeast two-hybrid system, we
    identified cellular hnRNP-K as an interacting protein with ASFV early protein p30. This interaction was
    further confirmed by an in vitro GST-fusion pull-down assay, using either p30 obtained from baculovirus
    system or ASFV infected cell extracts.</text>
    <annotation id="10">
        <infon key="type">ExperimentalMethod</infon>
        <infon key="PSIMI">0018</infon>
        <location offset="17916" length="123"/>
        <text>By using the yeast two-hybrid system, we identified cellular hnRNP-K as an interacting protein with
        ASFV early protein p30.</text>
    </annotation>
    <annotation id="11">
        <infon key="type">ExperimentalMethod</infon>
        <infon key="PSIMI">0096</infon>
        <location offset="18040" length="163"/>
        <text>This interaction was further confirmed by an in vitro GST-fusion pull-down assay, using either p30
        obtained from baculovirus system or ASFV infected cell extracts.</text>
    </annotation>
</passage>
```

Figure 4.1. A sample annotation from a paragraph of an article in the data set. Each annotation has an identifier that is incremented by one throughout the article. Moreover, the value of the 'type' infon is static and set to 'ExperimentalMethod' for all annotations. The value of the 'PSIMI' infon is set to the PSI-MI identifier of the interaction detection method. The 'text' tag holds the annotated sentence(s). The 'location' tag holds the position of the annotated portion in the article with the 'offset' and 'length' attributes.

The articles are annotated by considering 103 interaction detection methods [76] in the PSI-MI ontology (the nodes under 'MI:0045' [38] which defines 'experimental interaction detection'). The annotation statistics for the 35 interaction detection methods that are annotated in at least one article in the data set are shown in Table 4.1. The PSI-MI identifiers of the methods, their canonical names in the PSI-MI ontology, the number of articles each method occurs in, as well as the total number of passages annotated for each method are presented in the table. 15 interaction detection methods are annotated in only one article and 7 methods are annotated in only one passage. The most common methods at the article-level (i.e., annotated in the highest number of different articles) are 'pull down', 'coimmunoprecipitation', 'two hybrid', 'anti bait coimmunoprecipitation', and 'anti tag coimmunoprecipitation'. The most common methods at the passage-level are 'two hybrid', 'coimmunoprecipitation', 'pull down', 'nuclear magnetic resonance', 'chromatin immunoprecipitation assay', 'anti bait coimmunoprecipitation', and 'x-ray crystallography'.

Table 4.1. List of experimental interaction detection methods which are annotated in at least one article in the manually annotated data set.

Id	Name	Articles	Passages
MI:0004	affinity chromatography technology	3	5
MI:0006	anti bait coimmunoprecipitation	8	23
MI:0007	anti tag coimmunoprecipitation	7	14
MI:0014	adenylate cyclase complementation	2	2
MI:0017	classical fluorescence spectroscopy	1	4
MI:0018	two hybrid	10	54
MI:0019	coimmunoprecipitation	14	49
MI:0029	cosedimentation through density gradient	1	1
MI:0030	cross-linking study	2	5
MI:0040	electron microscopy	1	4
MI:0053	fluorescence polarization spectroscopy	1	1
MI:0054	fluorescence-activated cell sorting	3	4
MI:0055	fluorescent resonance energy transfer	3	10

Table 4.1. List of experimental methods in data set (cont.)

Id	Name	Articles	Passages
MI:0065	isothermal titration calorimetry	3	8
MI:0071	molecular sieving	4	9
MI:0077	nuclear magnetic resonance	4	27
MI:0081	peptide array	1	4
MI:0096	pull down	15	43
MI:0104	static light scattering	1	1
MI:0107	surface plasmon resonance	2	3
MI:0114	x-ray crystallography	5	21
MI:0276	blue native page	1	2
MI:0402	chromatin immunoprecipitation assay	5	24
MI:0411	enzyme linked immunosorbent assay	2	3
MI:0412	electrophoretic mobility supershift assay	1	2
MI:0413	electrophoretic mobility shift assay	1	6
MI:0416	fluorescence microscopy	5	15
MI:0419	gtpase assay	2	4
MI:0423	in-gel kinase assay	1	1
MI:0426	light microscopy	1	1
MI:0663	confocal microscopy	3	6
MI:0676	tandem affinity purification	1	4
MI:0809	bimolecular fluorescence complementation	1	8
MI:0858	immunodepleted coimmunoprecipitation	1	1
MI:0889	acetylase assay	1	1

5. METHODOLOGY

An information retrieval based system for identifying passages that describe an experimental method as evidence for physical PPI is developed. The overall workflow of the system is shown in Figure 5.1. The system pipeline takes a BioC article as input, processes it, and returns the article with the annotated passages for experimental methods in BioC format as output. *The BioC Java library* [44, 77] is used to read, modify, and re-create the BioC files.

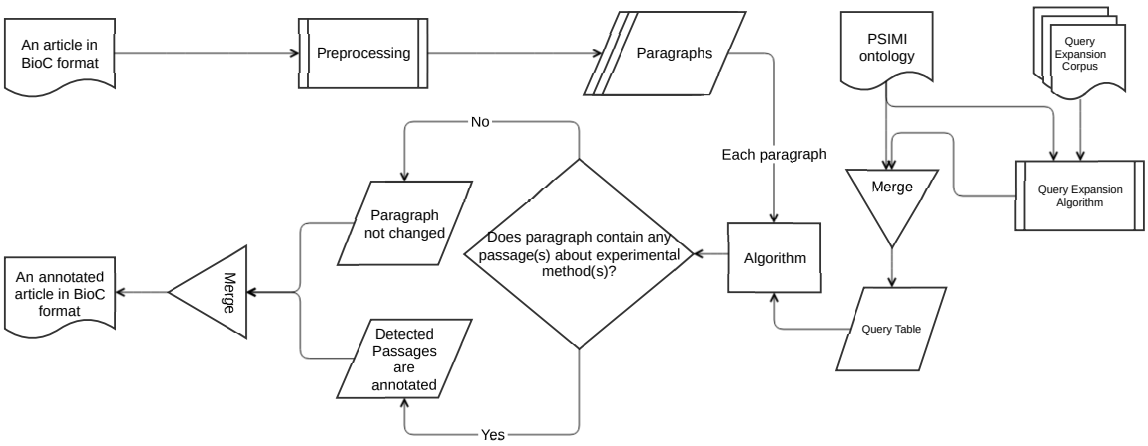


Figure 5.1. Overall system workflow

In the preprocessing step a rule-based sentence splitting method, which we developed based on the period followed by a space pattern, is used. The infon types such as 'title', 'table caption', 'table', 'ref', 'footnote', and 'front' are excluded, since the text of some of these infon types are not sentences, but may contain experimental method relevant keywords. In order to reduce the number of false positives, paragraphs tagged with these infon types are not used for query matching. Moreover, even if paragraphs are tagged with the infon types that we do not exclude, they are not used for query matching if they comprise less than five words. We observe that such short paragraphs with infon types that we do not exclude, in general result due to incorrect tag assignment during BioC format conversion. For example, a header may be tagged with 'paragraph' infon instead of 'title', and the text may be 'Pull-down Experiment Re-

sults'. The *Stanford CoreNLP toolkit* [78–80] is used to tokenize the sentences. At the tokenization phase, punctuation marks, braces, left and right parentheses, brackets, digits, floats etc. are removed from the sentences.

Three query matching based algorithms are designed to retrieve passages that describe specific experimental methods. All three algorithms share the same main idea that a query is generated for each experimental method and included in the query table. The queries in the query table are used to match against the paragraphs in the input article to annotate the passages with experimental methods. The article is returned from the pipeline either unchanged or annotated for the passages with the matching experimental methods.

Each algorithm is described in detail in the following sections.

5.1. Baseline for Query Matching

The baseline algorithm defines an initial query for each experimental method by using the names of the experimental method in the PSI-MI ontology. An example of the initial queries for 'affinity chromatography technology', 'two hybrid' and 'pull down' experimental methods are shown in Table 5.1. While 'pull down' only has its name without any synonyms in the ontology, 'affinity chromatography technology' and 'two hybrid' have more than one synonyms. The algorithm uses the terms in the initial queries of experimental methods to detect relevant passages. The baseline algorithm is also designed to provide a base for the construction of the other two algorithms by expanding the initial queries while determining the minimum performance line to be improved.

The sentences in the paragraphs are matched against the query table of the initial queries for each experimental method. The initial queries contain terms which can be unigrams, bigrams, or trigrams (word-level). If a sentence contains a term from the initial query of an experimental method, the sentence is annotated with that experimental method. If there are successive sentences with the same annotation,

Table 5.1. Initial queries for the 'affinity chromatography technology' (MI:0004), 'two hybrid' (MI:0018) and 'pull down' (MI:0096) experimental methods.

MI:0004	**MI:0018**	**MI:0096**
affinity chromatography technology	two hybrid	pull down
affinity chrom	two-hybrid	
affinity purification	yeast two hybrid	
	2 hybrid	
	2-hybrid	
	y2h	
	classical two hybrid	
	gal4 transcription regeneration	
	2h	

they are concatenated under one annotation tag. As a result, sentences or groups of sentences (passages) in paragraphs are annotated for experimental methods.

Two algorithms are developed on top of the baseline for expanding the initial query. The first algorithm, is a supervised approach and uses a training set of articles annotated for passages with experimental methods. The most salient query terms are selected based on the term frequency - relevance frequency (tf.rf) term weighting metric [81]. The second algorithm is an unsupervised approach and utilizes a large unlabeled corpus for query expansion based on the word embeddings of the initial query terms.

5.2. Term frequency - Relevance Frequency based Query Generation

The texts under the 'annotation' tags in the paragraphs (see Fig. 4.1) of the manually annotated BioC articles in our data set were used as input for the tf.rf method. These texts were filtered according to each experimental method, split into sentences, and tokenized. The frequency of each token was calculated and token-frequency tuples were prepared. These tuples were used to calculate the weight of each token with the tf.rf method as follows.

$$tf.rf = tf * \log_2(2 + \frac{a}{max(1, c)}) \tag{5.1}$$

tf is the number of times the token occurs in the passages annotated for the given experimental method (i.e., passages in the positive category), a is the number of passages in the positive category that contain the token, and c is the number of passages in the negative category (i.e., passages annotated with other experimental methods) that contain the token. The intuition behind relevance frequency (rf) is that a term that occurs more in the positive category compared to the negative category has more discriminating power.

For each experimental method the terms are ranked by their tf.rf weights and manually examined to create the first tier tf.rf and second tier tf.rf term lists and the initial query of that experimental method is expanded by these lists. The first tier tf.rf list consists of high scored relevant tf.rf terms, whereas the second tier tf.rf list consists of lower scored, yet still relevant terms. An example expanded query for the 'pull down' is shown in Table 5.2. We also investigated selecting the first and second tier term lists automatically. Table 5.3 shows the expanded query for the 'pull down' generated by selecting the top 7 terms based on their tf.rf scores as first tier terms, and the next top 7 terms as second tier terms. Similarly, Table 5.4 shows the expanded query when the top 10 terms based on their tf.rf scores are selected as first tier terms and the next top 10 terms are selected as second tier terms. The names of the experimental methods are excluded from the first and second tier lists even if they have high tf.rf weights, since they are already included in the initial query.

Table 5.2. Expanded query for the 'pull down' experimental method. The names are extracted from the PSI-MI ontology. The Tier 1 and Tier 2 terms are extracted manually based on tf.rf weights.

Names	**Tier 1 Terms**	**Tier 2 Terms**
pull down	pull-down	flag-tagged
	down	pull
	pulled	pulled-down
	gst	gst-fusion
	his-tagged	glutathione
	s-transferase	glutathione-sepharose
	affinity	

The sentences in the paragraphs are matched against the created queries for each experimental method. First, the names list of the expanded query is used. The names list contains terms which can be unigrams, bigram, or trigram (word-level). On the other hand, the first and second tier lists only consist of unigrams. The names found in the sentence have weight of 1.0. The terms in the first and second tier lists are searched in the sentences. A matching term from the first tier list is assigned the weight of 0.50, whereas a matching term from the second tier list is assigned the weight of 0.25. These weights have been set without tuning, but heuristically by giving full weight to a name/synonym in the PSI-MI ontology, half of this weight to a term from the first tier list, and 25% of this weight to a term from the second tier list. The threshold for selecting a sentence as relevant to an experimental method is set as 1.0. That means, existence of a name or a synonym of an experimental method in the sentence is enough to annotate the sentence with the corresponding experimental method, but if there is no name or synonym in the sentence, at least a Tier 1 term and two Tier 2 terms or two Tier 1 terms or four Tier 2 terms are need for annotation. The previous and next sentences of the selected sentence are also processed to check whether they are relevant to the same experimental method or not. If the previous and next sentences

Table 5.3. Expanded query for the 'pull down' experimental method. The names are extracted from the PSI-MI ontology. The Tier 1 and Tier 2 terms are constructed automatically from the first 7 and second 7 top terms of tf.rf weights.

Names	**Tier 1 Terms**	**Tier 2 Terms**
pull down	pull-down	binding
	gst	gst-hnrnp-k
	rab5	recombinant
	appl1	interaction
	down	his-tagged
	proteins	protein
	melk	mutations

Table 5.4. Expanded query for the 'pull down' experimental method. The names are extracted from the PSI-MI ontology. The Tier 1 and Tier 2 terms are constructed automatically from the first 10 and second 10 top terms of tf.rf weights.

Names	**Tier 1 Terms**	**Tier 2 Terms**
pull down	pull-down	interaction
	gst	his-tagged
	rab5	protein
	appl1	mutations
	down	pull
	proteins	used
	melk	gtp
	binding	assay
	gst-hnrnp-k	fusion
	recombinant	figure

of the annotated sentence obtain the highest score for the same experimental method and if this score is greater than or equal to 0.50 (at least a Tier 1 term or two Tier 2 terms), they are annotated with the same experimental method. All the successive sentences with the same annotation are concatenated under one annotation tag. As a result, sentences or groups of sentences (passages) in paragraphs are annotated for experimental methods.

5.3. Word Embeddings based Query Generation

We used *word2vec* to expand the initial experimental method queries consisting of the PSI-MI ontology terms by using word embeddings learned from a large unlabeled biomedical corpus. A set of 691,558 full text articles from the *PMC Open Access* [82] database is used as unlabeled data. All the articles are passed from a preprocessing pipeline. The Stanford CoreNLP tool is used for conversion of the data to lower case, tokenization and sentence splitting. Then, the punctuation marks are removed using a manually prepared list of punctuation marks. The numeric and non-ascii characters are also removed [83]. After the preprocessing steps, all 691,558 articles are merged into a single text file in order to use as input for *word2vec*.

Since experimental method names generally consist of multiple words, vectors for words as well as phrases, which consist of up to four words (bigram, trigrams and fourgrams), are required. The phrases are obtained by running *word2phrase* [84], which uses bigram statistics to form phrases, twice (i.e., consecutively) with the default parameters on the preprocessed unlabeled data. The minimum word occurrence count is set as 5, the threshold parameter is set as 200 and 100 for the first and second runs, respectively. The phrases are treated as individual tokens like words (during training). The resulting data set contains 2,241,223,681 total and 3,229,270 unique tokens. This data set is given as unlabeled input data to the *word2vec* tool, which is run with the *Hierarchical Softmax* [85, 86] based training algorithm and the *Skip-Gram* [27] architecture with the suggested default settings for the parameters. Context window size, sub-sampling rate, and training iteration count are set as 10, $1e-4$, and 15, respectively. Minimum word occurrence count is set as 10. In other words, words

appearing less than 10 times are removed. As a result, word vectors with size 200 are generated.

The word vectors of an experimental method's names (and synonyms) in the PSI-MI ontology are used to expand the initial query for the experimental method. The word vectors for some terms of the initial queries could not be constructed by *word2vec* because of the insufficient data related to those terms in the input unlabeled data set. For example, for the initial query of the 'two hybrid' (MI:0018) experimental method consisting of the terms 'two hybrid', 'two-hybrid', 'yeast two hybrid', '2 hybrid', '2-hybrid', 'y2h', 'classical two hybrid', 'gal4 transcription regeneration', and '2h', only the vectors of the '2-hybrid', 'two-hybrid', 'y2h', and '2h' terms were constructed. For each term (which has a word vector) in the initial query, the top 100 terms whose word vectors are most similar (in terms of cosine similarity) to the word vector of the initial query term are retrieved by using the modified version of the *distance* [87] component of *word2vec*. Then, for each term in the initial queries, the top 100 similar terms are manually analyzed. It is observed that for ambiguous initial query terms such as the '2h' initial query term of the 'two hybrid' experimental method, non-relevant terms with high cosine similarity scores are retrieved. Therefore, such ambiguous terms are removed from the initial queries of the corresponding experimental methods. The list of terms that are removed are (MI:0016, cd), (MI:0018, 2h), (MI:0053, fps), (MI:0055, ret), (MI:0099, spa), (MI:0104, sly), (MI:0112, myth), (MI:0114, x-ray), (MI:0226, ice), (MI:0419, gtpase), (MI:0428, microscopy), (MI:0437, trihybrid), (MI:0676, tap), (MI:0728, kiss), and (MI:0825, x-ray).

For each term in the initial query of an experimental method, the most similar 100 terms are obtained. If the initial query contains k terms, then a pool of $k * 100$ terms is obtained for the corresponding experimental method. A list of terms for each experimental method is created from its pool of terms by removing; (i) the duplicate terms by leaving the term with the highest cosine similarity score, (ii) the terms that are already listed as a name or synonym of an experimental method in the PSI-MI ontology (see the combine function in Figure 5.3). The list of combined terms is further filtered (after expanding the initial query with terms whose scores are higher than the main

threshold of the annotation process (see the following sections)) by removing; (iii) the terms which contain a name from an initial query of another experimental method, e.g., 'gst-pull-down-assay' and 'pull' are removed from the 'two hybrid' results, (iv) the terms which have a higher cosine score in the list of another experimental method, e.g., 'coprecipitated, 0.82' is removed from the results of 'coimmunoprecipitation' (MI:0019), since it is already in the results of 'pull down' experimental method with 0.83 score (see the clean function in Figure 5.4). The terms, which already have a substring with higher score in the list, are cleaned from the list (see the filter function in Figure 5.5). The remaining terms in the list of each experimental method are included to its initial query as expansion terms together with their cosine similarity scores. The cosine similarity scores of the initial query terms are set as 1.0. This procedure is summarized in Figure 5.2.

```
procedure QUERY_EXPANSION
    combinedLists ← COMBINE_RESULTS
    for each expMethod do
        expandedQuery := initialQuery
        combinedList ← combinedLists[expMethod]
        cleanedList ← CLEAN_RESULTS(expMethod, combinedList)
        filteredList ← FILTER_LONGER_TERMS(expMethod, cleanedList)
        add filteredList to expandedQuery
    end for
end procedure
```

Figure 5.2. The query expansion algorithm of wor2vec approach.

As an example consider the 'two hybrid' experimental method. The initial query terms that have word vectors for this method are '2-hybrid', 'two-hybrid', and 'y2h'. The term 'yeast two-hybrid assays' is among the most similar 100 terms for each of these three initial query terms, where the corresponding cosine similarity scores are 0.773, 0.838, and 0.826, respectively. Therefore, the term is included into the list of the 'two hybrid' method with the cosine similarity score 0.838. The other two

```
function COMBINE_RESULTS
    combinedLists ← ∅
    for each expMethod do
        the initialQuery of expMethod
        the initialQueries of any experimental method
        the combinedList of expMethod ← ∅
        for each item i of initialQuery do
            the top100terms of i from word2vec
            for each term in top100terms do
                if term is not in initialQueries and
                    term is not in combinedList with higher score then
                    add term to combinedList
                end if
            end for
        end for
        add (expMethod, combinedList) to combinedLists
    end for
    return combinedLists
end function
```

Figure 5.3. The algorithm for combining the word2vec results of each experimental method into one list.

occurrences of the term are eliminated. The term 'gst pull down' is eliminated while cleaning the list of the 'two hybrid' method, since it contains the term 'pull down' which is a name of the pull down experimental method (MI:0096). Likewise, the term 'bifc' is not included into the list, since it is listed as a synonym of the experimental method 'bimolecular fluorescence complementation' (MI:0809) in the PSI-MI ontology. Moreover, the term 'tap-tag' with cosine score 0.739 is eliminated, since it is listed in the list of 'tandem affinity purification' (MI:0676) experimental method with cosine score 0.804. The term 'yeast two-hybrid screening' with cosine score 0.853 is eliminated since 'two-hybrid screening' is a substring of 'yeast two-hybrid screening', and its score

```
function CLEAN_RESULTS(expMethod, list)
    the cleanedList of expMethod ← ∅
    the initialQuery of expMethod
    the expandedInitialQuery ← initialQuery + terms from list
                                    whose score higher than mainThreshold
    the expandedInitialQueries of other methods other than expMethod
    the combinedLists of other methods other than expMethod
    for each term in list do
        if term does not contain any item from expandedinitialQueries and
           any expandedinitialQueries item does not contain term and
           term is not already in combinedLists with higher score then
            add term to cleanedList
        end if
    end for
    return cleanedList
end function
```

Figure 5.4. The algorithm for cleaning the given list of an experimental method.

is higher (0.860). The final expanded queries for the 'two hybrid' and 'pull down' experimental methods are shown in Table 5.5 and Table 5.6. As a result, for each experimental method, an expanded query with a different size is obtained.

```
function FILTER_LONGER_TERMS(expMethod, list)
    the filteredList of expMethod ← ∅
    for each term in list do
        if list does not contain any item which is a substring of
            term with higher score then
            add term to filteredList
        end if
    end for
    return filteredList
end function
```

Figure 5.5. The algorithm for filtering the longer terms with lower scores from the given list of an experimental method.

Table 5.5. The expanded query terms of 'two hybrid' (MI:0018) are shown in bold. The italic terms are eliminated from the word2vec results in the cleaning operation as explained in Figure 5.4. The terms with score 1.0 are the initial query items (name or synonyms). The terms which already contain (after splitting with space) any name or synonym are also eliminated, so italicized.

Terms	Scores
two hybrid	**1.0**
two-hybrid	**1.0**
yeast two hybrid	**1.0**
2 hybrid	**1.0**
2-hybrid	**1.0**
y2h	**1.0**
classical two hybrid	**1.0**
gal4 transcription regeneration	**1.0**
yeast two-hybrid	*0.91416*
two-hybrid system	*0.874584*
y2h system	*0.869421*

Table 5.5. Expanded query using word embeddings for 'two hybrid' (cont.)

Terms	Scores
two-hybrid experiments	*0.865643*
yeast-two-hybrid	**0.865173**
two-hybrid analysis	*0.851562*
y2h assay	*0.845296*
yeast-two hybrid	**0.844966**
two-hybrid assays	*0.844023*
y2h screens	*0.843109*
two-hybrid assay	*0.842222*
y2h assays	*0.84187*
yeast two-hybrid assays	*0.837754*
yeast two-hybrid y2h	*0.832441*
y2h experiments	*0.830541*
yeast two-hybrid system	*0.826506*
two-hybrid screening	*0.825039*
yeast two-hybrid assay	*0.824025*
yeast-2-hybrid	**0.823851**
yeast 2-hybrid	*0.820577*
y2h screen	*0.81859*
two-hybrid screens	*0.810895*
y2h screening	*0.807603*
y2h interactions	*0.806635*
yeast two-hybrid screens	*0.794152*
two-hybrid interaction	*0.791841*
y2h interaction	*0.787684*
ap-ms	**0.78406**
two hybrid y2h	*0.780946*
high-throughput yeast two-hybrid	*0.778971*
y2h-based	**0.775805**
two-hybrid screen	*0.769459*

Table 5.5. Expanded query using word embeddings for 'two hybrid' (cont.)

Terms	Scores
yeast-two-hybrid y2h	*0.766024*
large-scale yeast two-hybrid	*0.76319*
yeast two-hybrid screening	*0.758309*
protein-protein interaction	**0.756688**
split-ubiquitin	**0.7553**
yeast two-hybrid y2h assays	*0.75367*
yeast two-hybrid y2h screens	*0.749456*
bait-prey	**0.748784**
yeast 2-hybrid assays	*0.745503*
yeast two-hybrid y2h assay	*0.742103*
large-scale y2h	*0.740383*
yeast-two-hybrid experiments	*0.739971*
two-hybrid interactions	*0.739497*
interactors	**0.737921**
yeast two-hybrid screenings	*0.737534*
interacting proteins	**0.73329**
yeast two-hybrid screen	*0.730519*
tap-ms	**0.728466**
literature-curated interactions	**0.725529**
cty10-5d	**0.725047**
yll049wp	**0.715317**
mammalian two-hybrid	*0.713345*
y2h bait	*0.711965*
lexa-based	**0.710877**
ap/ms	**0.708912**
lexa fusions	**0.708365**
bait prey	**0.707578**
yeast-two-hybrid assay	*0.707011*
interaction partners	**0.706699**

Table 5.5. Expanded query using word embeddings for 'two hybrid' (cont.)

Terms	Scores
yeast-two-hybrid system	*0.706006*
jnm1p	**0.705212**
bait	**0.704263**
y3h	**0.703098**
bait plasmid	*0.700969*
putative interactors	*0.700491*
matchmaker gold	**0.698625**
mating-based	**0.696109**
bait construct	*0.695515*
co-ap	**0.695304**
yeast co-transformation	**0.695194**
mbsus	**0.694528**
interacting partners	**0.694077**
protein-interaction	**0.692748**
protein-protein interactions	**0.692531**
yeast strain y190	**0.692009**
matchmaker gal4	**0.691998**
ht-y2h	**0.6908**
domain gal4-ad	**0.689826**
yeast-two-hybrid assays	*0.689708*
bait/prey	**0.689438**
large-scale ap-ms	*0.68708*
y2h library screening	*0.686014*
biogrid	**0.685742**
prey-prey	**0.685525**
two-hybrid library	*0.683535*
high-throughput y2h	*0.679981*
yeast-two-hybrid screen	*0.678596*
yth assays	**0.678315**

Table 5.5. Expanded query using word embeddings for 'two hybrid' (cont.)

Terms	Scores
y2h screenings	*0.678036*
high-confidence interactions	**0.677115**
yeast-two hybrid assays	*0.655188*
glutathione s-transferase gst pulldown	*0.654585*
gst-pulldown assay	*0.646435*
yeast three-hybrid assay	**0.636332**
gst-pulldown experiments	*0.635499*
yeast two-hybrid co-immunoprecipitation	*0.635368*
vitro pull-down experiments	*0.634054*
gst-pull-down assay	*0.63086*
gst-pull-down	*0.628426*
unc-89 ig1	**0.62706**
bait constructs	*0.626819*
gst-pull-down assays	*0.626504*
l40 yeast	**0.625285**
yeast strain mav203	**0.624612**
glutathione s-transferase gst tagged	**0.624112**
gst pulldowns	*0.623283*
y3h assays	*0.622409*
lexa-esc4	**0.618557**
eif2bε	**0.618421**

Table 5.6. The expanded query terms of 'pull down' (MI:0096) are shown in bold. The italic terms are eliminated from the word2vec results in the cleaning operation as explained in Figure 5.4. The terms with score 1.0 are the initial query items (name or synonyms). The terms whose scores are higher than the main threshold (0.9) are counted as synonyms of 'pull down'. The terms which already contain (after splitting with space) any name or synonym are also eliminated, so italicized.

Terms	Scores
pull down	**1.0**
pull-down	**0.939108**
pulldown	**0.92231**
pulled down	**0.913027**
pulldowns	**0.887387**
pull-down assays	*0.8829*
pull-down assay	*0.881704*
pulldown assays	*0.879614*
pull-downs	**0.878394**
pulled-down	**0.877428**
pull-down experiments	*0.873226*
pull downs	**0.869779**
gst alone	**0.858877**
pulldown experiments	*0.855896*
pulldown assay	*0.846789*
immobilized gst	**0.84599**
pull-down experiment	*0.845928*
co-precipitated	**0.844448**
pull down assay	*0.839032*
coprecipitated	**0.833142**
pull down endogenous	*0.832411*
gst-fused	**0.830135**
pulls down	**0.829674**

Table 5.6. Expanded query using word embeddings for 'pull down' (cont.)

Terms	Scores
gst-tagged	**0.827178**
gst pull down	*0.82398*
pull down assays	*0.8225*
purified gst	**0.81814**
glutathione beads	**0.817414**
gst beads	**0.814973**
hek293 cell lysates	**0.812967**
efficiently pulled down	*0.809786*
pull down experiments	*0.809765*
bacterially purified	**0.804833**
glutathione-sepharose beads	**0.80482**
gst alone lane	*0.803281*
purified gst-tagged	*0.798264*
anti-ha beads	**0.798137**
pulldown experiment	*0.796522*
affinity beads	**0.794702**
gst fusions	**0.794422**
co-precipitate	**0.793779**
strep-tactin beads	**0.793182**
flag beads	**0.791584**
gst-βpix	**0.790978**
s-protein beads	**0.789618**
vitro translated	**0.789464**
bacterially expressed	**0.789107**
gst tagged	**0.788444**
glutathione sepharose beads	**0.787931**
glutathione-sepharose	**0.787101**
recombinant gst-tagged	*0.786871*
immunoprecipitate flag-tagged	**0.785623**

Table 5.6. Expanded query using word embeddings for 'pull down' (cont.)

Terms	Scores
gst-pdcd4	**0.78539**
did not co-precipitate	*0.784861*
anti-flag resin	**0.783796**
gst-14-3-3β	**0.776625**

The generated expanded queries are used to identify the passages describing experimental methods in full text articles. Given a full text article in BioC format, the sentences in the paragraphs are matched against the expanded queries of each experimental method. The threshold for selecting a sentence as relevant to an experimental method is set to a certain value, e.g. 0.9 (for the word embeddings based approach) and 1.0 (for the baseline and tf.rf based approaches). If the query score for a sentence is greater than or equal to that threshold, the sentence is annotated with the experimental method for which it scored highest. The previous and next sentences of the selected sentence are also processed to check whether they are relevant to the same experimental method or not. If the previous and next sentences of the annotated sentence obtain the highest score for the same experimental method and if this score is greater than or equal to a certain value, e.g. 0.5 (for the baseline and tf.rf based approaches) and 0.65 (for the word embeddings based approach), they are annotated with the same experimental method. All the successive sentences with the same annotation are concatenated under one annotation tag. As a result, sentences or groups of sentences (passages) in paragraphs are annotated for experimental methods.

6. EXPERIMENTAL RESULTS

6.1. Evaluation

6.1.1. Jaccard Index for Passage Similarity

The Jaccard index [88] is a statistic to measure similarity of given finite sets by taking the ratio of the size of the intersection and size of the union of the given sets as shown in Equation 6.1. In contrast, the Jaccard distance, which measures dissimilarity of given sets, is obtained by subtracting the Jaccard coefficient from 1 as shown in Equation 6.2.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{6.1}$$

$$d_J(A,B) = 1 - J(A,B) \tag{6.2}$$

In our case, we use the Jaccard index to calculate the similarity of passages. Each passage can be thought as an ordered sequence of characters (string). When the Jaccard similarity of two passages is measured, the character length of each passage ($|A|$ and $|B|$) and the character length of the intersection of the passages ($|A \cap B|$) are calculated. The intersection of two passages can be in one of the following cases; (i) one passage can cover the other one, so the shorter passage is the intersection, (ii) the two passages can be completely the same, so any of them is the intersection, (iii) if case (i) or case (ii) are not satisfied, the longest common substring of the two passages is extracted as the intersection. After the length of each passage and the length of their intersection are calculated, we can measure the Jaccard similarity of the two passages according to Equation 6.1. The Jaccard index is between [0, 1]. The Jaccard indexes of two exactly the same passages is 1.0, whereas it is 0.0 for completely different passages.

6.1.2. Evaluation Measures

The output articles of the system are compared to the manually annotated versions to evaluate the performance of the system. Each output article of the system is compared to its manually annotated version at paragraph level. Two annotations should have the same experimental method id and (fully or partially) common text to be evaluated as matched annotations. If this is not the case, the annotations are assessed as non-matched. The possible cases are listed below;

(i) If a paragraph contains annotated passages in the manually annotated article, but the corresponding paragraph in the system output article does not have any annotated passages, this corresponds to the case of false negative (FN). The Jaccard distance for each annotated passage in the manually annotated paragraph is 1.0. The sum of those Jaccard distances is added to the total FN score of the evaluation.

(ii) If a paragraph does not have any annotated passages in the manually annotated article, but the corresponding paragraph in the system output article has annotated passages, this corresponds to the case of false positive (FP). The Jaccard distance for each annotated passage in the system output paragraph is 1.0. The sum of those Jaccard distances is added to the total FP score of the evaluation.

(iii) If a paragraph has annotated passages both in the manually annotated article and the system output article, but some of these annotations are non-matched (i.e., either the experimental method ID does not match or the annotated passages do not have any overlapping text), then these correspond to the cases of false negative and/or false positive. For each non-matched passage annotation in the manually annotated article, the false negative score is updated as defined in case (i) above. At the same time, for each non-matched passage annotation in the system output article, the false positive score is updated as defined in case (ii) above.

(iv) If a paragraph has matched annotated passages both in the manually annotated article and the system output article, this corresponds to the case of true positive

(TP). In this case, the Jaccard indexes of those passage pairs are added to the total TP score of the evaluation. In case of exact match, the Jaccard index for a passage pair is 1.0. However, if a passage pair matches partially in terms of common text, after its Jaccard index is added to the TP, we calculate the 'Partial Jaccard Distance' (an adaption of Jaccard distance to this problem by us) for the unmatched parts of the passages as shown in Equations 6.5 and 6.6. The text of the manually annotated passage and the text of the system output passage are represented with M and S, respectively. The unmatched text portion can be part of either the manually annotated passage or system annotated passage. If it is part of the manually annotated passage, the partial Jaccard distance of the manually annotated passage (M) (Equation 6.5) is added to the FN score in the evaluation. Otherwise, if it is part of the system annotated passage, the partial Jaccard distance of the system annotated passage (S) (Equation 6.6) is added to the FP score. The partial Jaccard distance of a target passage from another passage is calculated by subtracting the Jaccard index of the passage pair from the normalized length of the target passage. The normalized length of the target passage of a passage pair can be calculated by taking the ratio of the character length of the target passage and the character length of the union of the passage pair as shown in Equations 6.3 and 6.4.

$$N_M(M,S) = \frac{|M|}{|M \cup S|} \tag{6.3}$$

$$N_S(M,S) = \frac{|S|}{|M \cup S|} \tag{6.4}$$

$$pd_{JM}(M,S) = N_M(M,S) - J(M,S) \tag{6.5}$$

$$pd_{JS}(M,S) = N_S(M,S) - J(M,S) \tag{6.6}$$

The following example, which is shown in Figure 6.1, covers the different cases mentioned above for the evaluation logic of the annotated passages. The paragraph in Figure 6.1 is taken from [89]. This article is also in our published manually annotated data set [90].

As shown in Figure 4A-C, co-expression of ATH1-CYFP/STM-NYFP, BLH3-NYFP/STM-CYFP and BLH9-NYFP/STM-CYFP resulted in reconstitution of YFP fluorescence. In contrast to results obtained with STM-GFP fusions, however, the BiFC signal exhibited a high preference for the nucleus whereas fluorescence was low in the cytoplasm. The BiFC results therefore, confirm that full-length STM and BLH proteins interact in plant cells, but more notably, that BLH/STM heterodimers are efficiently incorporated into the nuclear compartment. Expectedly, the GFP-ATH1, GFP-BLH3 and GFP-BLH9 fusion proteins exerted a preference for the nuclear compartment (Figure 4DF). The nuclear import of GFP-BLH9 in leek epidermal cells appears to be in contrast to previous results obtained in tobacco leaves after agroinfection with a C-terminal BLH9-GFP fusion construct. However, without analysing the cellular localization of the STM protein, Bhatt et al. (19) arrive at the same conclusion that STM/BLH9 heterodimers are efficiently imported into the plant cell nucleus. In contrast to bombardments with the ATH1 and BLH3 constructs, the number of BiFC-positive cells obtained with the BLH9/STM combination (e.g. BLH9-NYFP/STM-CYFP in Figure 4C) was generally low. This may indicate a weak interaction between STM and BLH9 as suggested by the yeast two-hybrid results, that chimeric BLH9-NYFP protein levels remain low in leak epidermal cells, or that steric constraints interfere with a reconstitution of YFP fluorescence. Reminiscent, however, of the MEIS/PBC situation in animals, the GFP and BiFC data show that the TALE-HD transcription factor STM is targeted into the nuclear compartment through interaction with three BLH partner proteins, which are members of the non-MEINOX class of TALE-HD proteins in plants.

Figure 6.1. An example paragraph which shows three different logic of passage evaluation. Manually annotated passages are underlined with red and green for 'bimolecular fluorescence complementation' (MI:0809) and 'two hybrid' (MI:0018) experimental methods, respectively. The annotated passages by the system are colored with blue and purple for 'bimolecular fluorescence complementation' (MI:0809) and 'two hybrid' (MI:0018) experimental methods, respectively.

The first annotation is an over-cover-annotation. The manual annotation text length is 371 characters, whereas the system annotation text length is 523 characters. The union length is 523 and matching (common) text length 371. Since the Jaccard index is $(371/523) = 0.709$, TP is increased with 0.709. FN is not changed (manual annotation is fully covered). The partial Jaccard distance is calculated as $((523 - 371)/523) = 0.291$ and FP is increased with 0.291. The second annotation is an under-cover-annotation. The manual annotation text length is 452 characters and system annotation text length is 258 characters. The union length is 452 and matching text length 258. Since the Jaccard index is $(258/452) = 0.571$, TP is increased with 0.571. FP is not changed. The partial Jaccard distance is calculated as $((452 - 258)/452) =$

0.429 and FN is increased with 0429. The last annotation is a full-cover-annotation. The manual and system annotations are the same. Thus, the Jaccard index is 1.0 and TP is increased with 1.0. FN and FP are not changed. After the calculation of the TP, FP and FN scores for the evaluation, Recall, Precision, F-measure are calculated as shown in Equations 6.7, 6.8 and 6.9, respectively.

$$Recall = \frac{TP}{TP + FN} \tag{6.7}$$

$$Precision = \frac{TP}{TP + FP} \tag{6.8}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6.9}$$

6.2. Results

The developed methods ('baseline', 'tf.rf' and 'word.embeddings') are evaluated under different configurations on the test set, which comprises 17 full text articles, and the performances of the methods are compared with each other as shown in Table 6.1. The baseline approach does not need a training or validation set, since the existence of a name or a synonym of an experimental method determines the result of the annotation for that sentence. There is no a training or parameter tuning phase. On the other hand, the tf.rf based approach is supervised and needs a training set to extract the Tier 1 and Tier 2 terms to expand the initial queries. The training set of 13 articles are used for that purpose in the tf.rf method. Like in the baseline, the threshold value, which determines whether a sentence should be annotated with an experimental method, is determined heuristically as explained in sections 5.1 and 5.2. The threshold value is not tuned. Since the 'word.embeddings' based method is an unsupervised approach, there is no need for a training set, so the 13 articles are used as validation set to determine the threshold values for the main target sentence and the previous and next sentences around the target sentence for annotation. The test

set of 17 articles is shared for all methods and the performances of the methods are measured on it. All 30 articles were initially annotated by two annotators with natural language processing background and the disagreements were resolved collaboratively. Additionally, the initial annotations of the test set of 17 articles were checked, validated, and corrected whenever necessary by a domain expert. These final annotations were used as the gold standard. The Inter Annotator Agreement (IAA) of the test set is computed by comparing the initial annotations of the two annotators (after resolving the disagreement between them) against the gold standard test set checked by the domain expert. Since the number of passages that neither the two annotators nor the domain expert annotator annotated (true negative cases) is undefined, instead of the Kappa measure, traditional information retrieval metrics can be used to calculate the agreement score [91]. Therefore, the evaluation approach described in Section 6.1 is used to measure recall, precision, and F-measure, which are 0.937, 0.787, and 0.856, respectively. These scores of inter annotator agreement can be considered as high compared to the manual annotation agreement scores obtained in general in the biomedical domain [92]. Thus, we believe the published annotated data set has high confidence to be used for further studies.

In the 'tf.rf' based approach, we experimented with three different configurations as explained in Section 5.2. The manual selection of the tf.rf terms is labeled as 'tf.rf.manual' and the automatic selections of the tf.rf terms are labeled as 'tf.rf.f7s7' and 'tf.rf.f10s10' in the result tables. 'tf.rf.f7s7' corresponds to the configuration when the first 7 and the second 7 terms are included in the Tier 1 and Tier 2 lists, respectively. Similarly, 'tf.rf.f10s10' corresponds to the configuration when the first 10 and the second 10 terms are included in the Tier 1 and Tier 2 lists, respectively. The results in Table 6.1 show that all three tf.rf configurations obtain similar recall levels, which are higher than the recall of the baseline. The precision values of the tf.rf configurations with automatically selected terms are much lower than the tf.rf configuration with manually selected term, all of which are lower than the precision of the baseline. The results also show that including more automatically selected terms to the tr.rf approach leads to only a slight increase in recall, but results in drastic decrease in precision and F-measure.

Table 6.1. Performance of the methods on the test set.

	Precision	**Recall**	**F-measure**
baseline	0.424	0.418	0.421
baseline.genia.ino	0.484	0.413	0.446
tf.rf.f7s7	0.120	0.508	0.194
tf.rf.f7s7.genia.ino	0.133	0.502	0.211
tf.rf.f10s10	0.068	0.512	0.119
tf.rf.f10s10.genia.ino	0.074	0.507	0.129
tf.rf.manual	0.315	0.508	0.389
tf.rf.manual.genia.ino	0.357	0.503	0.418
word2vec	0.321	0.618	0.422
word2vec.genia.ino	0.362	0.606	0.453
IAA	0.787	0.937	0.856

In the 'word.embeddings' based approach; the threshold for detecting the main target sentences is set as 0.9 and the threshold for detecting the before and after sentences of the main target sentences is set as 0.65. These thresholds were determined after running the system, on the validation set of 13 articles, with a range of different threshold values to determine the best setting in terms of highest F-measure performance. While improving the recall and F-measure measures (that means improvement in false negative and true positive scores), we try to improve (i.e., decrease) the false positive score by applying an extra constraint which mandates a passage to contain at least one protein and one interaction keyword, in the annotation decision step. The passages which explain the details of how an experiment was conducted (i.e., the experiment layout) without giving any information/explanation or evidence about the interacting proteins are prevented to be annotated by this constraint. Likewise, passages that only include some experimental method related keywords but do not give any detail about the proteins or the interaction are eliminated. The protein detection in the passages is done with the Genia Tagger [93, 94]. To detect the interaction keyword(s) in the passages, Interaction Network Ontology (INO) [68, 95, 96] is used. The terms

under the 'INO_0000006' annotation [97], defining the literature mining keywords, are extracted, the protein related interaction keywords are filtered manually [98]. As illustrated in Table 6.1, this constraint results in considerable improvement in precision and F-measure for all methods (represented with the 'genia.ino' extension) with a relatively smaller decrease in recall.

McNemar's test [99] is used to determine whether the differences between the results shown in Table 6.1 are statistically significant [100]. According to McNemar's test results, the differences between the following cases are statistically significant ($p < 0.05$):

- the 'tf.rf' based approaches and the 'baseline' method
- the 'genia.ino' extensions of the 'tf.rf' based approaches and the 'baseline.genia.ino' method
- the 'word2vec' based approach and the 'baseline' method
- the 'word2vec.genia.ino' based approach and the 'baseline.genia.ino' method
- the proposed methods (e.g., 'word2vec') and their 'genia.ino' extensions (e.g., 'word2vec.genia.ino')

The goal of the BioC Task was to develop a text mining system to assist biocurators. It has been shown that high recall is more desirable than high precision for a system aiming to assist biocurators in exhaustive curation [17]. It is relatively easier for a biocurator to filter the false positives retrieved by the system compared to manually finding the false negatives missed by the system. Therefore, we aimed at increasing the recall of the baseline method by query expansion without sacrificing much from precision. To the best of our knowledge, there does not exist a study on identifying passages describing protein-protein interaction detection methods. So, there are no previously reported state of the art results/scores for this task. Therefore, we try to improve the performance of the baseline approach. As shown in the result tables, the word embeddings based methods and the tf.rf based methods achieve higher recall scores than the baseline method at the expense of lower precision. In addition to higher recall, the word embeddings based approach achieves slightly higher F-measure than the baseline.

In addition, it outperforms the tf.rf based methods in terms of precision, recall, and F-measure, which shows that it is a promising approach for detecting passages with experimental methods.

6.3. Discussion

For a query matching based information retrieval system, it is crucial that queries are expanded with highly relevant terms in terms of both semantic and relational ways to increase the overall performance of the system. The amount of data used to extract the relevant terms from for query expansion affects the quality of the final expanded query. Supervised learning techniques need labeled data. Manual data labeling is a laborious and time consuming process. On the other hand, unsupervised learning techniques use unlabeled data. The amount of published articles in the biomedical domain constitute a large unlabeled data set that can be used for unsupervised learning. Therefore, compared to supervised techniques, unsupervised techniques can help to scan a larger set of articles in the literature for more experimental methods to extract valuable information such as repeating patterns, sentence rules, experiment specific terms, proteins etc. in the experimental method related passages. Nevertheless, if an experimental method is not mentioned enough in the literature, the effort for extracting information related to that experimental method by unsupervised techniques would not give encouraging results. For example, running *word2vec* on 691555 articles ($2,241,223,681$ words) produced word embeddings for the names of only 58 out of 103 experimental methods [101]. For infrequently used (uncommon) experimental methods, manually curated rules or queries created by domain experts can be used to enhance the retrieval results by expanding such rare experimental methods' queries with high quality terms.

There are some hard cases that need to be considered for improving the results. Experimental methods which are siblings or close to each other in the PSI-MI ontology have very similar names or synonyms, definitions and experimental details. For example, 'anti bait coimmunoprecipitation' (MI:0006) and 'anti tag coimmunoprecipitation' (MI:0007) are siblings, and 'coimmunoprecipitation' (MI:0019) is their parent. After

the queries are expanded with unsupervised techniques like the word embeddings based approach, specific patterns or rules may be defined to distinguish these kind of experimental methods from each other. The word embeddings based approach can not expand the queries of 'anti bait coimmunoprecipitation' and 'anti tag coimmunoprecipitation' methods, since their names are not used directly as a method name like 'two hybrid' in the text. Therefore, the system annotates a passage with 'coimmunoprecipitation', even if it is related to 'anti bait coimmunoprecipitation' or 'anti tag coimmunoprecipitation'. When the annotations, which have 0006 and 0007 values in the 'PSIMI' infon, are changed to 0019 in the manually annotated articles, the scores of 'word2vec.genia.ino' configuration are changed as shown in Table 6.2 ('word2vec.genia.ino.coip'). Moreover, if the system is evaluated in experimental method ID agnostic way (which means, the ID of the experimental method (i.e., the PSIMI infon) is not considered as long as the passage retrieved describes an experimental method), the scores increase considerably as shown in Table 6.2 ('word2vec.genia.ino.psimi').

Table 6.2. Additional results on the test set. The system evaluation when the 'anti bait coimmunoprecipitation' and 'anti tag coimmunoprecipitation' methods are regarded as 'coimmunoprecipitation' is shown with 'word2vec.genia.ino.coip'. The system evaluation without the requirement of experimental method ID matching is shown with 'word2vec.genia.ino.psimi'.

	Precision	**Recall**	**F-measure**
word2vec.genia.ino	0.362	0.606	0.453
word2vec.genia.ino.coip	0.390	0.645	0.486
word2vec.genia.ino.psimi	0.439	0.751	0.554

Ambiguous terms also require special focus. For example, the term 'cross-linking', which is a name of an experimental method for detecting physical PPIs (MI:0030), is also frequently used in articles to identify interactions between genes. As future work, we plan to investigate these edge cases and improve the performance of our system by better discriminating ambiguous terms and closely-related experimental methods.

The paragraphs mainly in the materials and methods sections of the articles contain experimental method names, but in general only give experimental layout (i.e., describe how the experiment was performed). Experimental layouts are not direct evidences of PPIs. In our data set, we did not annotate such paragraphs. However, since they may contain experimental method names or related terms, our system can annotate such parts in the articles and this situation causes to an increase in false positive rates. The performance of the system should be improved by defining the layout patterns or adding constraints into the algorithm to filter such cases in the future.

7. CONCLUSION

In this thesis, we have defined the problem of extracting passages with experimental interaction detection methods used to determine physical interactions between proteins from full text articles in the biomedical domain as an information retrieval search task. In order to extract passages describing experimental methods as evidence for physical protein-protein interactions, an initial query for each experimental method is prepared by utilizing the names and synonyms of that experimental method from the PSI-MI ontology. The baseline approach is based on matching those initial queries to the sentences in the paragraphs of articles. To improve the performance of a query matching based approach, the existing queries need to be expanded with more relevant terms. Therefore, we have proposed two new approaches based on the baseline approach. The first method is supervised and based on expanding the initial queries with the terms determined with the tf.rf metric on manually annotated passages. Our second approach is unsupervised and based on expanding the queries with terms determined using the word emmbeddings of the terms of the initial queries. We have also published a data set of 30 full text articles annotated for passages describing experimental methods used to detect physical PPI.

We have applied the tf.rf term weighting metric to passage level classification (instead of its original application to the article level text classification problem). We have started from the idea that determining important and discriminating terms for each experimental method, by assigning appropriate weights, should help the problem of extracting passages and classifying them to the experimental methods that they describe. The tf.rf based methods achieve higher recall scores than the baseline method at the expense of lower precision and F-measure. We also applied distributional semantic models, by using the word2vec tool and generating word embeddings of terms specific to each experimental method defined in the ontology to expand the queries, for the passage extraction problem in the biomedical domain. The word embeddings based approach results in only slight improvement in F-measure, but achieves considerably higher recall compared to the baseline. In addition, the word embeddings based

approach outperforms the tf.rf based approach in terms of precision, recall, and F-measure, which shows that utilizing the huge and growing biomedical literature within an unsupervised learning setting is an effective approach for information retrieval tasks in this domain. We have also shown that empowering unsupervised approaches with extra constraints or features like terms from domain specific ontologies (i.e., interaction keywords from INO) and specific named entities (i.e., protein names, determined with the GENIA Tagger, in the sentences) leads to improved performance.

The most challenging part of this study has been gathering enough labeled data. Manually annotating passages for multi-class of experimental methods in full text domain specific articles is a really difficult and time consuming task. The labeled data are needed for training, validation and as test (gold) set to develop, validate and test the proposed approaches. The lack of enough labeled data for supervised approaches (as training set), like in our tf.rf based approach, limits the learning ability of the proposed algorithms. The tf.rf based approach does not perform well in terms of precision and F-measure because of the small training set available to extract relevant terms for each experimental method. In addition, labeled data are needed (as validation set) to optimize the parameters of unsupervised algorithms, like in the word embeddings based approach. In the future, if we can access enough annotated full text articles we will improve our proposed solutions and test their generalization performance on a larger test set.

An automatic passage (context) extraction system could be a valuable tool for the biomedical domain. It can help to reduce the time consumed on curation for biocurators. In addition, it can help scientists reach relevant information in a structured format by facilitating the expansion of existing databases. We plan to enlarge the target experimental method list and manually annotated data set while trying to solve the edge cases defined in the discussion section.

REFERENCES

1. Phizicky, E. M. and S. Fields, "Protein-protein interactions: methods for detection and analysis.", *Microbiological reviews*, Vol. 59, No. 1, pp. 94–123, 1995.

2. "Pubmed Database Page", `http://www.ncbi.nlm.nih.gov/pubmed`, accessed at August 2016.

3. Lu, Z., "PubMed and beyond: a survey of web tools for searching biomedical literature", *Database*, Vol. 2011, p. baq036, 2011.

4. "Pubmed Central Database Page", `http://www.ncbi.nlm.nih.gov/pmc/`, accessed at August 2016.

5. Chatr-Aryamontri, A., B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell *et al.*, "The BioGRID interaction database: 2015 update", *Nucleic acids research*, Vol. 43, No. D1, pp. D470–D478, 2015.

6. Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia *et al.*, "IntAct: an open source molecular interaction database", *Nucleic acids research*, Vol. 32, No. suppl 1, pp. D452–D455, 2004.

7. Xenarios, I., L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions", *Nucleic acids research*, Vol. 30, No. 1, pp. 303–305, 2002.

8. Chatr-Aryamontri, A., A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli and G. Cesareni, "MINT: the Molecular INTeraction database", *Nucleic acids research*, Vol. 35, No. suppl 1, pp. D572–D574, 2007.

9. Bader, G. D., D. Betel and C. W. Hogue, "BIND: the biomolecular interaction network database", *Nucleic acids research*, Vol. 31, No. 1, pp. 248–250, 2003.

10. Baumgartner, W. A., K. B. Cohen, L. M. Fox, G. Acquaah-Mensah and L. Hunter, "Manual curation is not sufficient for annotation of genomic databases", *Bioinformatics*, Vol. 23, No. 13, pp. i41–i48, 2007.

11. Arighi, C. N., Z. Lu, M. Krallinger, K. B. Cohen, W. J. Wilbur, A. Valencia, L. Hirschman and C. H. Wu, "Overview of the BioCreative III workshop", *BMC bioinformatics*, Vol. 12, No. Suppl 8, p. S1, 2011.

12. Hirschman, L., A. Yeh, C. Blaschke and A. Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology", *BMC bioinformatics*, Vol. 6, No. Suppl 1, p. S1, 2005.

13. Krallinger, M., A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman and A. Valencia, "Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge", *Genome Biol*, Vol. 9, No. Suppl 2, p. S1, 2008.

14. Kim, J.-D., T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii, "Overview of BioNLP'09 Shared Task on Event Extraction", *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pp. 1–9, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009.

15. Kim, J.-D., S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen and J. Tsujii, "Overview of BioNLP shared task 2011", *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 1–6, Association for Computational Linguistics, 2011.

16. Nédellec, C., R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo and P. Zweigenbaum, "Overview of BioNLP shared task 2013", *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 1–7, 2013.

17. Krallinger, M., F. Leitner, C. Rodriguez-Penagos, A. Valencia *et al.*, "Overview of the protein-protein interaction annotation extraction task of BioCreative II", *Genome biology*, Vol. 9, No. Suppl 2, p. S4, 2008.

18. Tikk, D., P. Thomas, P. Palaga, J. Hakenberg and U. Leser, "A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature", *PLoS Comput Biol*, Vol. 6, No. 7, p. e1000837, 2010.

19. Krallinger, M., "Importance of negations and experimental qualifiers in biomedical literature", *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 46–49, Association for Computational Linguistics, 2010.

20. Chien, C.-T., P. L. Bartel, R. Sternglanz and S. Fields, "The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest.", *Proceedings of the National Academy of Sciences*, Vol. 88, No. 21, pp. 9578–9582, 1991.

21. "BioGrid Get3 and Sed5 Interaction Record", `http://thebiogrid.org/interaction/329516`, accessed at August 2016.

22. "Intact Get3 and Sed5 Interaction Record", `https://www.ebi.ac.uk/intact/interaction/EBI-1955838`, accessed at August 2016.

23. Schuldiner, M., J. Metz, V. Schmid, V. Denic, M. Rakwalska, H. D. Schmitt, B. Schwappach and J. S. Weissman, "The GET complex mediates insertion of tail-anchored proteins into the ER membrane", *Cell*, Vol. 134, No. 4, pp. 634–645, 2008.

24. Soulat, D., T. Bürckstümmer, S. Westermayer, A. Goncalves, A. Bauch, A. Stefanovic, O. Hantschel, K. L. Bennett, T. Decker and G. Superti-Furga, "The DEAD-box helicase DDX3X is a critical component of the TANK-binding kinase 1-dependent innate immune response", *The EMBO journal*, Vol. 27, No. 15, pp.

2135–2146, 2008.

25. Puig, O., F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm and B. Séraphin, "The tandem affinity purification (TAP) method: a general procedure of protein complex purification", *Methods*, Vol. 24, No. 3, pp. 218–229, 2001.

26. "Word2Vec Tool", `https://code.google.com/archive/p/word2vec/`, accessed at August 2016.

27. Mikolov, T., K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013.

28. Mikolov, T., W.-t. Yih and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations.", *HLT-NAACL*, pp. 746–751, 2013.

29. Aydın, F., Z. M. Hüsünbeyi and A. Ozgür, "Retrieving Passages Describing Experimental Methods using Ontology and Term Relevance based Query Matching", *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pp. 42–50, 2015.

30. "BioCreative V Challenge Workshop", `http://biocreative2015.org`, accessed at August 2016.

31. Aydın, F., Z. M. Hüsünbeyi and A. Ozgür, "Automatic Query Generation using Word Embeddings for Retrieving Passages Describing Experimental Methods", *Database*, Vol. 2016, 2016, under Review.

32. "Oxford Database Journal BioCreative Virtual Issue", `http://www.oxfordjournals.org/our_journals/databa/biocreative_virtual_issue.html`, accessed at August 2016.

33. "The Journal of Biological Databases and Curation", `http://database.`

oxfordjournals.org/, accessed at August 2016.

34. Kim, S., R. I. Doğan, A. Chatr-Aryamontri, C. S. Chang, R. Oughtred, J. Rust, R. Batista-Navarro, J. Carter, S. Ananiadou, S. Matos *et al.*, "BioCreative V BioC track overview: collaborative biocurator assistant task for BioGRID", *Database*, Vol. 2016, p. baw121, 2016.

35. "Human Proteome Organisation Proteomics Standards Initiative Molecular Interactions (HUPO PSI-MI)", http://www.psidev.info/groups/molecular-interactions, accessed at August 2016.

36. Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering *et al.*, "The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data", *Nature biotechnology*, Vol. 22, No. 2, pp. 177–183, 2004.

37. Kerrien, S., S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. F. Quinn, N. Vinod, G. D. Bader, I. Xenarios, J. Wojcik, D. Sherman *et al.*, "Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions", *BMC biology*, Vol. 5, No. 1, p. 44, 2007.

38. "Experimental Molecular Interaction Detection Ontology Record", http://purl.obolibrary.org/obo/MI_0045, accessed at August 2016.

39. "The European Bioinformatics Institute Molecular Interaction Ontology Lookup Service", http://www.ebi.ac.uk/ols/ontologies/mi, accessed at August 2016.

40. "Two Hybrid Ontology Record", http://www.ebi.ac.uk/ols/ontologies/mi/terms/graph?iri=http://purl.obolibrary.org/obo/MI_0018, accessed at August 2016.

41. "BioCreative Organization Page", http://www.biocreative.org/, accessed at August 2016.

42. "BioCreative Goals Page", http://www.biocreative.org/about/background/goals/, accessed at August 2016.

43. "BioCreative-V Track-1 BioC Task Page", http://www.biocreative.org/tasks/biocreative-v/track-1-bioc/, accessed at August 2016.

44. Comeau, D. C., R. I. Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii *et al.*, "BioC: a minimalist approach to interoperability for biomedical text processing", *Database*, Vol. 2013, p. bat064, 2013.

45. "BioC Sourceforge Page", http://bioc.sourceforge.net/, accessed at August 2016.

46. Kim, S., R. I. Dogan, A. Chatr-aryamontri, M. Tyers, W. J. Wilbur and D. C. Comeau, "Overview of BioCreative V BioC Track", *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pp. 1–9, 2015.

47. Miller, G. A. and W. G. Charles, "Contextual correlates of semantic similarity", *Language and cognitive processes*, Vol. 6, No. 1, pp. 1–28, 1991.

48. Harris, Z. S., "Distributional structure", *Word*, Vol. 10, No. 2-3, pp. 146–162, 1954.

49. Erk, K., "Vector space models of word meaning and phrase meaning: A survey", *Language and Linguistics Compass*, Vol. 6, No. 10, pp. 635–653, 2012.

50. Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American society for information science*, Vol. 41, No. 6, p. 391, 1990.

51. Schutze, H., "Dimensions of meaning", *Supercomputing'92., Proceedings*, pp. 787–796, IEEE, 1992.

52. Bengio, Y., R. Ducharme, P. Vincent and C. Jauvin, "A neural probabilistic language model", *journal of machine learning research*, Vol. 3, No. Feb, pp. 1137–1155, 2003.

53. Collobert, R. and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning", *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, ACM, 2008.

54. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", *Advances in neural information processing systems*, pp. 3111–3119, 2013.

55. Kappeler, T., S. Clematide, K. Kaljurand, G. Schneider and F. Rinaldi, "Towards automatic detection of experimental methods from biomedical literature", *Third International Symposium on Semantic Mining in Biomedicine (SMBM)*, 2008.

56. Lourenço, A., M. Conover, A. Wong, A. Nematzadeh, F. Pan, H. Shatkay and L. M. Rocha, "A linear classifier based on entity recognition tools and a statistical approach to method extraction in the protein-protein interaction literature", *BMC bioinformatics*, Vol. 12, No. 8, p. 1, 2011.

57. Jhamb, D., A. Krishnan, M. Palakal, Y. Pandit, M. J. Palakal and P. Duraiswamy, "Identification of protein interaction methods from biomedical literature", *Computational Advances in Bio and Medical Sciences (ICCABS), 2014 IEEE 4th International Conference on*, pp. 1–6, IEEE, 2014.

58. Danger, R., F. Pla, A. Molina and P. Rosso, "Towards a Protein–Protein Interaction information extraction system: Recognizing named entities", *Knowledge-Based Systems*, Vol. 57, pp. 104–118, 2014.

59. Matos, S., D. Campos and J. L. Oliveira, "Vector-space models and terminologies in gene normalization and document classification", *Proceedings of the BioCreative III Workshop*, pp. 119–124, Citeseer, 2010.

60. Schneider, G., S. Clematide and F. Rinaldi, "Detection of interaction articles and experimental methods in biomedical literature", *BMC bioinformatics*, Vol. 12, No. 8, p. 1, 2011.

61. Agarwal, S., F. Liu and H. Yu, "Simple and efficient machine learning frameworks for identifying protein-protein interaction relevant articles and experimental methods used to study the interactions", *BMC bioinformatics*, Vol. 12, No. Suppl 8, p. S10, 2011.

62. Wang, X., R. Rak, A. Restificar, C. Nobata, C. Rupp, R. T. B. Batista-Navarro, R. Nawaz and S. Ananiadou, "Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature", *BMC bioinformatics*, Vol. 12, No. Suppl 8, p. S11, 2011.

63. Rinaldi, F., T. Kappeler, K. Kaljurand, G. Schneider, M. Klenner, S. Clematide, M. Hess, J.-M. Von Allmen, P. Parisot, M. Romacker *et al.*, "Ontogene in biocreative ii", *Genome Biology*, Vol. 9, No. Suppl 2, p. S13, 2008.

64. Ehrler, F., J. Gobeill, I. Tbahriti and P. Ruch, "GeneTeam site report for BioCreative II: Customizing a simple toolkit for text mining in molecular biology", *Proc of the Second BioCreative Challenge Evaluation Workshop: Madrid, Spain*, pp. 199–207, 2007.

65. Wang, H., M. Huang and X. Zhu, "Extract interaction detection methods from the biological literature", *BMC bioinformatics*, Vol. 10, No. 1, p. 1, 2009.

66. Krallinger, M., M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli *et al.*, "The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text", *BMC bioinformatics*, Vol. 12, No. Suppl 8, p. S3, 2011.

67. Kolchinsky, A., A. Lourenço, H.-Y. Wu, L. Li and L. M. Rocha, "Extraction

of Pharmacokinetic Evidence of Drug–Drug Interactions from the Literature", *PLoS ONE*, Vol. 10, No. 5, pp. 1–24, 05 2015, http://dx.doi.org/10.1371%2Fjournal.pone.0122199, accessed at August 2016.

68. Hur, J., A. Özgür, Z. Xiang and Y. He, "Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions", *Journal of biomedical semantics*, Vol. 6, No. 1, p. 1, 2015.

69. "The Clinical Decision Support (CDS) Track of the Text Retrieval Conference (TREC 2014)", http://trec-cds.appspot.com/2014.html, accessed at August 2016.

70. Roberts, K., M. Simpson, D. Demner-Fushman, E. Voorhees and W. Hersh, "State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track", *Information Retrieval Journal*, Vol. 19, No. 1-2, pp. 113–148, 2016.

71. Chen, R.-C., D. Spina, W. B. Croft, M. Sanderson and F. Scholer, "Harnessing semantics for answer sentence retrieval", *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval*, pp. 21–27, ACM, 2015.

72. Diaz, F., B. Mitra and N. Craswell, "Query Expansion with Locally-Trained Word Embeddings", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL – Association for Computational Linguistics, June 2016, https://www.microsoft.com/en-us/research/publication/query-expansion-with-locally-trained-word-embeddings/, accessed at August 2016.

73. Grbovic, M., N. Djuric, V. Radosavljevic and N. Bhamidipati, "Search retargeting using directed query embeddings", *Proceedings of the 24th International Conference on World Wide Web*, pp. 37–38, ACM, 2015.

74. Yagcioglu, S., E. Erdem, A. Erdem and Çakıcı Ruket, "A Distributed Representation Based Query Expansion Approach for Image Captioning", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing*, ACL, 2015.

75. "Published Data Set of this Study", `https://github.com/ferhtaydn/biocemid/tree/thesis/files/published_dataset`, accessed at August 2016.

76. "List of 103 Experimental Methods Focused in this Study", `https://github.com/ferhtaydn/biocemid/blob/thesis/files/103_methods.txt`, accessed at August 2016.

77. "BioC Java Library - Version 1.0.1", `https://sourceforge.net/projects/bioc/files/BioC_Java_1.0.1.tar.gz/download`, accessed at August 2016.

78. "Stanford CoreNLP Toolkit", `http://stanfordnlp.github.io/CoreNLP/index.html`, accessed at August 2016.

79. "Stanford CoreNLP Toolkit - Version 3.6.0", `http://nlp.stanford.edu/software/stanford-corenlp-full-2015-12-09.zip`, accessed at August 2016.

80. Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit", *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014, `http://www.aclweb.org/anthology/P/P14/P14-5010`, accessed at August 2016.

81. Lan, M., C. L. Tan, J. Su and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 31, No. 4, pp. 721–735, 2009.

82. "The PMC Open Access Subset of journal literature at U.S. National Institutes

of Health's National Library of Medicine", http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/, 2015, accessed at August 2016.

83. "Data Cleaning Code of this Study", https://github.com/ferhtaydn/stopword_remover/tree/thesis, accessed at August 2016.

84. "Word2Phrase component of Word2vec Tool", http://word2vec.googlecode.com/svn/trunk/word2phrase.c, accessed at August 2016.

85. Morin, F. and Y. Bengio, "Hierarchical Probabilistic Neural Network Language Model", R. G. Cowell and Z. Ghahramani (Editors), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 246–252, Society for Artificial Intelligence and Statistics, 2005.

86. Mnih, A. and G. E. Hinton, "A scalable hierarchical distributed language model", *Advances in neural information processing systems*, pp. 1081–1088, 2009.

87. "Modified version of Word2vec's 'distance' component for this Study", https://github.com/ferhtaydn/word2vec_extension/blob/thesis/distance_files.c, accessed at August 2016.

88. Jaccard, P., "Nouvelles recherches sur la distribution florale.", *Bulletin de la Société Vaudoise des Sciences Naturelles*, Vol. 44, pp. 223–270, 1908.

89. Cole, M., C. Nolte and W. Werr, "Nuclear import of the transcription factor SHOOT MERISTEMLESS depends on heterodimerization with BLH proteins expressed in discrete sub-domains of the shoot apical meristem of Arabidopsis thaliana", *Nucleic acids research*, Vol. 34, No. 4, pp. 1281–1292, 2006.

90. "Article with Pubmed Id 16513846 annotated for this Study", https://github.com/ferhtaydn/biocemid/blob/thesis/files/published_dataset/16513846.xml, accessed at August 2016.

91. Hripcsak, G. and A. S. Rothschild, "Agreement, the f-measure, and reliability in information retrieval", *Journal of the American Medical Informatics Association*, Vol. 12, No. 3, pp. 296–298, 2005.

92. Zhang, Z., S. Chapman and F. Ciravegna, "A methodology towards effective and efficient manual document annotation: addressing annotator discrepancy and annotation quality", *International Conference on Knowledge Engineering and Knowledge Management*, pp. 301–315, Springer, 2010.

93. "Genia Tagger - Version 3.0.2", `http://www.nactem.ac.uk/tsujii/GENIA/tagger/`, accessed at August 2016.

94. Tsuruoka, Y., Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou and J. Tsujii, "Developing a robust part-of-speech tagger for biomedical text", *Panhellenic Conference on Informatics*, pp. 382–392, Springer, 2005.

95. "Interaction Network Ontology (INO)", `http://www.ino-ontology.org/`, `http://bioportal.bioontology.org/ontologies/INO`, accessed at August 2016.

96. Özgür, A., Z. Xiang, D. R. Radev and Y. He, "Mining of vaccine-associated IFN-γ gene interaction networks using the Vaccine Ontology", *Journal of biomedical semantics*, Vol. 2, No. 2, p. 1, 2011.

97. "Literature Mining Keywords Record in Interaction Network Ontology (INO)", `http://purl.obolibrary.org/obo/INO_0000006`, accessed at August 2016.

98. "Manually selected Literature Mining Keywords Record in Interaction Network Ontology (INO) for this Study", `https://github.com/ferhtaydn/biocemid/blob/thesis/files/ino/literature_mining_keywords_related_to_proteins.txt`, accessed at August 2016.

99. McNemar, Q., "Note on the sampling error of the difference between correlated proportions or percentages", *Psychometrika*, Vol. 12, No. 2, pp. 153–157, 1947.

100. “McNemar’s Test for Results”, https://github.com/ferhtaydn/biocemid/blob/thesis/files/results/mcnamer.md, accessed at November 2016.

101. “The list of experimental methods which have word embeddings in this study”, https://github.com/ferhtaydn/biocemid/tree/thesis/files/oa_word2vecs_pure_baseline, accessed at August 2016.