

MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup*

Xiaohua Zhou, Xiaodan Zhang, and Xiaohua Hu

College of Information Science & Technology, Drexel University
3141 Chestnut Street, Philadelphia, PA 19104
xiaohua.zhou@drexel.edu, {xzhang, thu}@ischool.drexel.edu

Abstract. Dictionary-based biological concept extraction is still the state-of-the-art approach to large-scale biomedical literature annotation and indexing. The *exact dictionary lookup* is a very simple approach, but always achieves low extraction recall because a biological term often has many variants while a dictionary is impossible to collect all of them. We propose a generic extraction approach, referred to as *approximate dictionary lookup*, to cope with term variations and implement it as an extraction system called *MaxMatcher*. The basic idea of this approach is to capture the significant words instead of all words to a particular concept. The new approach dramatically improves the extraction recall while maintaining the precision. In a comparative study on GENIA corpus, the recall of the new approach reaches a 57% recall while the *exact dictionary lookup* only achieves a 26% recall.

1 Introduction

A biological concept is a unique meaning in biological domain. It represents a set of synonymous terms. For example, *C0020538* is a concept about the symptom of hypertension in Universal Medical Language System (UMLS) [13]; it represents a set of synonymous terms including *high blood pressure*, *hypertension*, and *hypertensive disease*. In comparison with individual words, a concept is more meaningful; in comparison with multi-word phrases, a concept well solves polysemy and synonymy problems [12]. Therefore, using biological concepts can improve the performance of many applications such as large-scale biomedical literature retrieval, clustering, and summarization.

There are volumes of work addressing the issue of biological concept extraction in literature. However, most of them utilize the special naming conventions or patterns to identify a few types of biological concepts such as genes, proteins and cells [1, 3, 4, 7, 8, 9, 10]. In general, those approaches are designed for very specific types of concepts, and work efficiently and effectively if the types of biological concepts have unique naming patterns. Many large-scale biomedical applications such as literature retrieval, clustering, and summarization, however, are interested in many rather than a few types of biological concepts most of which do not have unique naming patterns.

* This research work is supported in part from the NSF Career grant (NSF IIS 0448023). NSF CCF 0514679 and the research grant from PA Dept of Health.

For example, UMLS covers 135 semantic types of biological concepts; a typical genomic IR system will index all of them.

The dictionary-based biological concept extraction is still the state-of-the-art approach to large-scale biomedical literature annotation and indexing [6, 11, 12]. Its major advantage over the pattern-based approach is that it not only recognizes names, but also identifies unique concept identities. Among dictionary-based approaches, the *exact dictionary lookup* is the simplest one, but always achieves low extraction recall because a biological term often has many variants such as morphological variants, syntactic variants, and semantic variants [2] while a dictionary is impossible to collect all of them.

In this paper, we propose a new approach, referred to as *approximate dictionary lookup*, to the biological concept extraction. The basic idea is to capture the significant words rather than all words of a concept. For example, the word *gyrb* is significant to the concept “*gyrb protein*”; we will recognize it as a concept name even if the word *protein* is not present. Using UMLS Metathesaurus [13] as the dictionary, we implement this approach as an extraction system called *MaxMatcher*. We test the new approach on GENIA corpus [14]. As expected, the new approach dramatically raises the recall from 26% to 58%.

2 The Concept Extraction Approach

To overcome the limitation of *exact dictionary lookup*, we introduce an approximate dictionary lookup technique. The basic idea of this technique is to capture significant words rather than all words in a concept name. For example, the word *gyrb* is obviously very significant to the concept “*gyrb protein*”; we treat it as a concept name even if the word *protein* is not present. So the problem is reduced to measuring the significance of any word to given concept names. In particular, we propose a relative significance score measure in this paper. Suppose a concept (c) has n concept names denoted as s_1, \dots, s_n , respectively. Let $N(w)$ denotes the number of concepts whose variant names contain word w , and let w_{ji} denotes the i -th word in the j -th variant name of the concept, the significance of w to the concept is defined as follows:

$$I(w, c) = \max\{I(w, s_j) \mid j \leq n\} \quad (2.1)$$

where :

$$I(w, s_j) = \begin{cases} 0 & w \notin s_j \\ \frac{1/N(w)}{\sum_i 1/N(w_{ji})} & w \in s_j \end{cases}$$

We use UMLS Metathesaurus 2005AA version [13] as the dictionary to train the significance score of each word to biological concepts containing that word. The UMLS Metathesaurus has a table called normalized string index, which record all normalized names of each concept. We remove normalized strings containing more than ten words and then use the remaining 2,573,244 strings to build the significance score matrix. A huge matrix, 509,170 rows (words) by 998,774 columns (concepts), is

```

Find next starting word  $t_s$ 
 $k = 0$ 
 $C = \{c \mid t_s \in T(c)\}$  /*  $T(c)$  is the set of words appearing in names of concept  $c$  */
For each  $c \in C$   $S_c = I(t_s, c)$  /*  $I(t_s, c)$  is the score of word  $t_s$  to concept  $c$  */
While next word  $t$  is not boundary word AND  $k < skip$ 
     $N = \{c \mid t \in T(c) \wedge c \in C\}$ 
    IF  $N = \emptyset$  Then  $k = k + 1$ 
    Else
         $C = N$ 
        For each  $c \in C$   $S_c = S_c + I(t, c)$ 
    End If
Wend
 $C = \{c \mid S_c > threshold \wedge c \in C\}$ 
If  $|C| > 0$  Then
    return concept name and candidate concepts  $c \in C$ 
End If

```

Fig. 1. The algorithm for extracting one concept name and its candidate concept IDs. The *threshold* is set to 0.95; the maximum number (*skip*) of skipped words is set to 1.

obtained. Because for each word, only a few concepts contain it, we use sparse matrix to make the storage and search more efficiently.

During the stage of extraction, we use a set of simple rules to identify the boundary of a concept candidate. A biological concept name should begin with a noun, a number, or an adjective while ending with a noun or a number; it can not contain any boundary words including (1) punctuations (except hyphen, period, and single quote), verbs, and conjunctions and prepositions (except “*of*”). In other words, whenever a boundary word is encountered, a candidate concept name reaches its end. The detailed searching algorithm is shown in Figure 1.

The major advantage of *approximate dictionary lookup* is that even if a concept name changes the word ordering a little bit, inserts or deletes a couple of insignificant words, it is still can be recognized. According to its definition, the significance score of a concept name should be equal to or greater than 1.0 if no word is missing. Thus, the threshold of significance score should be close to 1.0. If the threshold is too small, our approach may falsely recognize “*high pressure*” as the concept name “*high blood pressure*”; if it is too high, our approach may fail to recognize “*gyrb*” as “*gyrb protein*”. We found that 0.95 as the threshold gave good results for UMLS-based biological concept extraction. Our approach is able to recognize concept names with a couple of insertions such as articles, pronouns, and even nouns. The parameter *skip* controls the maximum number of insertions. We found that *skip*=1 gave good results.

The searching results are concept names and corresponding concept IDs. If two or more concept IDs are returned, we need to further figure out the meaning the extracted concept name refers to. The words surrounding the extracted concept name are often indicative to the meaning [5]. Thus, we take surrounding words (4 to the left and 4 to the right) as the context and use the same algorithm as shown in Figure 1 to disambiguate the meaning of the extracted concept name if necessary.

3 Experimental Results

We evaluate both efficiency and effectiveness of the *MaxMatcher*. The effectiveness is evaluated on GENIA 3.02 corpus [14] which consists of 2,000 human annotated PubMed abstracts. We compare the result of *MaxMatcher* with that of two other *exact dictionary lookup* systems, *BioAnnotator* [8] and *ExactMatcher*. *ExactMatcher* is implemented by us. The machine-extracted terms are compared with human annotations. Because human annotation is kind of subjective, we provide exact-match based evaluation and approximate-match based evaluation, following the evaluation method in [8]. For approximate-match, the human annotation should be the substring of the machine annotation, or the opposite.

The comparison among three systems is presented in Table 1. For exact-match, *MaxMatcher* performs significantly better than the other two systems in terms of both precision and recall. For approximate match, the precision of *MaxMatcher* is comparable to that of the other two systems while the recall is significantly better than that of the other two.

Table 1. The effectiveness comparison. *BioAnnotator* [8] actually tested several configurations. But only the configuration with only dictionaries (i.e. exact dictionary lookup) is compared. *BioAnnotator* was evaluated on GENIA 1.1 (containing 670 human annotated abstracts of research papers). The dictionary used for *BioAnnotator* also includes LocusLink and GeneAlias in addition to UMLS.

IE Systems	Exact Match Eva.			Approximate Match Eva.		
	Recall	Precision	F-score	Recall	Precision	F-score
MaxMatcher	57.73	54.97	56.32	75.18	71.60	73.35
ExactMatcher	26.63	31.45	28.84	61.56	72.69	66.66
BioAnnotator	20.27	44.58	27.87	39.75	87.67	54.70

For efficiency comparison, we download first 10,000 PubMed abstracts published in 2005 and count the time for annotating these abstracts by *MaxMatcher* and *ExactMatcher*, respectively. It takes 510 seconds for *MaxMatcher* to annotate all 10,000 PubMed abstracts; the average annotation speed is 19.6 abstracts per second. *ExactMatcher* is faster. It only costs 320 seconds to process those abstracts; the average annotation speed is 31.3 abstracts per second. However, *ExactMatcher* consumes much more memory (765Megabytes) than *MaxMatcher* (362 Megabytes).

4 Conclusions

Dictionary-based biological concept extraction is still the state-of-the-art approach to the large-scale biomedical literature annotation and indexing. The *exact dictionary lookup* is very simple but always achieves low extraction recall because biological terms often have many variants while a dictionary is impossible to collect all of them. In this paper, we propose a generic approach, referred to as *approximate dictionary lookup*, to cope with the biological concept variation. The basic idea of the new approach is to capture the significant words of a biological concept rather than all of

them. A comparative study on GENIA corpus shows that the new approach can dramatically improve the extraction recall while maintaining the precision. However, the extraction efficiency of the new approach goes down a little bit in comparison with the *exact dictionary lookup*.

References

1. Chang, J.T., Schütze, H., and Altman, R.B., "GAPSCORE: finding gene and protein names one word at a time", *Bioinformatics*, Vol. 20, No. 2, pp. 216-225, 2004.
2. Chiang, J.-H. and Yu, H.-C., "Literature extraction of protein functions using sentence pattern mining", *IEEE Transactions on Knowledge and Data Engineering*, 17(8), Aug. 2005 Page(s):1088 – 1098
3. Collier, N., Nobata, C., and Tsujii, J., "Extracting the names of genes and gene products with a Hidden Markov Model", *Proc. COLING 2000*, 201--207, 2000
4. Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T., "Toward information extraction: Identifying protein names from biological papers", In *Proceedings of Pacific Symposium on Biocomputing*, pages 707--718, Maui, Hawaii, January 1998.
5. Lesk, M., "Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream Cone", *Proceedings of the SIGDOC'86 Conference, ACM*, 1986.
6. Rindfleisch, T.C., Tanabe, L., and Weinstein, J.N., "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature", *Proceedings of Pacific Symposium on Bioinformatics*, Hawaii, USA, pp. 514-525, 2000.
7. Song, Y.-I., Kim, S.-B., and Rim, H.-C., "Terminology Indexing and Reweighting methods for Biomedical Text Retrieval", In *Proceedings of the SIGIR'04 Workshop on Search and Discovery in Bioinformatics*, Sheffield, UK, ACM, July 2004.
8. Subramaniam, L., Mukherjea, S., Kankar, P., Srivastava, B., Batra, V., Kamesam, P. and Kothari, R., "Information Extraction from Biomedical Literature: Methodology, Evaluation and an Application", In *the Proceedings of the ACM Conference on Information and Knowledge Management*, New Orleans, Louisiana, 2003.
9. Tanabe, L. and Wilbur, W., "Tagging gene and protein names in biomedical text", *Bioinformatics*, Vol. 18, No. 8, pp.1124-1132, 2002.
10. Zhou, G.-D., Zhang, J., Su, J., Shen, D., and Tan, C.-L., "Recognizing Names in Biomedical Texts: A Machine Learning Approach", *Bioinformatics*, 20(7), 1178-1190, 2004.
11. Zhou, X., Han, H., Chankai, I., Prestrud, A., and Brooks, A., "Converting Semi-structured Clinical Medical Records into Information and Knowledge", *Proceeding of The International Workshop on Biomedical Data Engineering (BMDE) in conjunction with the 21st International Conference on Data Engineering (ICDE)*, Tokyo, Japan, April 5-8, 2005.
12. Zhou, X., Hu, X. and Zhang, X., "Using Concept-based Indexing to Improve Language Modeling Approach to Genomic IR", *The 28th European Conference on Information Retrieval (ECIR' 2006)*, 10 - 12 April, 2006, London, UK.
13. UMLS, <http://www.nlm.nih.gov/research/umls/>
14. GENIA Corpus, <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>