A Maximum-Entropy Approach for Accurate Document Annotation in the Biomedical Domain

George Tsatsaronis*, Natalia Macari, Sunna Torge, Heiko Dietze, Michael Schroeder

Biotechnology Center (BIOTEC), Technische Universität Dresden, 01062, Dresden, Germany

ABSTRACT
Motivation: The increasing number of scientific literature on the Internet and the absence of efficient tools used for classifying and searching the documents are the two most important factors that influence the speed of the search and the quality of the results. Previous studies have shown that the usage of ontologies makes it possible to process document and query information at the semantic level, which greatly improves the search for the relevant information and makes one step further towards the Semantic Web. A fundamental step in these approaches is the annotation of documents with ontology concepts, which can also be seen as a classification task. In this paper we address this issue for the biomedical domain and present a new automated and robust method, based on a Maximum Entropy approach, for annotating biomedical literature documents with MeSH concepts, which provides very high F-measure. The experimental evaluation shows that the suggested robust to the ambiguity of terms, and can provide very good performance even when a very small number of training documents is used.

1 INTRODUCTION

With the rapid expansion of the Internet as a source of scientific and educational literature, the search for relevant information has become a difficult and time consuming process. The current state of the Internet can be characterized by weak structured data and, practically, the absence of relationships between data. Current search engines, such as Google and Yahoo, provide a keyword-based search, which takes into account mainly the surface string similarity between query and document terms, and often a simple synonym expansion, omitting other types of information about terms, such as polysemy and homonymy. In order to address this problem and improve search results, the usage of ontologies is suggested to allow for document annotation with ontology concepts. The usage of ontologies provides a content-based access to the data, which makes it possible to process information at the semantic level and significantly improve the search of relevant documents, as it has been shown by recent studies in the case of the search in the life sciences literature (Doms, 2008; Doms and Schroeder, 2005).

Some representative examples of such search engines for the biomedical domain are: (a) GoPubMed which uses the Gene Ontology (GO) and the Medical Subject Headings (MeSH) as background knowledge for indexing the biomedical literature stored in the PubMed database, and various text mining techniques and algorithms (stemming, tokenization, synonym detection) for the identification of relevant ontology entities in PubMed abstracts, (b) semedico, which provides access to semantic metadata about MEDLINE abstracts using the JULIE Lab text mining engine and MeSH as a knowledge base, and (c) novoseek, which uses external available data and contextual term information to identify key biomedical terms in biomedical literature documents. However, in all cases the challenges that arise are several and difficult to resolve; more precisely: (i) the amount of scientific documents to be annotated and indexed is very large, as PubMed documents grow really fast in number, (ii) the presence of ambiguous concepts constitutes the classification (annotation) process a challenging task, and, (iii) the classifier model used needs to be trained and tuned specifically for this domain, in order to achieve the best possible results, and in tandem needs to be fast and robust to address challenges (i) and (ii) respectively.

Fig. 1. Left: number of PubMed articles (blue line) indexed over the period 1965-2010 and logarithmic trend (red line). Right: number of PubMed articles (blue line), plotted with the number of MeSH annotated documents (red line).

* To whom correspondence should be addressed: george.tsatsaronis@biotec.tu-dresden.de.

1 http://www.gopubmed.com/web/gopubmed/
2 http://www.semedico.org
3 http://www.julielab.de
Fig. 2. Pie showing the ambiguous MeSH terms, examining 4,078 terms, and consulting three dictionaries/thesauri.

As a proof of concept for (i), we present in Figure 1 the growth of PubMed documents over the period 1965 - 2010. The figure shows clearly that new PubMed documents are nowadays doubled within the past 20 years (left), as also discussed by Biglu (2007). The exponential trend (red line) also shows that this tendency continues. In parallel, we can observe that the annotated documents with MeSH concepts (red line) attempts to keep up with the document growth (right). For this purpose, the Medical Text Indexer System is used, which makes the annotation process semi-automatic and improves the efficiency of indexing PubMed articles. This constitutes as fundamental the need for fast and accurate automated annotation methods with MeSH concepts, so that the growth of PubMed documents can be followed with respective concept annotations.

As a proof of concept for (ii), we have selected randomly a set of 4,078 MeSH, which are the terms under the roots: diseases, anatomy, and psychology. In these terms we will also base our analysis and our experimental evaluation. For all of them we have measured the number of different meanings that these terms may carry, consulting three very popular thesauri/lexica, namely the WordNet thesaurus for the English language, the Wikipedia encyclopedia (English version), which is currently the largest electronic encyclopedia available, and the UMLS thesaurus, which is also focused in our examined domain. The measurements shown in the pie of Figure 2, reveal that 23.3% of the examined terms are ambiguous, i.e., they have more than one meaning. Another interesting finding is the coverage of the non-domain specific lexica, i.e., WordNet and Wikipedia, which is 78% combined. In fact only 22% of the examined have entries only in the domain specific UMLS thesaurus. In order to stress out the implications of the existence of such ambiguous terms in the annotation process, we have furthermore analyzed the number of different documents these 4,078 terms appear literally in GoPubMed, as well as in another popular and general purpose search engine, namely Yahoo. The aim of this analysis is to show how the number of documents that these terms appear literally varies, depending on their number of entries in the two used lexica. In Figure 3 we present four plots showing the results of this analysis.

Fig. 3. Scatter plots of number of documents where the terms appear literally in GoPubMed (horizontal axis), and Yahoo (vertical axis). Red lines show medians.

The top left figure shows for all the terms the number of documents in which each of the examined term appears literally in the GoPubMed (horizontal axis) and the Yahoo (vertical axis) indexed documents. The figure shows that the difference on the number of the retrieved documents comparing the results returned by GoPubMed and Yahoo is several orders of magnitude. A typical term appears literally in almost 5,000 GoPubMed documents and in 1 million Yahoo documents. The remaining three plots highlight respectively the terms for which there is no entry in the majority of the used lexica (yellow), the terms for which there is exactly one entry in the majority of the used lextica (red), and the terms which are ambiguous in according to the majority of the used lexica. It is obvious from the plots, that there is a shift of the placement of the terms from left to right and, in parallel, from bottom to top as the number of entries increase. This fact shows that the ambiguous terms may appear in a very large number of documents (contexts), larger compared to the rest of the terms, and, thus, any context-based model for document annotation will have to handle a lot of noise for those terms, highlighting the need for a very robust annotator.

2 APPROACH

The approach that we follow for automated document annotation of biomedical literature documents with MeSH concepts creates a context model for each and every concept of the used ontology, which characterizes the term and that consists of the lexical tokens taken from related PubMed articles' abstracts. The approach uses the notion of Maximum Entropy, whose principle is to measure the uncertainty of each class (also known as entropy), expressed by information that we do not have about the classes occupied by
the data. Given the fact that the Maximum Entropy (Max-Ent) approach has been applied successfully in the past to several natural language and computational linguistic tasks, such as word sense disambiguation (Doms, 2008), part of speech tagging, prepositional phrase attachment, and named entity recognition (Ratnaparkhi, 1998), but also to gene annotation (Raychauduri et al., 2002), and to mining patient medication status (Pakhomov et al., 2002), in this work we decided to adopt this approach in order to investigate its performance in the task of document annotation for the biomedical domain. The MaxEnt method is insensitive to noisy data and capable to process incomplete data such as sparse data or data with missing attributes. In addition, the MaxEnt models can be trained on massive data sets (Mann et al., 2009), and their implementation is publicly available through open source projects, such as OpenNLP4.

In Figure 4 we show in detail how we apply MaxEnt for the annotation of documents with MeSH concepts. The algorithm is separated into two parts: training and testing. For each MeSH term we measure the values of pre-selected features by examining PubMed documents. The features in our case are of four types: (1) lexical tokens from the titles of PubMed documents, (2) lexical tokens from the abstracts of PubMed documents, (3) name of the journal in which the respective documents were published, and (4) year of the published documents. The algorithm constructs a context model for each of the terms, trained on a pre-selected set of positive and negative examples. For the training part, the weights of the features are maximized using iteratively re-weighted least squares (IRLS). The classes on which the classifier is trained are always two for each constructed model, i.e., for each term: positive, denoted with 1, and negative, denoted with 0. Once the feature weights for each class are maximized and known for each term mj in M (βj1 and βj0 respectively), the testing procedure can be applied, which decides for each term mj separately whether it should annotate the instance ti (positive class), or not (negative class). For this reason, a classification threshold using a parameter δ is used.

3 RESULTS

For our experimental setup we used 4,078 MeSH terms, under the MeSH roots: diseases, anatomy, and psychology. This selection is not random, as psychology is considered to have difficult terms for annotation, because many terms are general, diseases is considered to have easy terms, and anatomy has an unknown difficulty. Thus, the selection spans across all levels of annotation difficulty. All of the experiments shown next were conducted using 10-fold cross validation, and in all cases we measure average precision, recall and F-Measure based on the classification results. In all cases, only the title and the abstract of each document were used for the lexical features (i.e., the two of the four features used by MaxEnt), as explained in the previous section. The δ parameter was set to the value that was found optimal in the validation set (10% of the training was always kept as validation set). This value was 0.1.

Table 1. Results of annotation for two methods, Exact Matching and MaxEnt. Results on ambiguous terms are also shown separately.

| Method       | All Terms | Ambiguous Terms |
|--------------|-----------|-----------------|
|              | P        | R     | F     | P  | R   | F   |
| Exact Matching | 52.3     | 22.1  | 23.9  | 45.4 | 37  | 34.8|
| MaxEnt        | 99.4     | 86.8  | 92.4  | 99.3 | 86.8| 92.4|

Table 1 shows the results for our method (MaxEnt) as well as a simple baseline technique for annotation, which is the use of exact matching. Exact Matching searches for the ex-

---

4 http://opennlp.sourceforge.net/index.html
act or stemmed appearance of each of the terms in the abstract or the title of the documents. In case it is found, the document is annotated with that term. The table shows that the MaxEnt approach gives an F-Measure of 92.4% for all the terms of our experiment, which is almost four times larger than the F-Measure of the Exact Matching approach (23.9% respectively). The most interesting observations arise from the separate study of the ambiguous terms, i.e., in this case the terms with more than one entry in UMLS, which are included, however, in the results of all terms shown in Table 1. Naturally, the Exact Matching approach drops its precision in those terms, by almost 8 percentage points (p.p.), and increases its recall by almost 15 p.p.. MaxEnt manages to retain high performance in those terms, always higher from Exact Matching. Its precision and recall remain almost the same in the ambiguous terms. Regarding the performance for the individual MeSH branches, the MaxEnt F-Measure was 93.52% for anatomy, 92.21% for diseases and 91.35% for psychology.

Figure 5 (top left) shows the F-Measure of MaxEnt for increasing number of training documents. As shown, MaxEnt can perform really well, even with few hundreds of training documents per term. Top right shows the distribution of the F-Measure values in the ambiguous terms. In the majority of the cases, the F-Measure is really high, more than 90%. The two bottom figures show F-Measures obtained when using each feature type individually. As shown, title and year are the most important features, while journal is very important when a large number of training documents is used. We also present the F-Measures when several combinations of features are explored (bottom right). The results show again that year is very important (blue and black lines), since, if it is omitted (green and red lines), the performance drops significantly. Overall, the results show that MaxEnt can annotate documents successfully with MeSH terms, and with very few training documents needed. The results also show that MaxEnt produces robust models that are not affected in precision and F-Measure by the ambiguity of the terms.

4 CONCLUSIONS AND FUTURE WORK

In this work we introduced a novel approach for annotating documents of the biomedical literature with concepts from the MeSH ontology. The approach is based on the use of Maximum Entropy (MaxEnt) classifiers to perform the annotation. For each of the terms, a MaxEnt model is trained and it can be applied to any document in order to decide whether it should be annotated with the respective term or not. We performed a thorough experimental evaluation on the application of the proposed MaxEnt approach on a selected set of 4,078 MeSH terms that were used to annotate PubMed documents. We showed that the used feature types (title, abstract, year, and journal) are sufficient for producing high accuracy annotations. The results showed that the proposed approach was able to annotate PubMed documents with an average precision of 99.4%, average recall of 86.8%, and average F-Measure of 92.4%. Regarding the tuning of the used parameters, we found that a delta value of 0.1 produces the best results, and that even few training documents are sufficient to achieve very good performance. As a future work, we plan to investigate the connection of the ambiguity of terms to the semantic search procedure and the ranking of documents.

REFERENCES

Biglu, M.H. (2007) The editorial policy of languages is being changed in Medline. Acimed, 16(3).
Doms, A. and Schroeder, M. (2005) GoPubMed: Exploring PubMed with the Gene Ontology. Nucleic Acids Research, 33.
Doms, A. (2008) GoPubMed: Ontology-based literature search for the life sciences. PhD Thesis, Technical University of Dresden.
Mann, G., et.al. (2009) Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models. Advances in Neural Information Processing Systems, 22.
Pakhomov, S.V., et al. (2002) Maximum entropy modeling for mining patient medication status from free text. Proc. of AMIA Symposium, pp. 587-591.
Ratnaparkhi, A. (1998) Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD Thesis, University of Pennsylvania.
Raychaudhuri, S., et al. (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Res. 12(1).