Classifying Biological Full-Text Articles for Multi-Database Curation

Wen-Juan Hou, Chih Lee and Hsin-Hsi Chen
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
{wjhou, clee}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

In this paper, we propose an approach for identifying curatable articles from a large document set. This system considers three parts of an article (title and abstract, MeSH terms, and captions) as its three individual representations and utilizes two domain-specific resources (UMLS and a tumor name list) to reveal the deep knowledge contained in the article. An SVM classifier is trained and cross-validation is employed to find the best combination of representations. The experimental results show overall high performance.

1 Introduction

Organism databases play a crucial role in genomic and proteomic research. It stores the up-to-date profile of each gene of the species interested. For example, the Mouse Genome Database (MGD) provides essential integration of experimental knowledge for the mouse system with information annotated from both literature and online sources (Bult *et al.*, 2004). To provide biomedical scientists with easy access to complete and accurate information, curators have to constantly update databases with new information. With the rapidly growing rate of publication, it is impossible for curators to read every published article. Since fully automated curation systems have not met the strict requirement of high accuracy and recall, database curators still have to read some (if not all) of the articles sent to them. Therefore, it will be very helpful if a classification system can correctly identify the curatable or relevant articles in a large number of biological articles.

Recently, several attempts have been made to classify documents from biomedical domain (Hirschman *et al*., 2002). Couto *et al*. (2004) used the information extracted from related web resources to classify biomedical literature. Hou *et al*. (2005) used the reference corpus to help classifying gene annotation. The Genomics Track (http://ir.ohsu.edu/genomics) of TREC 2004 and 2005 organized categorization tasks. The former focused on simplified GO terms while the latter included the triage for "tumor biology", "embryologic gene expression", "alleles of mutant phenotypes" and "GO" articles. The increase of the numbers of participants at Genomics Track shows that biological classification problems attracted much attention.

This paper employs the domain-specific knowledge and knowledge learned from full-text articles to classify biological text. Given a collection of articles, various methods are explored to extract features to represent a document. We use the experimental data provided by the TREC 2005 Genomics Track to evaluate different methods.

The rest of this paper is organized as follows. Section 2 sketches the overview of the system architecture. Section 3 specifies the test bed used to evaluate the proposed methods. The details of the proposed system are explained in Section 4. The experimental results are shown and discussed in Section 5. Finally, we make conclusions and present some further work.

2 System Overview

Figure 1 shows the overall architecture of the proposed system. At first, we preprocess each training article, and divide it into three parts, including (1) title and abstract, (2) MeSH terms assigned to this article, and (3) captions of figures and tables. They are denoted as "Abstract", "MeSH", and "Caption" in this paper, respectively. Each part is considered as a representation of an article. With the help of domain-specific knowledge, we obtain more detail representations of an article. In the model selection phase, we perform feature ranking on each representation of an article and employ cross-validation to determine the number of features to be kept. Moreover, we use cross-validation to obtain the best combination of all the representations. Finally, a support vector machine (SVM) (Vapnik, 1995; Hsu *et al.*, 2003) classifier is obtained.

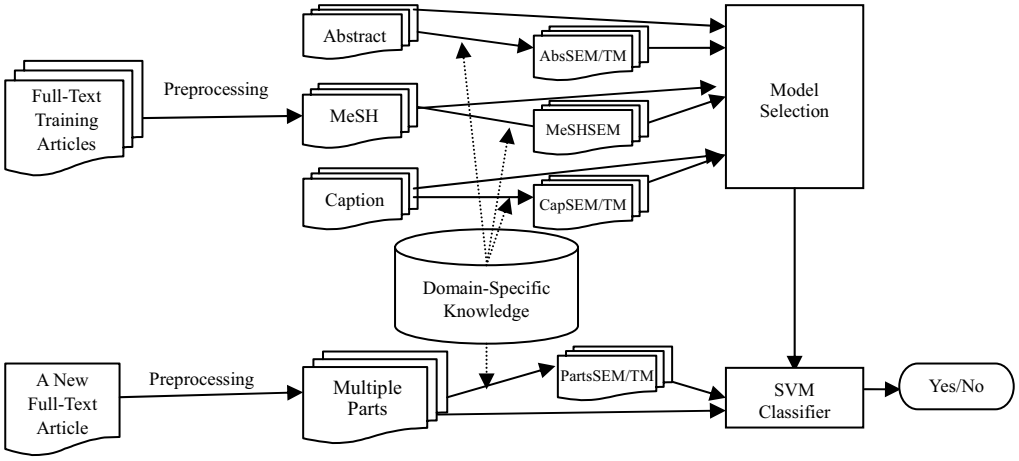

Figure 1. System Architecture

3 Experimental Data

We train classifiers for classifying biomedical articles on the Categorization Task of the TREC 2005 Genomics Track. The task uses data from the Mouse Genome Informatics (MGI) system (http://www.informatics.jax.org/) for four categorization tasks, including tumor biology, embryologic gene expression, alleles of mutant phenotypes and GO annotation. Given a document and a category, we have to identify whether it is relevant to the given category.

The document set consists of some full-text data obtained from three journals, i.e., Journal of Biological Chemistry, Journal of Cell Biology and Proceedings of the National Academy of Science in 2002 and 2003. There are 5,837 training documents and 6,043 testing documents.

4 Methods

4.1 Document Preprocessing

In the preprocessing phase, we perform acronym expansion on the articles, remove the remaining tags from the articles and extract three parts of interest from each article. Abbreviations are often used to replace long terms in writing articles, but it is possible that several long terms share the same short form, especially for gene/protein names. To avoid ambiguity and enhance clarity, the acronym expansion operation replaces every tagged abbreviation with its long form followed by itself in a pair of parentheses.

4.2 Employing Domain-Specific Knowledge

With the help of domain-specific knowledge, we can extract the deeper knowledge in an article. For example, with a gene name dictionary, we can identify the gene names contained in an article. Moreover, by further consulting organism databases, we can get the properties of the genes. Two domain-specific resources are exploited in this study. One is the Unified Medical Language System (UMLS) (Humphreys *et al.*, 1998) and the other is a list of tumor names obtained from Mouse Tumor Biology Database (MTB)[1].

UMLS contains a huge dictionary of biomedical terms – the UMLS Metathesaurus and defines a hierarchy of semantic types – the UMLS Semantic Network. Each concept in the Metathesaurus contains a set of strings, which are variants of each other and belong to one or more semantic types in the Semantic Network. Therefore, given a string, we can obtain a set of semantic types to which it belongs. Then we obtain another representation of the article by gathering the semantic types found in the part of the article. Consequently, we get another three much deeper representations of an article after this step. They are denoted as "AbstractSEM", "MeSHSEM" and "CaptionSEM".

We use the list of tumor names on the Tumor task. We first tokenize all the tumor names and stem each unique token. With the resulting list of unique stemmed tokens, we use it as a filter to remove the tokens not in the list from the "Abstract" and "Caption", which produce "AbstractTM" and "CaptionTM".

4.3 Model Selection

As mentioned above, we generate several representations for an article. In this section, we explain how feature selection is done and how the best combination of the representations

[1] http://tumor.informatics.jax.org/mtbwi/tumorSearch.do

of an article is obtained.

For each representation, we first rank all the tokens in the training documents via the chi-square test of independence. Postulating the ranking perfectly reflects the effectiveness of the tokens in classification, we then decide the number of tokens to be used in SVM classification by 4-fold cross-validation. In cross-validation, we use the TF*IDF weighting scheme. Each feature vector is then normalized to a unit vector. We set C_+ to $u_r * C_-$ because of the relatively small number of positive examples, where C_+ and C_- are the penalty constants on positive and negative examples in SVMs. After that, we obtain the optimal number of tokens and the corresponding SVM parameters C_- and *gamma*, a parameter in the radial basis kernel. In the rest of this paper, "Abstract30" denotes the "Abstract" representation with top-30 tokens, "CaptionSEM10" denotes "CaptionSEM" with top-10 tokens, and so forth.

After feature selection is done for each representation, we try to find the best combination by the following algorithm.

Given the candidate representations with selected features, we start with an initial set containing some or zero representation. For each iteration, we add one representation to the set by picking the one that enhances the cross-validation performance the most. The iteration stops when we have exhausted all the representations or adding more representation to the set doesn't improve the cross-validation performance.

For classifying the documents with better features, we run the algorithm twice. We first start with an empty set and obtain the best combination of the basic three representations, e.g., "Abstract10", "MeSH30" and "Caption10". Then, starting with this combination, we attempt to incorporate the three semantic representations, e.g., "Abstract30SEM", "MeSH30SEM" and "Caption10SEM", and obtain the final combination. Instead of using this algorithm to incorporate the "AbstractTM" and "CaptionTM" representations, we use them to replace their unfiltered counterparts "Abstract" and "Caption" when the cross-validation performance is better.

5 Results and Discussions

Table 1 lists the cross-validation results of each representation for each category (in Normalized Utility (NU)[2] measure). For category Allele, "Caption" and "AbstractSEM" perform the best among the basic and semantic representations, respectively. For category Expression, "Caption" plays an important role in identifying relevant documents, which agrees with the finding by the winner of KDD CUP 2002 task 1 (Regev *et al.*, 2002). Similarly, MeSH terms are crucial to the GO category, which are used by top-performing teams (Dayanik *et al.*, 2004; Fujita, 2004) in TREC Genomics 2004. For category Tumor, MeSH terms are important, but after semantic type extraction, "AbstractSEM" exhibits relatively high cross-validation performance. Since only 10 features are selected for the "AbstractSEM", using this representation alone may be susceptible to over-fitting. Finally, by comparing the performance of the "AbstractTM" and "Abstract", we find the list of tumor names helpful for filtering abstracts.

We list the results for the test data in Table 2. Column "Experiment" identifies our proposed methods. We show six experiments in Table 2: one for Allele (AL), one for Expression (EX), one for GO (GO) and three for Tumor (TU, TN and TS). Column "cv NU" shows the cross-validation NU measure, "NU" shows the performance on the test data and column "Combination" lists the combination of the representations used for each experiment. In this table, "M30" is the abbreviation for "MeSH30", "CS10" is for "CaptionSEM10", and so on. The combinations for the first 4 experiments, i.e., AL, EX, GO and TU, are obtained by the algorithm described in Section 4.3, while the combination for TN is obtained by substituting "AbstractTM30" for "Abstract30" in the combination for TU. The experiment TS only uses the "AbstractSEM10" because its cross-validation performance beats all other combinations for the Tumor category.

The combinations of the first 5 experiments illustrate that adding other inferior representations to the best one enhances the performance, which implies that the inferior ones may contain important exclusive information. The cross-validation performance fairly predicts the performance on the test data, except for the last experiment TS, which relies on only 10 features and is therefore susceptible to over-fitting.

[2] Please refer to the TREC 2005 Genomics Track Protocol (http://ir.ohsu.edu/genomics/2005protocol.html).

	Allele	Expression	GO	Tumor
	# Tokens / NU	# Tokens / NU	# Tokens / NU	# Tokens / NU
Abstract	10 / 0.7707	10 / 0.5586	10 / 0.4411	10 / 0.8055
MeSH	10 / 0.7965	10 / 0.6044	**10 / 0.4968**	**30 / 0.8106**
Caption	**10 / 0.8179**	**10 / 0.7192**	10 / 0.4091	10 / 0.7644
AbstractSEM	**10 / 0.7209**	10 / 0.4811	10 / 0.3493	**10 / 0.8814**
MeSHSEM	10 / 0.6942	10 / 0.4563	**10 / 0.4403**	10 / 0.7047
CaptionSEM	30 / 0.6789	**10 / 0.5433**	10 / 0.2551	30 / 0.7160
AbstractTM				**30 / 0.8325**
CaptionTM				10 / 0.7498

Table 1. Partial Cross-validation Results.

Experiment	cv NU	NU	Recall	Precision	F-score	Combination
AL (for Allele)	0.8717	0.8423	0.9488	0.3439	0.5048	M30+C10+A10+CS10+AS10+MS10
EX (for Expression)	0.7691	0.7515	0.8190	0.1593	0.2667	M10+C10+CS10+MS10
GO (for GO)	0.5402	0.5332	0.8803	0.1873	0.3089	M10+C10+MS10
TU (for Tumor)	0.8742	0.8299	0.9000	0.0526	0.0994	M30+C30+A30+AS10+CS30
TN (for Tumor)	0.8764	0.8747	0.9500	0.0518	0.0982	M30+C30+AT30+AS10+CS30
TS (for Tumor)	0.8814	0.5699	0.6500	0.0339	0.0645	AS10

Table 2. Evaluation Results.

Subtask	NU (Best/Median)	Recall (Best/Median)	Precision (Best/Median)	F-score (Best/Median)
Allele	0.8710/0.7773	0.9337/0.8720	0.4669/0.3153	0.6225/0.5010
Expression	0.8711/0.6413	0.9333/0.7286	0.1899/0.1164	0.3156/0.2005
GO Annotation	0.5870/0.4575	0.8861/0.5656	0.2122/0.3223	0.3424/0.4107
Tumor	0.9433/0.7610	1.0000/0.9500	0.0709/0.0213	0.1325/0.0417

Table 3. Best and Median Results for Each Subtask on TREC 2005 (Hersh *et al*., 2005).

To compare with our performance, we list the best and median results for each subtask on the genomics classification task of TREC 2005 in Table 3. Comparing to Tables 2 and 3, it shows our experimental results have overall high performance.

6 Conclusions and Further Work

In this paper, we demonstrate how our system is constructed. Three parts of an article are extracted to represent its content. We incorporate two domain-specific resources, i.e., UMLS and a list of tumor names. For each categorization work, we propose an algorithm to get the best combination of the representations and train an SVM classifier out of this combination. Evaluation results show overall high performance in this study.

Except for MeSH terms, we can try other sections in the article, e.g., Results, Discussions and Conclusions as targets of feature extraction besides the abstract and captions in the future. Finally, we will try to make use of other available domain-specific resources in hope of enhancing the performance of this system.

Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC94-2213-E-002-033 and NSC94-2752-E-001-001-PAE.

References

Bult, C.J., Blake, J.A., Richardson, J.E., Kadin, J.A., Eppig, J.T. and the Mouse Genome Database Group. The Mouse Genome Database (MGD): Integrating Biology with the Genome. *Nucleic Acids Research*, 32, D476–D481, 2004.

Couto, F.M., Martins, B. and Silva, M.J. Classifying Biological Articles Using Web Resources. *Proceedings of the 2004 ACM Symposium on Applied Computing*, 111-115, 2004.

Dayanik, A., Fradkin, D., Genkin, A., Kantor, P., Lewis, D.D., Madigan, D. and Menkov, V. DIMACS at the TREC 2004 Genomics Track. *Proceedings of the Thirteenth Text Retrieval Conference*, 2004.

Fujita, S., Revisiting Again Document Length Hypotheses TREC-2004 Genomics Track Experiments at Patolis. *Proceedings of the Thirteenth Text Retrieval Conference*, 2004.

Hersh, W., Cohen, A., Yang, J., Bhuptiraju, R.T., Toberts, P. and Hearst, M. TREC 2005 Genomics Track Overview. Proceedings *of the Fourteenth Text Retrieval Conference*, 2005.

Hirschman, L., Park, J., Tsujii, J., Wong, L. and Wu, C.H. Accomplishments and Challenges in Literature Data Mining for Biology. *Bioinformatics*, 18(12): 1553-1561, 2002.

Hou, W.J., Lee, C., Lin, K.H.Y. and Chen, H.H. A Relevance Detection Approach to Gene Annotation. *Proceedings of the First International Symposium on Semantic Mining in Biomedicine*, http://ceur-ws.org, 148: 15-23, 2005.

Hsu, C.W., Chang, C.C. and Lin, C.J. A Practical Guide to Support Vector Classification. http://www.csie.ntu.edu.tw /~cjlin/libsvm/index.html, 2003.

Humphreys, B.L., Lindberg, D.A., Schoolman, H.M. and Barnett, G.O. The Unified Medical Language System: an Informatics Research Collaboration. *Journal of American Medical Information Association*, 5(1):1-11, 1998.

Regev, Y., Finkelstein-Landau, M. and Feldman, R. Rule-based Extraction of Experimental Evidence in the Biomedical Domain - the KDD Cup (Task 1). *SIGKDD Explorations*, 4(2):90-92, 2002.

Vapnik, V. *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.