Biomedical concept extraction based on combining the content-based and word order similarities

Duy Dinh, Lynda Tamine

Biomedical concept extraction based on combining the content-based and word order similarities

Duy Dinh
University of Toulouse
118 route de Narbonne
Toulouse, France
dinh@irit.fr

Lynda Tamine
University of Toulouse
118 route de Narbonne
Toulouse, France
tamine@irit.fr

ABSTRACT

It is well known that the main objective of conceptual retrieval models is to go beyond simple term matching by relaxing term independence assumption through concept recognition. In this paper, we present an approach of semantic indexing and retrieval of biomedical documents through the process of identifying domain concepts extracted from the Medical Subject Headings (MeSH) thesaurus. Our indexing approach relies on a purely statistical vector space model, which represents medical documents and MeSH concepts as term vectors. By leveraging a combination of the bag-of-word concept representation and word positions in the textual features, we demonstrate that our mapping method is able to extract valuable concepts from documents. The output of this semantic mapping serves as the input to our relevance document scoring in response to a query. Experiments on the OHSUMED collection show that our semantic indexing method significantly outperforms state-of-art baselines that employ word or term statistics.

Keywords

Concept Recognition, Semantic Indexing, Document Expansion, Biomedical Information Retrieval

1. INTRODUCTION

Traditional Information Retrieval (IR) systems rely on matching keywords from queries to those from documents under the basic assumption of term independence. This leads to the well known *bag-of-word* based indexing and retrieval models that provide poor search results. Conceptual indexing and retrieval models, i.e., the use of concepts in ontologies or thesauri, is the extension of bag-of-word based models. The centerpiece of conceptual approaches is that independent words are not able to capture the document semantic content and that a suitable solution to this problem is to reach the conceptual level of information contents. Several types of knowledge could be exploited for deriving semantic document representations such as knowledge about the search task, knowledge about the problem, knowledge about the user's intent, knowledge about the domain, etc. We focus here on the use of this latter to derive the semantic kernels of biomedical documents. In the biomedical domain, such knowledge bases (e.g., Medical Subject Headings (MeSH), Unified Medical Language System (UMLS), International Classification of Diseases (ICD), etc.) exist and are so far maintained by several biomedical research groups. They entail generally domain concepts at various levels of specificity. In biomedical IR, there have been many works dealing with conceptual indexing by mapping free text to medical ontologies leading to the challenging problem of concept identification or extraction [1–7, 10].

In this paper, we propose a novel method for extracting key concepts from biomedical documents using the MeSH resource. More specifically, we use an IR-based approach [10] for both MeSH concept categorization and document relevance estimation. Our main contribution consists in representing the document's semantic kernel as the top relevant concepts extracted by measuring the concept relevance for the document. Our basic assumption behind concept relevance is that a list of document words is more likely to map a concept that (1) both shares a maximum number of words either among its preferred or non-preferred terms derived from all of its possible entries; (2) the words tend to appear in the same order so to cover the same meaning. Thus we propose to combine two features: *content-based similarity* and *word order similarity* between a document and each MeSH concept using the cosine measure and the Spearman rank correlation. We incorporate the semantic indexing approach into the semantic retrieval model by (1) adopting our semantic-based method to extract the most representative concepts from documents; (2) expanding documents with keywords extracted from identified concepts; (3) computing the document relevance score.

The remainder of this paper is organized as follows. Section 2 summarizes related work. Section 3 details our semantic document indexing and retrieval framework. Experiments and results are presented in section 4. Section 5 concludes the paper and outlines directions for future work.

2. RELATED WORK

Automatic concept extraction from medical text is a challenging task because of many reasons. First, terms representing biomedical concepts are usually comprised of multiple words that lead to a more specific meaning as a whole. For instance, "cancer of brain" is a neoplasm of the intracranial components of the central nervous system but the word

SAC'11 March 21-25, 2011, TaiChung, Taiwan.

"cancer" alone could be any type of malignant growth or tumour. Second, there are several synonymous terms applied to the same concept. For example, terms such as "avian flu", "avian influenza", "fowl plague"... can be used to indicate the concept "Influenza in Birds" as defined in MeSH. Third, many abbreviations/acronyms can be used to refer to different concepts, e.g, "APS" may refer to a *gene* or *protein* in UMLS. This work focuses on the two first challenging points of automatic concept extraction for indexing and document retrieval purposes. Automatic indexing can be basically referred to as the assignment of a number of *terms* denoting *concepts* to a document. Each concept is represented by a unique preferred term, used for indexing, and one or many non-preferred terms, used for retrieval. Intuitively speaking, the document concept extraction within a thesaurus can be defined as a classification problem as follows:

- *classify document D into a ranked list of n concepts* $C_{i1}, C_{i2}, \ldots, C_{in} \subset C_1, C_2, \ldots, C_N$ *where* $C_1, C_2, \ldots, C_N$ *is a set of semantic categories belonging to the thesaurus;*
- *the meaning of each semantic category is defined by the labels along the path from the root to the corresponding node in a poly-hierarchical structure*;
- *document D consists of textual features of words.*

Works on biomedical concept extraction have been extensively studied in the literature [1–7, 10]. Current approaches to concept extraction for identifying biomedical concepts can be subdivided into four categories: *rule-based* [4, 5], *machine learning* [2, 7], *dictionary-based* [1, 8] and *statistical* approaches [6, 10]. Rule-based approaches generally rely on formation patterns naming structures for specific concept classes using lexical and morphological properties [4,5]. Such approaches are known to be extremely time-consuming for development, and moreover their application to other entities is usually difficult. Machine learning (ML) methods use manually annotated corpus for training classifiers, which basically try to learn several features for binding terms from free text to predefined classes. The work cited in [2] use Hidden Markov models (HMM) and specific orthographic features for discovering terms belonging to a set of ten classes. Each term candidate was assigned a class of the most similar term from the training set, with respect to the orthographic similarity. Support vector machines (SVM) have been a powerful tool for supervised ML. The work cited in [7] trained multi-class SVMs on the manually annotated GENIA corpus for the task of named entity recognition. More precisely, their method aims at predicting composite tags indicating named entities based on position-dependent features (e.g., POS, prefix and suffix features), as well as a word cache capturing similarities of patterns with a common keyword, and HMM state features in order to address the data sparseness problem. However, ML methods are faced to some difficulties when training data are not available, e.g., instances of RNAs in the GENIA corpus. Dictionary-based methods for concept extraction use existing terminological resources to map free text to entries in a dictionary. MTI (Medical Text Indexer) [1] integrates several methods of concept recommendation for indexing MEDLINE documents. MTI uses a knowledge intensive approach based on symbolic, and computational linguistic techniques for identifying biomedical concepts. It provides at the first stage a list of UMLS concepts and then restricts to MeSH concepts, which are finally used to represent the semantics of the document. The work cited in [8] suggested a method based on an approximate string matching to recognize gene and protein names and their variations. In their approach, both protein dictionaries and target text are encoded using the nucleotide code (a four-letter encoding over the A, C, G, T alphabet). Statistical approaches have been proposed to address the recognition of general terms. For example, the work cited in [3] has proposed a method called *C/NC* value for recognizing technical terms used in Digital Libraries. It has been also used to recognize terms from biomedical literature [6]. The *C/NC* value is a domain-independent method combining statistical and linguistic information for the extraction of multi-word and nested terms. The work cited in [10] introduced a retrieval-based system for MeSH classification. For each MeSH term, its synonyms and description are indexed as a single document in a retrieval index. A piece of text, the query to the retrieval system, is classified with the best ranked MeSH documents.

Our work presented in this paper belongs to this last category of work. Our approach differs from previous works in two important ways: first, using an IR point of view, we estimate concept relevance for a document by combining document/query concept matching degree and document/query concept correlation, as clues for achieving concept extraction accuracy. Besides, we propose to expand the documents with words belonging to the extracted concepts.

3. OUR CONCEPTUAL INDEXING AND RETRIEVAL APPROACH

In this work, we design a conceptual indexing and retrieval framework that incorporate identified terms denoting concepts from MeSH in an attempt to highlight the subject matter(s) of documents for improving the IR effectiveness. Our system architecture consists of three main components detailed below: (1) MeSH categorization, (2) document-to-concept mapping and (3) Document expansion and retrieval.

3.1 MeSH categorization

We suppose that a MeSH concept can be thought of as a document containing biomedical terms. Indeed, in MeSH, each concept is described by a *main heading* (preferred term), one or many *concept entries* (non-preferred terms), qualifiers, scope notes, etc. Main headings and concept entries constitute together the most common indexing and retrieval features used in the domain.

Let's denote $Entries(C)$ the set of preferred and non-preferred terms denoting concept C. According to our approach, MeSH is viewed as a collection of textual concepts that can be indexed according to the vector space model [11]. Formally, each concept C is represented as a basic keyword vector: $C = (c_1, c_2, \ldots c_{N_c})$, where N_c is the number of unique words in MeSH, c_j is a weight measuring the aboutness of word w_j in C. We propose to adapt the BM25 weighting schema [9] for concept weighting as follows:

$$c_j = tfc_j * \frac{log\frac{N-n_j+0.5}{n_j+0.5}}{k_1 * ((1-b) + b\frac{cl}{avgcl}) + tfc_j} \quad (1)$$

where tfc_j is the number of occurrences of word w_j in concept C, N is the total number of concepts in MeSH the-

saurus, n_j is the number of concepts containing at least one occurrence of word w_j in its textual fields, cl is the length of concept C (i.e. total number of distinct words occurring in its textual features), and $avcl$ is the average concept length in MeSH thesaurus, k_1, and b are tuning parameters.

3.2 Document-to-concept mapping: how to extract key concepts?

In our approach, the document-to-concept mapping is formalized as an IR task. In other words, given a document, the mapping leads to the selection of the most relevant MeSH concepts using a content-based similarity measure. Furthermore, in order to take into account the importance of the word order while matching an entry to a bounded multi-word terms issued from a document, we propose to leverage the content-based similarity between document and concept using a rank correlation based matching. Our strategy, which is mainly based on ranking concepts extracted from documents using a combined score, involves three steps detailed below: (1) computing a content-based matching score, (2) computing a rank correlation based score, (3) selecting the document semantic kernel by ranking the concepts according to their combined score.

1. ***Computing a content-based matching score***. According to our IR based approach, the top-ranked relevant concepts issued from MeSH are assigned to the document. Formally, we compute for each concept vector C (*cf.* section 3.1) a content-based cosine similarity w.r.t the document D, denoted $Sim(C, D)$, as follows:

$$Sim(C, D) = \frac{\sum_{j=1}^{N_c} c_j * d_j}{\sqrt{\sum_{j=1}^{N_c} c_j^2} * \sqrt{\sum_{j=1}^{N_c} d_j^2}} \quad (2)$$

where N_c is the total number of concepts in MeSH, c_j is the weight of word w_j in concept C computed using formula 1, d_j is the weight of word w_j in document D computed using an appropriate weighting schema.

2. ***Computing a rank correlation coefficient***. The candidate concepts extracted from step 1 are re-ranked according to a correlation measure that estimates how much the word order of a MeSH entry is correlated to the order of words in the document. For this aim, we propose to measure the word order correlation between the concept entry and the document both represented by word position vectors. Formally, the correlation measure is computed using the Spearman operator as follows: let document $D = (w_{d_1}, w_{d_2}, \ldots, w_{d_L})$ be the ranked word based vector according to the average position of related occurrences in document D, i.e., w_{d_i} is the document word in D such that $p\bar{o}s(occs(w_{d_i})) < p\bar{o}s(occs(w_{d_{i+1}}))\ \forall i = 1 \ldots L-1$, where $occs(w_{d_i})$ is the set of occurrences of word w_{d_i} in document D, L is the total number of unique words in document D. Similarly, let $E = (w_{e_1}, w_{e_2}, \ldots, w_{e_{L'}})$ be the ranked word based vector according to the average position of related occurrences in concept entry E, where L' is the concept entry length. We denote the set of words in D as $words(D) = \{w_{d_1}, w_{d_2}, \ldots w_{d_L}\}$ and in concept entry E as $words(E) = \{w_{e_1}, w_{e_2}, \ldots, w_{e_{L'}}\}$.

First, in order to avoid false rank bias, when measuring the word order correlation, a portion of the document window bounded by the first and last word occurrences shared by the concept entry E and the document is captured and normalized as follows: $D_w = (w_{d_w}, w_{d_{w+1}}, \ldots, w_{d_W})$ where $words(D_w) \subset words(D)$, $w_{d_w} \in words(E), w_{d_{W+1}} \notin words(E)$. Afterwards, the Spearman correlation coefficient is used to compute the word rank correlation between words in document D and concept entry E:

$$\rho(E, D) = 1 - \frac{6 * \sum_i^T [rank(w_i, D_w) - rank(w_i, E)]^2}{T * (T^2 - 1)} \quad (3)$$

where $rank(w_i, D_w)$ (resp. $rank(w_i, E)$) is the word order or rank of word w_i according to $p\bar{o}s(occs(w_{d_i}))$ in D_w (resp. E), $T = |words\,(D_w) \cap (words\,(E))|$ is the number of shared words between document D and concept entry E. We simply assume that the rank of an absent word in D_w or E is assigned a default value $r_0 > T$. The coefficient $\rho(E, D)$ allows measuring the degree of agreement between two word rankings in E and D_w. If the agreement between two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1. If the disagreement between them is perfect (i.e., one ranking is the reverse of the other) the coefficient has value -1. For all other arrangements the value lies in between -1 and 1, and increasing values imply increasing agreement between them. In order to consider each significant entry separately, we practically compute:

$$\rho(C, D) = Max_{E \in Entries(C)}\ \rho(E, D) \quad (4)$$

where $Entries(C)$ refers to both preferred terms or non-preferred terms belonging to concept C.

3. ***Selecting the semantic document kernel***. Finally the content based similarity score and the correlation score between concept C and document D are combined in order to compute the overall relevance score $Rel(C, D)$ as follows:

$$Rel(C, D) = (1 + Sim(C, D)) * (1 + \rho(C, D)) \quad (5)$$

The N top-ranked concepts with the highest scores are selected as the semantic index kernel of document D. The optimal value of the parameter N is experimentally tuned.

3.3 Document expansion and retrieval

The document expansion stage aims at increasing the degree of word overlap between user queries and observed documents. Here, we use words figuring in the main entries for normalizing the document content in an attempt to resolve the synonymy problem. Next, our objective is to compute the relevance score of the expanded documents with respect to each query. We hypothesize that expanded terms denoting concepts are somehow less relevant than original terms in the document (namely hypothesis H) because terms identified using the IR approach may return some irrelevant information (noise) w.r.t the query, thus the score of a given term in the document D is computed as:

$$score(t, D) = \left\{ \begin{array}{ll} (1 - \alpha) * w_0(t, D) & \text{if } t \text{ is an expanded term} \\ w_0(t, D) & \text{otherwise} \end{array} \right\} \quad (6)$$

where $w_0(t, D)$ is the original document term score computed by the BM25 weighting model. $\alpha \in [0..1)$ is a decay factor by which the score of the expanded terms is reduced.

Finally, the relevance score of the document D with respect to the query Q, namely $RSV(Q, D)$, is given by:

$$RSV(Q, D) \quad = \quad \sum_{t_i \in Q} score(t_i, D) \quad (7)$$

where t_i is the query term, $score(t_i, D)$ is the final document term score computed using formula 6.

4. EXPERIMENTAL EVALUATION

4.1 Test Collection

We used the OHSUMED test collection, which is a MEDLINE subcollection used for biomedical IR, under the Terrier IR platform (`http://terrier.org/`). Each document has been annotated by human experts (physicians) with a set of MeSH concepts revealing the subject matter(s) of the document. Some statistical characteristics of the collection are depicted in Table 1. For measuring the IR effectiveness, we used *P@10, P@20* representing respectively the mean precision values at the top 10, 20 returned documents and *MAP* representing the *Mean Average Precision* calculated over all topics.

Number of documents	348,566
Average document length	100 tokens
Number of queries	63
Average query length	12 terms
Average number of relevant docs/query	50

Table 1: Test collection statistics

4.2 Experimental Setup

The purpose of our experimental evaluation is to determine the utility of our MeSH concept extraction method by measuring the impact of exploiting them for document expansion on the retrieval effectiveness. Therefore, we carried out two series of experiments: the first one is based on the classical indexing of titles and/or abstracts using the state-of-the-art weighting scheme OKAPI BM25 [9], as the baseline, denoted *BM25*. The second one concerns our conceptual indexing approach and consists of four scenarios:

1. the first one concerns the document expansion using concepts[1] manually assigned by human experts, denoted $Manual$,
2. the second one concerns the document expansion using concepts identified by the MTI tool [1], denoted MTI,
3. the third one concerns the document expansion using concepts identified by the cosine content-based text-to-concept mapping, denoted $Cosine$,
4. the last one concerns the document expansion using concepts identified by the combination of the cosine content-based and the Spearman rank correlation between word occurrences in document and concept entries (see formula 5), denoted $Combination$.

4.3 Results and discussion

We now present the experimental results of four document expansion strategies: $Manual, MTI, Cosine$ and $Combination$. At the first stage, we aim to measure the impact of the document kernel size on the IR effectiveness by tuning the number of extracted concepts; and at the second stage, we will measure the IR effectiveness using the optimal number of extracted concepts and our proposed semantic term weighting schema (see formula 6).

[1]only preferred terms are used for document expansion

4.3.1 Impact of the document kernel size

As mentioned in section 3.2, the number of identified concepts, namely N, has an important impact on the IR effectiveness and must be tuned experimentally. At this stage, terms, both original or those derived from the extraction method (expanded terms), are weighted using the state-of-the-art BM25 model [9]. In the $Manual$ approach, a dozen terms or descriptors among more than 24,000 MeSH main headings (e.g. "Affective Symptoms", "Life Style", etc.) representing the subject matter(s) of the article were manually selected by the NLM human indexers for assigning to each abstract. We don't aim to vary the number of descriptors that have been appropriately selected by human indexers. In an automatic setting, we firstly tuned the number of identified concepts from 0 to 50, with a step of 5. MTI took two months for achieving the concept extraction task on the OHSUMED collection while extracting only 25 concepts maximum by default for each abstract. Our concept extraction methods are able to extract any given number of concepts from the document. The concept extraction task using one of our two methods on the OHSUMED collection only took around four days and nights. Table 2 shows the MAP results achieved for three document expansion methods ($MTI, Cosine$ and $Combination$) when varying the number of concepts used for document expansion. The results of the $Manual$ method are presented in Table 4. MTI is not appropriate for document expansion when N is linearly increased because the IR effectiveness are dramatically decreased after expanding a limited number of concepts to documents. This may be due to the term over-generation problem suffered by the MetaMap approach when trying to increase the recall by generating a set of lexical variants of a given term and then map to the UMLS concepts before restricting to MeSH concepts. Therefore, we only retain $N = 5$ for MTI in the next experiments. The $Cosine$ method is better than the baseline $BM25$ in terms of MAP by 3.24% at $N = 25$. In addition, the $Combination$ method shows an improvement of 9.48% in terms of MAP over the baseline at the same value $N = 25$. The performance is decreased when N gets over 25 concepts. This could be explained by the fact that the more the number of valuable terms expanded to the document is, the higher the probability that query terms match the document ones will increase. However, we could not expand an unlimited concepts to the document because the semantics of the document will be dramatically changed by adding irrelevant terms to the document.

N	MTI	Cosine	Correlation
0	0.2595	0.2595	0.2595
5	**0.2448** (-5.66)	0.2424 (-6.59)	0.2609 (+0.54)
10	0.2395 (-7.71)	0.2497 (-3.78)	0.2677 (+3.16)
15	0.2344 (-9.67)	0.2566 (-1.12)	0.2735 (+5.39)
20	0.2331 (-10.17)	0.2626 (+1.19)	0.2800 (+7.90)
25	0.2316 (-10.75)	**0.2679** (+3.24)	**0.2841** (+9.48)
30	N/A	0.2561 (-1.13)	0.2682 (+3.35)
35	N/A	0.2537 (-2.24)	0.2536 (-2.27)
40	N/A	0.2499 (-3.69)	0.2501 (-3.62)
45	N/A	0.2473 (-4.70)	0.2417 (-6.86)
50	N/A	0.2447 (-5.70)	0.2405 (-7.32)

Table 2: MAP (% change) by varying $N \in [0..50]$

4.3.2 Retrieval effectiveness evaluation

After determining the optimal value of parameter N, we aim to measure the IR effectiveness of each indexing method. Here, we consider weighting terms in the original document with a higher score than those figuring in the expanded document by adopting hypothesis H (see formula 6). Table 3 shows the MAP values obtained by varying parameter α in the interval [0..1] with step of 0.1. Our best results are obtained at $\alpha_{Cosine} = 0.10, \alpha_{Correlation} = 0.10$. The selected value of α could be interpreted as some of the expanded terms are not relevant for expanding the document. Thus, their score should be reduced by 10% of the original score.

α	MTI	Cosine	Combination
0.0	0.2448 (-5.56)	0.2679 (+3.24)	0.2841 (+9.48)
0.1	0.2347 (-9.16)	**0.2758** (+6.28)	**0.2910** (+12.14)
0.2	0.2220 (-14.45)	0.2739 (+5.55)	0.2905 (+11.95)
0.3	0.2086 (-19.61)	0.2696 (+3.89)	0.2864 (+10.37)
0.4	0.1947 (-24.97)	0.2653 (+2.24)	0.2810 (+8.29)
0.5	0.1814 (-30.10)	0.2635 (+1.54)	0.27.66 (+6.59)
0.6	0.1677 (-35.38)	0.2575 (-0.77)	0.2724 (+4.97)
0.7	0.1537 (-40.77)	0.2541 (-2.08)	0.2684 (+3.43)
0.8	0.1416 (-45.43)	0.2506 (-3.43)	0.2661 (+2.54)
0.9	0.1303 (-49.79)	0.2481 (-4.39)	0.2634 (+1.50)

Table 3: MAP (% change) by varying $\alpha \in [0..1)$

Table 4 depicts the IR effectiveness of the $Manual, MTI$ and our two semantic indexing approaches. We observe that in an automatic setting, our best indexing method, namely *Combination*, gives the highest improvement rate (+12.14%) while the *Cosine* method only gives +6.28% in terms of MAP over the baseline $BM25$. This proves the interest to take into account the word order correlation during the concept extraction process. The MTI concept extraction method does not help to improve the IR effectiveness in the case of document expansion. Furthermore, we see that the $Manual$, *Cosine* and *Combination* methods consistently outperform the baseline. Although the gain of the *Combination* method is a little bit smaller than the $Manual$ method in terms of MAP (12.14% *vs.* 13.87%), the former is better than the latter in terms of P@10 and P@20.

	P@10	P@20	MAP
BM25	0.4365	0.3722	0.2595
Manual	0.4508 (+2.90)	0.3992 (+7.25)††	0.2955 (+13.87)
MTI	0.4127 (-5.45)††	0.3587 (-3.63)††	0.2473 (-4.70)†††
Cosine	0.4619 (+5.82)	0.4079 (+9.60)†††	0.2758 (+6.28)
Comb.	0.4683 (+7.29)†	0.4135 (+11.10)††	0.2910 (+12.14)†††

Table 4: IR effectiveness (% change) over the 63 queries

In order to show how our indexing approach is statistically significant, we computed the paired-sample T-tests between means of each ranking obtained by each indexing method and the baseline $BM25$. We assume that the difference between two given rankings is significant if $p < 0.05$ (noted †), very significant if $p < 0.01$ (noted ††) and extremely significant if $p < 0.001$ (noted †††). As shown in table 4, the paired-sample T-test ($M = 3.13\%, t = 5.28, df = 62, p = 0.00000175$) shows that our best concept extraction approach (*Combination*) for document expansion and indexing is extremely statistically significant compared to the baseline. Thus, we conclude that conceptual indexing and searching in conjunction with an efficient way of identifying an appropriate number of concepts would significantly improve the biomedical IR performance.

5. CONCLUSION

In this paper, we have proposed two methods of concept identification from the MeSH thesaurus for improving biomedical information indexing and retrieval. Our approach relies mainly on turning the concept mapping into a concept retrieval task by means of concept relevance scoring. The best scoring method combines the content-based similarity and the word order correlation between word occurrences in documents and concept entries. It shows that the word order correlation is a potential source of evidence that could be incorporated into a conceptual indexing and retrieval process. Future work will focus on improving the concept weighting schema using evidence issued from specific semantic features derived from the centrality of the concepts in the poly-hierarchical structure. Besides, we aim to decline the mapping score of concepts in the weighting schema in order to enhance the best concepts in the retrieval step.

6. REFERENCES

[1] A. R. Aronson, J. G. Mork, S. M. H. CW Gay, and W. J. Rogers. The nlm indexing initiative's medical text indexer. In *Medinfo 2004*, pages 268–272, 2004.

[2] N. Collier, C. Nobata, and J.-i. Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *18th international conference on Computational linguistics*, pages 201–207, 2000.

[3] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, August 2000.

[4] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi. Toward information extraction: identifying protein names from biological papers. In *Pacific Symposium on Biocomputing*, pages 707–718, 1998.

[5] R. Gaizauskas, G. Demetriou, and K. Humphreys. Term recognition and classification in biological science journal articles. In *Computional Terminology for Medical and Biological Applications Workshop*, pages 37–44, 2000.

[6] A. Hliaoutakis, K. Zervanou, and E. G. M. Petrakis. The amtex approach in the medical document indexing and retrieval application. *Data Knowledge Engineering*, 68(3):380–392, 2009.

[7] J. Kazama, T. Makino, and Y. Ohta. Tuning support vector machines for biomedical named entity recognition. In *Workshop on Natural Language Processing in the Biomedical Domain*, pages 1–8, 2002.

[8] M. Krauthammer, A. Rzhetsky, and *et al.* Using blast for identifying gene and protein names in journal articles. In *Gene*, pages 245–252, 2000.

[9] S. E. Robertson, S. Walker, and M. Hancock-Beaulieu. Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive. In *TREC-7*, pages 199–210, 1998.

[10] P. Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658–664, March 2006.

[11] G. Salton. A vector space model for information retrieval. *CACM*, 18(11):613–620, 1975.