Event-centric Document Similarity for Biomedical Literature

Brita Keller
Institute of Computer Science
Heidelberg University
Heidelberg, Germany
bkeller5@stud.uni-heidelberg.de

Jannik Strötgen
Institute of Computer Science
Heidelberg University
Heidelberg, Germany
stroetgen@uni-hd.de

Michael Gertz
Institute of Computer Science
Heidelberg University
Heidelberg, Germany
gertz@uni-hd.de

Abstract

Identifying similar documents for a given query document helps users to explore large document collections. However, most existing techniques are based on the vector space model and handle documents only as bags of words. Thus, more complex information that can be used for calculating similarities is not taken into account. For example, events play an important role in the biomedical literature and could be valuable to identify similar documents. In this paper, we present an event-centric document similarity model for biomedical literature and demonstrate the effectiveness of our approach based on experiments using the GENIA corpus.

1 Introduction

In the biomedical domain, events such as protein-protein interactions play an important role. Thus, there is a lot of research on automatically extracting biomedical entities and events from documents. For instance, there have been research challenges such as BioCreative and the bioNLP shared tasks 2009 and 2011. While BioCreative 2004 (Hirschman et al., 2005) concentrated on the extraction and normalization of entities and functional annotations, the subsequent BioCreative challenges (Krallinger et al., 2008; Leitner et al., 2010; Arighi et al., 2011) and the bioNLP shared tasks (Kim et al., 2009; Kim et al., 2011) furthermore addressed protein-protein-interaction and event extraction tasks.

In addition, due to the rapidly increasing number of publications in the biomedical domain – more than 850,000 citations were added to PubMed in 2011 (PubMed, 2012) – there is an increasing need for methods to explore large document collections and to easily access the embedded information. One such method that helps users to search in and explore large document collections like PubMed is to organize documents with respect to their similarity. While there are many methods to calculate the similarities of documents, most of them are based on the vector space model with documents being viewed as bags of words. More complex information such as biomedical events can hardly be included in the process for calculating document similarity. An example of a typical approach to identifying similar documents is to retrieve related citations for a given document directly provided by the PubMed interface[1]. Although the MeSH (medical subject heading) indexing terms are used in addition to the words of the documents, a lot of information is still not utilized for calculating similarities, e.g., information about the events mentioned in the documents.

Further shortcomings of bag of word-based techniques in general are that they cannot deal with ambiguity issues such as synonymy and polysemy of entity names and relation types, a further important issue in biomedical publications. For example, the two events (a) *overexpressing NF-IL-6* and (b) *C/EBPbeta expression* are very similar with *NF-IL-6* and *C/EBPbeta* being synonyms and *overexpressing* and *expression* both being descriptions for gene expressions. However, the similarity between these two events cannot be discovered using word-based methods. Assuming two documents containing (a)

[1] PubMed Related Citations Algorithm, http://ii.nlm.nih.gov/MTI/related.shtml

and (b), respectively, similarities between the documents can only be identified if the used similarity model addresses these shortcomings, e.g., by taking event information directly into account.

Recently, a model for calculating event-centric document similarities has been suggested for standard language documents such as Wikipedia articles (Strötgen et al., 2011). In this paper, we adapt this model to the biomedical domain. We present a novel approach to explore biomedical document collections in terms of the similarity of documents based on the events described in the documents. As we will show in our evaluation, our model identifies document similarities that cannot be identified by standard models and thus could be used as an alternative or complementary to existing models. Note that although events of the original model are defined as pairs of temporal and spatial information extracted from documents, they share key characteristics, which are crucial for the event-centric similarity model, with biomedical events consisting of one or more entities (e.g., proteins) and an event type (e.g., binding): (i) the components of both types of events can be normalized, and (ii) all components can be organized hierarchically. However, in contrast to spatio-temporal events, biomedical events can consist of more than two components and are much more complex. Thus, far-reaching adaptations have to be made to the original approach to calculate event-centric document similarities for biomedical documents.

The remainder of the paper is structured as follows. After surveying related work, we describe the original spatio-temporal event similarity model in Section 3. In Section 4, we present our adaptations and the event-centric similarity model for biomedical documents. In Section 5, we evaluate our approach and compare the results to standard similarity models. We conclude the paper in Section 6.

2 Related Work

There are many approaches to the computation of document similarity, among them the three classic models (set-theoretic models, algebraic models, probabilistic models) (Baeza-Yates and Ribeiro-Neto, 1999). Set-theoretic models such as the standard Boolean model treat documents as sets of words and phrases, and the similarities are computed by using set-theoretic operations on these sets. In algebraic models like the Vector Space Model both documents and queries are represented as vectors while a scalar value is used to represent the similarity of the query vector and the document vector. Numerous models based on and extending the Vector Space Model have been developed, e.g., Latent Semantic Indexing (LSI) (Deerwester et al., 1990). LSI uses singular value decomposition to analyze conceptual contents. In probabilistic models such as the binary independence retrieval model similarities are determined by computing the probabilities that a document is relevant for a given query.

Considering the rapidly growing number of digitalized publications in meta-databases like PubMed, gaining better and up-to-date access to similar documents is clearly of importance. One approach uses noun-phrases derived by the sentences of a paper for navigating biomedical literature on PubMed (Srikrishna and Coram, 2011). By associating a paper with its citations, the navigation of PubMed results becomes more transparent. Another approach enriches PubMed with sentence level co-citations (SLCs) based on citations within a single sentence, assuming that articles with a smaller citation distance in the same paper are more related (Tran et al., 2009). Lin and Wilbur presented a probabilistic topic-based model for content similarity underlying the related article search feature in PubMed that tries to determine "relatedness" rather than to estimate relevance like previous probabilistic retrieval models (Lin and Wilbur, 2007). Additionally, many Web-tools have been developed to enable a quick and efficient search and to retrieve relevant publications (Lu, 2011), e.g., RefMed and MedlineRanker.

An approach not limited to the biomedical domain or the PubMed database is context-aware citation recommendation (He et al., 2010). This system can be used for bibliography recommendations or for a ranked set of relevant citations to a specific placeholder using the words in its neighborhood as local context or the title and abstract as global context.

In contrast to our proposed event-centric document similarity model, none of these approaches considers deep semantics such as textually described events to calculate similarities between documents, since the text is treated as bag of words, disregarding grammar and word order.

3 Spatio-temporal Event Similarity

In this section, we describe the original event-centric similarity model before introducing our adaptation to the biomedical domain in Section 4. In contrast to standard vector space based models, the event-centric document similarity model does not handle documents as bags of words, but uses information about the events extracted from the documents.

3.1 Spatio-temporal Events

Motivated by the fact that events usually occur at a specific time and place, Strötgen et al. (2011) simply define an event as a combination of a temporal and a geographic expression if they co-occur in the same sentence. Both types of expressions can be extracted from documents using temporal taggers and geo-taggers, respectively. After extracting these co-occurrences, every document d is represented as a set of events in the form of so called *document event profiles*, denoted $dep(d)$, containing events of the form $e_i = \langle t_i, g_i \rangle$, with t_i being a temporal expression and g_i being a (geo)spatial expression (Strötgen et al., 2011). However, instead of just using the expressions referring to the time and place of an event, the similarity model is based on two key characteristics of geographic and temporal expressions:

- Both types of information can be normalized to some standard format. Thus, two expressions referring to the same point in time or location have the same value in the standard format.
- Both types of information can be organized hierarchically due to the different granularities of temporal and spatial information.

Based on these characteristics, it is possible to compare two events using so-called temporal and geographic *mappings*. One temporal (geographic) mapping step maps a temporal (geographic) expression to the next coarser granularity. Thus, when comparing two events, the geographic and the temporal components of both events are either identical, or they can be mapped to coarser granularities until they are equal, or are unequal if the highest level of the hierarchy is reached before both components match. For example, assuming the granularities day, month, and year for the temporal hierarchy, and city and country for the geographic hierarchy, we can compare the two events $e_1 = \langle$2012-09-03, Zurich-Switzerland$\rangle$ and $e_2 = \langle$2012-09, Basel-Switzerland$\rangle$ in the following way:

- $map_t(e_1(t)) = e_2(t)$
- $map_g(e_1(g)) = map_g(e_2(g))$

Thus, one temporal (map_t) and two geographic mapping steps (map_g) have to be applied to make e_1 and e_2 match each other. Furthermore, one temporal and two geographic values had to be mapped. Using the total number of mapping steps (α), the maximum number of values per dimension involved in the mapping process (β), and the number of mappings, which are still possible after the mapping process (α_{poss}), in Strötgen et al. (2011) the similarity between the two events is calculated as follows:

$$sim_e(e_1, e_2) := \frac{1}{(1+\alpha)^\beta} (\alpha_{poss} + 1) \quad (1)$$

Using this similarity function, several intuitive requirements for event similarity are satisfied. For example, the less mappings are needed (due to α, β) and the more fine-grained the events are (due to α_{poss}), the higher $sim_e(e_1, e_2)$. When describing the similarity function for biomedical events in Section 4, we will further detail the characteristics of sim_e. In the next section, we describe how such event similarities can be aggregated to compare sets of events, i.e., documents, in the original event-centric document similarity model.

3.2 Event-centric Document Similarity

For calculating similarities between two documents d_1 and d_2, Strötgen et al. (2011) build the cross-product of the m events in the document event profile $dep(d_1)$ and the n events in $dep(d_2)$ and calculate the event similarity for every event pair as described in the previous section. These event similarities are aggregated and, in addition, a cardinality normalization is performed:

$$sim_e(d_1, d_2) := \frac{\sum_{i=0}^{m} \sum_{j=0}^{n} sim_e(e_i, e_j)}{min\{m, n\}} \quad (2)$$

This calculation satisfies several requirements. For example, the more matching events there are in d_1 and d_2, the higher $sim_e(d_1, d_2)$. Since these requirements are similar to our requirements for calculating

biomedical event-centric document similarities, we do not further discuss them here, but detail them in the next section.

4 Biomedical Event Similarity

Our proposed approach for calculating similarities between biomedical events and biomedical documents containing such events is directly based on the similarity model described in the previous section. In this section, we detail our adaptations to the model to fit the biomedical domain.

4.1 Biomedical Events

In contrast to the simple definition of spatio-temporal events, biomedical events are much more complex. In addition, their heterogeneous structure with differing numbers of event components makes it more difficult to compare two biomedical events than two spatio-temporal events.

For our approach, we use the GENIA definition of a biomedical event (Kim et al., 2006). An event can thus be defined as having up to two *"themes"* and/or up to two *"causes"* and one *"event-type"*. Formally, $e_i = \langle t_i^{(1)}, t_i^{(2)}, c_i^{(1)}, c_i^{(2)}, et_i \rangle$ with a *theme* $(t^{(1)}, t^{(2)})$ containing a biological entity whose properties are changed by an event, and a *cause* $(c^{(1)}, c^{(2)})$ containing a biological entity, which affects the way of occurrence of an event. The *"event-type"* (et) represents the biomedical relationship (e.g., binding or phosphorylation). In the GENIA corpus, only dynamic relationships are annotated, i.e., at least one biological entity of a relationship has to be altered by the occurring event regarding its properties or location to qualify for an annotation (Ohta et al., 2006). Note that other definitions of biomedical events could be used with our model as long as the following characteristics of the events are satisfied:

- All components of an event can be normalized.
- All components of an event can be associated with concepts in a hierarchy.

The GENIA event annotation is based on two ontologies, the GENIA term ontology and the GENIA event ontology (Kim et al., 2008). The term ontology consists of biological entities (e.g., proteins, DNA, RNA), which can be categorized as either *themes* or *causes*. The event ontology contains biological processes and molecular functions (e.g., Positive_regulation, DNA_modification), i.e., the *event-types*.

Since both GENIA ontologies are organized hierarchically, they can be used to calculate similarities of events in the same way as for spatio-temporal events described in the previous section. Accordingly, we use mapping functions to map the event components (map_t for *themes*, map_c for *causes*, and map_{et} for *event-types*) to the next coarser granularity in the hierarchies. The ambiguity of biomedical events can be handled by normalizing the event-types and the biomedical entities using NER tools such as ProMiner (Hanisch et al., 2005) or Geno (Wermter et al., 2009). Finally, the events can be extracted from documents using systems such as those that participated in the bioNLP shared task on event extraction (Kim et al., 2011). Thus, due to the hierarchical organization and the possibility of normalizing event-types and entities, both requirements for adapting the original model to the biomedical domain are satisfied. Therefore, documents can be represented using document event profiles containing events of the form $e_i = \langle t_i^{(1)}, t_i^{(2)}, c_i^{(1)}, c_i^{(2)}, et_i \rangle$.

Note that although biomedical events could be extracted automatically from the documents, we use the events annotated in the GENIA corpus to demonstrate our approach. However, not all events annotated in the GENIA corpus are suitable for our event-centric document similarity model. In particular, only events that satisfy the following requirements become part of a document event profile: (i) all event components have to be part of either the term or the event hierarchy, (ii) a *clue-type* has to be annotated in the document to describe the *event-type*. In addition, to keep the similarity model as simple as possible, we do not consider nested events, whose *theme(s)* and/or *cause(s)* consist of other events. Finally, there are some annotation errors in the corpus (e.g., terms are not part of the terminal classes of the hierarchy). Thus, in total, the modified GENIA corpus contains 12,873 events distributed over 997 documents.

4.2 Similarities between Biomedical Events

Since all entities of the events in the GENIA corpus are of the finest granularity, all entities in the term

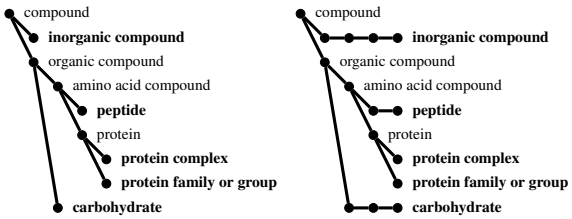


Figure 1: Parts of the original GENIA ontology (left) and adapted ontology with additional nodes for virtual classes (right).

hierarchy are terminal classes. Thus, every mapping from one entity does not result in another entity, but in a non-terminal class. However, the term and event hierarchies are not balanced. To treat all entities in a similar way and to avoid problems with calculating similarities between events, all terminal classes were moved to the same depth in the hierarchies by introducing empty virtual classes (see Figure 1). All annotated terms remain unchanged in the terminal classes, but now a fixed maximum number of mapping steps from the leaves to the roots of the hierarchies is possible.

Like in the original model, the event similarity algorithm takes two events e_1 and e_2 and does a pairwise comparison of their components (*theme(s)*, *cause(s)*, *event-type*). If no equality exists between two components, they will both be mapped to coarser granularities in the corresponding hierarchy until a match is found. Obviously, the more distant two elements are in the hierarchy, the more mapping steps are needed to achieve equality and the lower the resulting similarity score $sim_e(e_1, e_2)$. The procedure is quite straightforward using uniformly structured events with not only the same number of components but also the same type of components. However, for every event $e_i = \langle t_i^{(1)}, t_i^{(2)}, c_i^{(1)}, c_i^{(2)}, et_i \rangle$, only the *event-type* et_i is mandatory. In addition, not all possible variants of components occur in the corpus. Thus, calculating the event similarity sim_e between two biomedical events is much more challenging than between two spatio-temporal events.

In contrast to the original model, a mapping to equality between two events with the same structure is always possible due to the presence of root elements in both hierarchies, although a mapping to the highest level will be penalized and thus results in a low similarity score. However, two events do not have to consist of the same number and types of components. Thus, the similarity function has to be extended so that two non-uniform events can be compared, too. Intuitively, two events with different numbers and types of components should result in a lower similarity score than two events with the same types of components. For this, we introduce a new parameter capturing the number of incomplete element-pairs (γ), which can be used together with α (number of mapping steps), β (maximum number of mapped values per component), and α_{poss} (possible mappings after mapping process) to calculate the similarity score $sim_e(e_1, e_2)$.

Before adapting the formula $sim_e(e_1, e_2)$ to compare biomedical events (cf. Equation 1, Section 3.1), we first list important requirements that should be satisfied by the new similarity function:

R1 *The more similar e_1 and e_2, the higher sim_e.*

R2 *The fewer mapping steps are needed, the higher sim_e.*

R3 *The more similar the component-pairs, the higher sim_e.*

R4 *The more incomplete component-pairs, the lower sim_e.*

Based on these requirements, we adapted sim_e as follows:

$$sim_e(e_1, e_2) := \frac{1}{(1+\alpha)^{\beta\,(\gamma+1)}}\,(\alpha_{poss}+1) \quad (3)$$

This similarity function returns the highest score ($max(sim(e_1, e_2)) = 71$ due to the GENIA hierarchies) if two events contain all components and all components directly match without any mapping (R1). Furthermore, the lower α, the higher sim_e (R2). Component-pairs that have to be mapped are additionally penalized since in the case of a mismatch both values have to be mapped, i.e., $\beta = 2$, which lowers sim_e. In addition, the more similar a component-pair, the higher α_{poss} and thus the higher sim_e (R3). Finally, incomplete component-pairs are heavily penalized by $\gamma + 1$, especially if the other components are not equal in both events (R4).

Table 1 shows an example of two events being mapped to equality: $e_1 = \langle {t_1}^{(1)}, et_1 \rangle$ ("activation of

d_h	$t_1^{(1)}$	$t_2^{(1)}$	α_1	β_1	α_{poss_1}
-	NF-kappa B	MAP Kinase	0	0	14
6	Protein _complex	Protein_family _or_group	2	2	12
5*	Protein	Protein	4	2	10
d_h	et_1	et_2	α_2	β_2	α_{poss_2}
-*	activation	activation	0	0	14

Table 1: Mapping of themes and event-types of e_1 and e_2. d_h: current depth in hierarchy ($d_h = 6$ at leaf- and $d_h = 0$ at root level). * indicates reached equality.

NF-kappa B") and $e_2 = \langle {t_2}^{(1)}, et_2 \rangle$ ("activation of MAP kinase"). The comparison of the two *themes* starts at *clue-type*-level with $\alpha_{poss_1} = 14$ with seven possible mapping steps per theme. "NF-kappa B" and "MAP kinase" are not an exact match. Therefore, the two corresponding leaf classes of the hierarchy are compared ($\alpha_1 = 2$, $\beta_1 = 2$). Still not matching, mapping to the next level leads to $\alpha_1 = 4$ and results in equality at the fifth level of the hierarchy in the Protein class (cf. Figure 1). The two *event-types* match directly on *clue-type*-level. Therefore, α_2, β_2 and α_{poss_2} stay unchanged, leading to $\alpha = 4$, $\beta = 2$, and $\gamma = 0$ (same type and number of components in both events) and $\alpha_{poss} = 24$. Using Equation 3 the event similarity thus is $sim_e(e_1, e_2) = 1$.

4.3 "Bio-event-centric" Document Similarity

With the event-similarity equation in place, we can now calculate the similarity of two documents d_1 and d_2 containing biomedical events by comparing their document event profiles, i.e., two sets of events. The requirements for the document similarity stay the same as in the original model:

D1 *The more matching events there are in d_1 and d_2, the higher $sim_e(d_1, d_2)$.*

D2 *The more non-matching events there are in d_1 and d_2, the lower $sim_e(d_1, d_2)$.*

D3 *If only one document contains additional events, this should not be penalized as much as if both documents contain additional non-matching events.*

D4 *The more similar the events in d_1 and d_2, the higher $sim_e(d_1, d_2)$.*

Thus, we can use the original equation (Equation 2) to compare two documents with using Equation 3 as $sim_e(e_i, e_j)$:

$$sim_e(d_1, d_2) := \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} sim_e(e_i, e_j)}{min\{m, n\}} \quad (4)$$

Although Equation 4 is used in the original model and was demonstrated to work well, we propose an additional way of combining individual event similarities. Since aggregating the similarities of the cross-product of all events in both document event profiles and normalizing the similarity score only by the minimum number of events in the event profiles ($min\{m, n\}$), this similarity function prefers documents containing long documents – especially if mapping is performed up to the root level, as done in our case. Assuming a document d_1 with m events. Then, documents containing $m + x_1$ events result in higher similarity scores than those containing $m+x_2$ events ($x_1 > x_2$) even if there are many non-similar and only few similar events among the additional x_1 events. To avoid preferring long documents, we propose the following alternative equation:

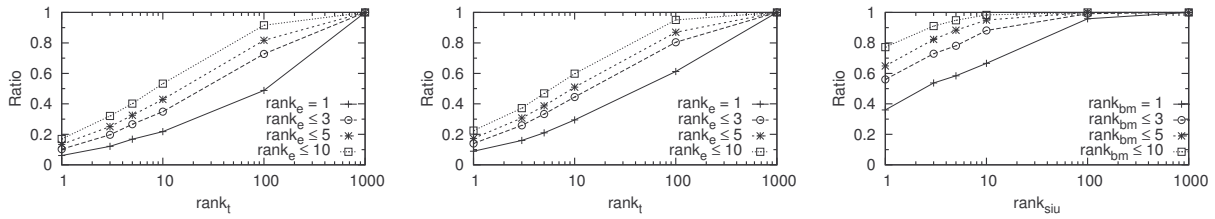
$$sim_e(d_1, d_2) := \sqrt{\frac{1}{m+n} \sum_{i=1}^{m+n} {bestScore_i}^2} \quad (5)$$

In contrast to Equation 4, only the highest similarity score $bestScore_i$ of every event i in d_1 and d_2, i.e., $m+n$ scores, are part of the sum; all other similarity scores of event pairs are discarded. By using the root mean square, larger scores have more impact on the result than smaller ones, and partial matches are punished more than one single mismatch.

In our evaluation described in the next section, we use both equations, compare their results, and compare our event-centric document similarity approach for biomedical literature with a standard document similarity method.

5 Evaluation

In this section, we describe the evaluation results of our event-centric document similarity approach on the GENIA corpus. Since the entities in the GENIA corpus are not normalized, we performed a simple



(a) Ratio of documents that are ranked top n according to sim_t for different $rank_{e-siu}$ values.

(b) Ratio of documents that are ranked top n according to sim_t for different $rank_{e-bm}$ values.

(c) Ratio of documents that are ranked top n according to sim_{siu} for different $rank_{bm}$ values.

Figure 2: Comparison between similarity measures.

normalization based on the dictionary of the Moara project (Neves et al., 2010) trying to map all synonyms to the same corresponding Entrez Gene IDs.

5.1 Comparison to Standard Document Similarities

For our evaluation, we ran both aggregation strategies, the sum-it-up-strategy (siu) using equations (3) and (4) and the best-match-strategy (bm) using equations (3) and (5). In contrast to standard document similarity models, our model is not term- but event-based. For this, we compare the results with a standard similarity measure (tf-idf with cosine similarity: sim_t). In Figure 2, we compare sim_{e-siu} with sim_t (a) and sim_{e-bm} with sim_t (b). Using the most similar 1, 3, 5, and 10 ranked documents for a query document, we calculate the ratio of how often the same documents are within the top-n most similar documents with respect to sim_t. Both event-centric models identify other documents as being similar than the term-based model does. For example, only for about 20% (30%) of the most similar documents with respect to sim_{e-siu} (sim_{e-bm}), the same document is within the most similar 10 documents using the term-based similarity model.

In contrast, as shown in Figure 2(c), the results of sim_{e-siu} and sim_{e-bm} are very similar and there is often only a slightly different ranking order. For example, 90% of the documents that are within the 5 most similar documents for a given query document using sim_{e-siu} are within the 5 most similar documents using sim_{e-bm}.

5.2 Manual Evaluation

To demonstrate that the similarities identified by the event-centric document similarity model are also valid and not just different from those identified by term-based models, we manually evaluated whether document pairs that are identified as similar using sim_{e-siu} and sim_{e-bm} are indeed similar. For this, we randomly selected five query documents and compared them to their five most similar documents, i.e., we analyzed if the documents had the same main topics with respect to the described events.

For the siu-strategy and the bm-strategy, we found a close similarity to the query document in 56% (64%) of the documents. While some of the other documents could be regarded as similar with deep domain knowledge, we only rated these documents to be similar, for which the event-similarity was obvious. In addition, we found that there were many documents that did not have 5 very similar documents with respect to the described events. Note, however, that the corpus consists only of 1000 documents, and when calculating event-centric document similarities on a larger corpus, the chance for more documents describing similar events is much higher.

6 Conclusions and Ongoing Work

In this paper, we presented a novel approach to determine documents similar to a given query document based on the event information extracted from the documents. The key idea is that components of events (themes, causes, event-type) can be associated with hierarchies, which enable the effective computation of similarity scores for pairs of events and their components, respectively, as well as documents containing sets of events. We are currently extending the model to consider more context information about extracted events, beyond sentence level. We are also applying our technique for different document clustering and classification approaches.

References

Cecilia N Arighi, Zhiyong Lu, Martin Krallinger, Kevin B Cohen, W J Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy H Wu. 2011. Overview of the BioCreative III Workshop. *BMC Bioinformatics*, 12(Suppl 8):S1.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press Books.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407.

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. 2005. ProMiner: Rule-based Protein and Gene Entity Recognition. *BMC Bioinformatics*, 6(Suppl 1):S14.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware Citation Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 421–430. ACM.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of BioCreAtIvE: Critical Assessment of Information Extraction for Biology. *BMC Bioinformatics*, 6(Suppl 1):S1.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. 2006. Genia Corpus Manual – Encoding Schemes for the Corpus and Annotation. Technical report, University of Tokyo, Kogakuin University, University of Manchester, National Centre for Text Mining.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus Annotation for Mining Biomedical Events from Literature. *BMC Bioinformatics*.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 1–9. ACL.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the Protein-Protein Interaction Annotation Extraction Task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.

Florian Leitner, Scott A. Mardis, Martin Krallinger, Gianni Cesareni, Lynette A. Hirschman, and Alfonso Valencia. 2010. An Overview of BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):385–399.

Jimmy Lin and W John Wilbur. 2007. PubMed Related Articles: A Probabilistic Topic-based Model for Content Similarity. *BMC Bioinformatics*, 8:423.

Zhiyong Lu. 2011. Pubmed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. *Database (Oxford)*, baq036.

Moara Project. 2012. `http://moara.dacya.ucm.es/index.html`. Website. [Online; accessed 10-August-2011].

Mariana Neves, Jose-Maria Carazo, and Alberto Pascual-Montano. 2010. Moara: A Java Library for Extracting and Normalizing Gene and Protein Mentions. *BMC Bioinformatics*, 11(1):157.

Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2006. Guidelines for Event Annotation. Technical report, University of Tokyo.

PubMed. 2012. `http://www.ncbi.nlm.nih.gov/pubmed/`. Website. [Online; accessed May 23, 2011].

Devabhaktuni Srikrishna and Marc A. Coram. 2011. Using Noun Phrases for Navigating Biomedical Literature on Pubmed: How Many Updates Are We Losing Track of? *Plos One*, 6(9):e24920.

Jannik Strötgen, Michael Gertz, and Conny Junghans. 2011. An Event-centric Model for Multilingual Document Similarity. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 953–962. ACM.

Nam Tran, Pedro Alves, Shuangge Ma, and Michael Krauthammer. 2009. Enriching PubMed Related Article Search with Sentence Level Co-citations. In *AMIA Annual Symposium Proceedings*, pages 650-654.

Joachim Wermter, Katrin Tomanek, and Udo Hahn. 2009. High-performance Gene Name Normalization with GENO. *Bioinformatics*, 25(6):815–821.