Inferring Grant Support Types from Online Biomedical Articles

Jongwoo Kim, Daniel X. Le, and George R. Thoma
National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
{jongkim, danle, and gthoma}@mail.nih.gov

Abstract

The category of institution or organization underwriting the research reported in a scientific article is a required field (Grant Support type) in the bibliographic record of that article in the MEDLINE® database. We describe a system based on a combination of a Naive Bayes classifier and heuristic rules that automatically infers the Grant Support types from article text. Testing the performance of the system on 2,000 biomedical articles shows Precision, Recall, and F-Measure exceeding 95%.

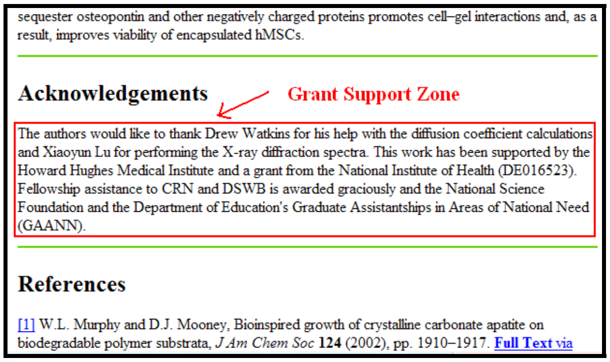
1. Introduction

The U.S. National Library of Medicine (NLM) maintains MEDLINE®, a heavily used bibliographic database of 17 million citations to the biomedical journal literature. Each citation consists of bibliographic information such as article title, author names, affiliations, grant numbers, grant support, etc. While NLM receives most such citations in XML format directly from journal publishers, key bibliographic information is often missing, requiring manual entry.

The *type* of Grant Support (GS) is typically one such missing item. GS identifies the funding sources that support the research being reported. There are fourteen types of funding sources that MEDLINE currently reports. This is important information that granting organizations and policymakers use to track research support.

Figure 1 shows a typical article that includes a text zone with GS information ("GS zone"), bounded here by a box. Of the three sentences in this zone, only the second and third sentences contain GS information ("GS sentences"). We will analyze this example in more detail in the next section.

To find GS information manually, professional indexers have to carefully search article text to find GS sentences. These sentences are usually located in the first or last page of an article, but can occur anywhere. Then, from organization names in these sentences, they deduce GS types. The indexers must therefore develop this skill by memorizing the relationships between organization names and GS types. The work is labor-intensive and often error-prone; hence our interest in an automated approach.



sequester osteopontin and other negatively charged proteins promotes cell–gel interactions and, as a result, improves viability of encapsulated hMSCs.

Acknowledgements

The authors would like to thank Drew Watkins for his help with the diffusion coefficient calculations and Xiaoyun Lu for performing the X-ray diffraction spectra. This work has been supported by the Howard Hughes Medical Institute and a grant from the National Institute of Health (DE016523). Fellowship assistance to CRN and DSWB is awarded graciously and the National Science Foundation and the Department of Education's Graduate Assistantships in Areas of National Need (GAANN).

References

[1] W.L. Murphy and D.J. Mooney, Bioinspired growth of crystalline carbonate apatite on biodegradable polymer substrata, *J Am Chem Soc* **124** (2002), pp. 1910–1917. Full Text via

Figure 1. An Example of Grant Support information.

Our approach is a two-step process, first the classification of GS sentences, and then the categorization of GS types from these sentences.

In previous work, we had developed rule-based algorithms to classify GS sentences in HTML-formatted articles using a combination of key words [1, 2]. However, this process was found to create over- or under-classification problems when authors use unusual words to express GS information. We therefore focus on a machine learning approach in our current work. The Naïve Bayes classifier is a commonly used technique for text classification [3]. Due to its simplicity, efficiency, and speed, it is widely used in document classification. Madgan [4] used the classifier to filter spam emails using frequently-occurring words in the emails as features. McCallum *et al.* [5] implemented two models of the classifier, Multi-variate Bernoulli and Multinomial models, and compared their performance using several types of documents including Web documents. They used an entropy theory to select word features. Rennie *et al.* [6] modified Multinomial model of the classifier to improve its performance in classifying documents such as newsgroups, newswire articles, etc. Tanabe *et al.* [7] used the classifier to classify documents with or without gene names in MEDLINE abstracts. Nobata *et al.* [8] also used Naïve Bayes to identify medical terms in the MEDLINE abstract. We therefore adapt a Naïve Bayes classifier to classify GS sentences.

Once sentences are classified, an automatic categorizer is employed to infer GS types using organizational names appearing in GS sentences. However, such names may not always represent granting organizations. Analysis of these GS sentences is needed to recognize actual granting organizations. Another issue is the handling of unknown organizational names in these sentences that frequently generate over-categorization problems. Categorization of GS types is a more complicated problem than Named-entity-recognition [9], and we do not find any related work or effective machine learning algorithms for its solution. We have therefore developed several heuristic rules to address this problem.

The paper is organized as follows. The definition of GS is given in Section 2. The details of our method using the Naïve Bayes classifier and heuristic rules are presented in Section 3. We report experimental results in Section 4, and conclusions in Section 5.

2. Grant Support (GS)

2.1 GS and GS types

GS identifies the funding sources for the research reported in an article. A GS sentence usually contains organizational names, words that suggest funding support (e.g., "supported", "funded", "financed", etc), and/or grant numbers. There are fourteen types of GS corresponding to the funding sources, as shown in Table 1. For example, "NIH Extramural" signifies support provided by one of the sub-organizations of the National Institutes of Health (NIH). There are also funding organizations outside the US government, such as Wellcome Trust, HHMI, and seven other UK organizations which are classified as "Non US Gov". Therefore, when Wellcome Trust appears in a GS sentence, the GS types assigned are "Wellcome Trust" and "Non US Gov".

We now apply the categorization to the GS sentences in Figure 1. Consider the second sentence:

> "This work has been supported by the Howard Hughes Medical Institute and a grant from the National Institutes of Health (DE016523)."

Here, Howard Hughes Medical Institute and the National Institutes of Health are funding organizations and DE016523 is the NIH grant number. The GS types in this case would be "NIH-Extramural" (because of NIH), "Howard Hughes Medical Institute", and "Non US Gov" (because HHMI is outside the government.)

Consider also the third sentence:

> "Fellowship assistance to CRN and DSWB is awarded graciously and the National Science Foundation and the Department of Education's Graduate Assistantships in Areas of National Need (GAANN)."

Since National Science Foundation and Department of Education are government agencies outside the Public Health Service, the GS type is "US Gov Non-PHS".

As shown in these examples, a GS sentence is similar to an affiliation since they both use organizational names. ("Affiliation" is a bibliographic item in MEDLINE signifying the authors' institution.) Consequently, a GS sentence classifier can mislabel an affiliation as a GS sentence, or vice versa.

Table 1. Grant Support types

Grant Support Type	Comment
NIH Extramural	Support to outside organizations from one of the institutes at the National Institutes of Health (NIH) (NLM, NCI, etc.)
NIH Intramural	Support from NIH institutes for internal research.
US Gov PHS	Support from a government agency other than NIH, but within the Public Health Service (PHS), such as FDA, CDC, ATSDR, HIS, HRSA, SAMHSA, AHRQ.
US Gov Non-PHS	Support from a government agency other than PHS, such as NSF, USDA, DoE, etc.
Non US Gov	Support from universities, companies, private institutions, foreign countries, and other organizations not related to US government. [*It also includes "Wellcome Trust", "HHMI" and seven British organizations below.]
Wellcome Trust	Support from Wellcome Trust
Howard Hughes Medical Institute (HHMI)	Support from Howard Hughes Medical Institute
Arthritis Research Campaign (UK)	Support from Arthritis Research Campaign (UK)
Biotechnology and Biological Sciences Research Council (UK)	Support from Biotechnology and Biological Sciences Research Council (UK)
British Heart Foundation	Support from British Heart Foundation
Cancer Research UK	Support from Cancer Research UK
Chief Scientist Office (UK)	Support from Chief Scientist Office (UK)
Department of Health (UK)	Support from Department of Health (UK)
Medical Research Council (UK)	Support from Medical Research Council (UK)

2.2 Grant number

GS types may be inferred if a Grant number is mentioned, for example, “DE016523” in the second sentence in Figure 1. The format of a grant number is unique: a two-letter code followed by a five or six-digit number. In this example, the two-letter code “DE” stands for the National Institute of Dental Research within NIH. Thus, the grant number itself is key information that can be used to infer the GS type in the absence of an organizational name. For instance, suppose the authors of our example write the second sentence without mentioning a funding organization, as follows:

> “This work has been supported by the Howard Hughes Medical Institute and a grant DE016523.”

In this case, we would examine the grant number in order to infer the corresponding GS type.

Fortunately, all US PHS organizations share the same grant number format [10]. It consists of six parts as shown in Table 2. The most important parts are the administering organization code (two-letter code) and the serial number (five to six-digit number). Since every organization in US PHS uses a different administering organization code, we can use it to infer the funding organization and the corresponding GS type.

Table 2. Official format of Grant numbers of US Public Health Service

Part	**Explanation**	**Example** **2R01 LM12345 02S1A2**
Application Type	A single-digit code identifying the type of application received and processed.	2 (a supplemental request for additional fund)
Activity Code	A three-digit code identifying a specific category of extramural activity.	R01 (Research Project)
Administering Organization Code	A two-letter code identifying the first major-level subdivision.	LM (National Library of Medicine)
Serial Number	A five (or six)-digit number assigned sequentially to a series with an institute, center, or division.	12345
Suffix Grant Year	A two-digit number indicating the segment or budget period of a project.	02 (grants in their second year)
Suffix Other	A four digit code composed of Supplement (S), Amendment (A), or Allowance(X).	S1A2

3. Our approach

As seen in Figure 1, a zone may contain both GS and non-GS sentences. Therefore, correct classification of GS sentences in a zone is very important. Our classification procedure has the following three steps.

1. Segment a zone into sentences using delimiters.
2. Classify GS sentences using our Naïve Bayes classifier.
3. Categorize GS types from each GS sentence using heuristic rules (GS type classifier).

In Step 1, the three punctuation marks (“!”, “?”, and “.”) are used as delimiters to segment a zone into sentences. Our Naïve Bayes classifier and GS type classifier, in Steps 2 and 3, respectively are discussed next.

3.1 Naïve Bayes classifier

Assume that we have a binary feature vector from a sentence $\boldsymbol{x}=(x_1, x_2, x_3, \ldots, x_m)$ where m is the dimension of the vector and $x_i= 0$ or 1 means absence or presence of the ith feature (feature refers to word in our case) in the vector. Assume there are two classes C_r and C_n: relevant and non-relevant classes. In this paper, GS sentences belong to C_r and Non-GS sentences belong to C_n. The decision function can be written as

$$P(\boldsymbol{x}|C_r)\,P(C_r) > P(\boldsymbol{x}|C_n)\,P(C_n) \qquad (1)$$

where $P(C_j)$ is the prior probability of C_j.

Assume that feature x_i in feature vector $\boldsymbol{x}=(x_1, x_2, \ldots, x_m)$ is stochastically independent. Let us define p_i as the probability of occurrence of a word (suitable as a feature) in a sentence that is in a relevant class, and q_i as the probability of such a word in a non-relevant sentence. Then, $P(\boldsymbol{x}|C_j)$ can be rewritten as

$$P(\boldsymbol{x} \mid C_r) = \prod_{i=1}^{m} p_i^{x_i}(1-p_i)^{1-x_i} \qquad (2)$$

$$P(\boldsymbol{x} \mid C_n) = \prod_{i=1}^{m} q_i^{x_i}(1-q_i)^{1-x_i} \qquad (3)$$

where $p_i = P(x_i=1|C_r)$ and $q_i = P(x_i=1|C_n)$.

When we insert Equations (2) and (3) into Equation (1), take logs, and move the right term to the left, we have the following linear decision function $G(\boldsymbol{x})$:

$$G(\boldsymbol{x}) = \sum_{i=1}^{m} \log\frac{p_i(1-q_i)}{q_i(1-p_i)}x_i + \sum_{i=1}^{m}\log\frac{(1-p_i)}{(1-q_i)} + \log\frac{P(C_r)}{P(C_n)} \qquad (4)$$

When $G(\boldsymbol{x})$ is positive, $\boldsymbol{x}$ belongs to C_r. If not, $\boldsymbol{x}$ belongs to C_n. We use this equation to classify the GS sentence. To decide on feature selection, the following equation is used [11] to extract features more related to the relevant class.

$$|\log \frac{p_i(1-q_i)}{q_i(1-p_i)}| \quad \geq \quad t \qquad (5)$$

When a feature candidate x_i satisfies the above criterion (greater than or equal to the threshold t) in Equation (5), we choose x_i as one of the features in $\boldsymbol{x}=(x_1, x_2, x_3, \ldots , x_m)$. We use t=1 in our experiment, though this may be varied in future work. In this paper, x_i stands for a word (a frequently occurring one) selected from sentences with and without GS.

3.2 GS type classifier

After a sentence is classified as a GS sentence by the Naïve Bayes classifier, a string matching method looks for organization names that suggest a particular GS type, and if a known name is found in the sentence, this sentence is categorized as the corresponding GS type. The matching is done against fourteen organization name lists. Table 1 shows examples of such names. For instance, the organization names for the GS type "NIH Extramural" are NIH, NLM, NCI, and so on.

Thirteen complete lists are created for GS types in Table 1, but for "Non US Gov" a comprehensive list is impractical, especially in the case of universities, schools, private companies, institutes, and other non-governmental organizations. Instead we create a list of terms that suggest these entities, as shown in the Organization list in the first row of Table 3. On the other hand, we do create lists of the largest 140 pharmaceutical companies, American states, and 155 countries, as indicated in the table.

As mentioned in Section 2, grant numbers contain clues suggesting GS types, particularly the administering organization codes, as shown in Table 2. We have therefore built lists of these codes that are part of grant numbers associated with NIH and PHS organizations.

Table 3. Four word lists for "Non US Gov"

Word List	Words
Organization	University, School, Center, Institute, Council, Hospital, Inc, Co., etc.
Pharmaceutical Company	Johnson & Johnson , Pfizer, Bayer, Novartis, AstraZeneca, Merck, Wyeth, etc.
US State	Alabama, California, Delaware, Florida, Georgia, Kansas, Maryland, Missouri, etc.
Country	Korea, Albania, Belgium, Canada, China, Germany, Hungary, Japan, UK, etc.

The string matching method that uses organization names and word lists cannot always categorize GS types correctly. It is susceptible to over or under categorization problems.

The first example in Table 4 shows the difficulty in inferring "Non US Gov" support from the sentence, since there is no information about "CBRS" or "CEA" in our word lists. This is an example prone to under or over categorization problem.

In the second example, we recognize "LM23456" as a NIH grant number because it follows the PHS/NIH grant format, and "LM" is the administering organization code of NIH's National Library of Medicine. This allows an inference of the GS type "NIH Extramural". While "University of Missouri" suggests "Non US Gov", the presence of the NIH grant number confirms that the research is supported by NIH's Extramural funds.

To resolve this kind of possible misclassification, we have developed a procedure to incrementally delete neighboring words before and after each organization name or grant number found in a GS sentence, using the following symbols and words to define the scope of each deletion: ",", ";", ":", "/", "by", and "and". This procedure is outlined in Table 5. In Step 1, the sentence is classified as "NIH Extramural" due to the NLM grant "LM 23456". Therefore, we remove "grant from University of Missouri (LM 23456)" because "by" and "," occur before and after "LM 23456". This step eliminates over-categorization because "University of Missouri" belongs to "Non US Gov". In Step 2, the sentence is classified as "US Gov PHS" due to the keyword "CDC". "CDC" is now removed from the sentence. In Step 3, we remove all other organization names that belong to "US Gov PHS" (FDA, in this case), leaving no organization word in the final step.

Table 6 shows the complete list of rules used in the classifier. The first rule categorizes the sentence as "NIH Extramural" or "US Gov PHS" support when there is a relevant grant number. Then, the rule deletes words around the grant number within the sentence using the scope delimiters.

The other rules are similar to the first one and are applied in the sequence specified in the table.

Table 4. Examples of Grant Support sentences

GS Type	Example of a sentence
Non US Gov	This research was supported by CBRS and CEA research programs.
NIH Extramural	This work was supported by grant from University of Missouri (**LM 23456**).

Table 5. An example of the deleting procedure.

Step	GS Sentence	GS Type
1	This work was supported by grant from University of Missouri (**LM 23456**), FDA, and CDC.	NIH Extramural
2	This work was supported by, FDA, and **CDC.**	US Gov PHS
3	This work was supported by, **FDA**, and.	US Gov PHS
4	This work was supported by,, and.	None

Table 6. Rules for the GS type classifier

No	Rules
1	If a sentence has a grant number from *X*, Categorize the sentence as *X*. Delete words in the sentence using the grant number. *X* ∈ {“NIH Extramural”, “US Gov PHS”}
2	If a sentence has an organization name from *X*, Categorize the sentence as *X*. *X*∈ {“NIH Extramural”,“NIH Intramural”,“US Gov PHS”} Delete words in the sentence using the organization name.
3	If a sentence has an organization name from *X*, Categorize the sentence as *X*. Delete words in the sentence using the organization name. *X* ∈ {all remaining GS types excluding “Non US Gov”}
4	If a sentence has an organization name or a word from *X*, Categorize the sentence as *X*. Delete words in the sentence using the organization name. *X* ∈ {“Non US Gov”}
5	If a sentence has “the” followed by a word with all capital characters or with a capital character in the first character, Categorize the sentence as *X*. Delete words in the sentence using the word. *X* ∈ {“Non US Gov”}
6	If a sentence was not labeled by the previous rules but, has a support word (“supported”, “granted”, etc.) Categorize the sentence as *X*. *X* ∈ {“Non US Gov”}

4. Experimental results

4.1 Naïve Bayes classifier

To train the Naïve Bayes classifier, we select 23,500 sentences from medical articles published in 2006. 5,142 of these sentences contain GS information (relevant class) and 18,538 sentences do not (non-relevant class). To obtain features for the classifier, we collect 6,870 of the *most frequently occurring* words in these sentences, and select 4,721 words as “general” features using the criterion expressed in Equation (5). In addition, we use three “special features”. Examples of both types of features are shown in Table 7. To test the classifier, we select a different set of 2,000 medical articles whose citations, including the GS type field, already exist in MEDLINE.

In the procedure shown in Section 3, it is assumed that the article is already segmented into text zones, following which the Naïve Bayes classifier reads and classifies every sentence in these zones. When any of the sentences in a zone is classified as a GS sentence, we label the entire zone as a “GS zone”, allowing us to test the performance of the classifier at the zone level.

The decision function used by the classifier as expressed in Equation (4) requires probability figures (p_i and q_i) that reflect the relative occurrence of the word features in GS and non-GS sentences respectively. These are computed from the words in these sentences, and shown in Table 7. For example, the word “health” occurs in about 54% of the GS sentences in the relevant class, while it appears in 4% of the non-GS sentences. These probability figures are keys in computing the decision function for classifying sentences, and labeling the zones.

Experimental results appear in Tables 8, 9, and 10. Training results (Table 8) show 72 under-classification and 74 over-classification errors in the training set of 23,500 sentences. Testing results (Table 9) show only 2 under-classification and 61 over-classification errors in the test set of 302,268 zones. As shown in Table 10, the classifier performs well in both training and testing: the measured Precision, Recall, and F-measure exceeding 98% (training) and 97% (testing). Since the testing results are computed at the zone level, and a zone may contain more than one sentence, we may assume that the actual results are better if computed at the sentence level.

Table 7. Some word features and corresponding p_i and q_i

Feature Type	Feature	p_i	q_i
Special	Granting Organization (NIH, FDA, CDC, etc.)	0.863579	0.016217
	Support Word (supported, funded, etc.)	0.898459	0.001307
	Grant Word (grant, fund, scholarship, etc.)	0.924425	0.001046
General	national	0.657009	0.022625
	health	0.540160	0.042583
	work	0.537157	0.001438
	foundation	0.167522	0.004368
	cancer	0.105706	0.046637

4.2 GS type classifier

We use the same 2,000 articles as in the previous step to test the performance of the GS type classifier. We measure performance at the article level, counting every misclassification of the GS type as an error. As shown in Table 11, there are 27 under-categorization and 90 over-categorization errors. All 1,879 articles with a GS type and 4 articles without one are classified correctly. Table 12 shows that all three measures of performance exceed 95%.

Table 8. Training results (Naïve Bayes classifier)

Sentences (Total:23,500)	True	False
Relevant (5142)	5,070	72
Non-Relevant (18,538)	74	18,284

Table 9. Testing results (Naïve Bayes classifier)

Zones (Total:302,268)	True	False
Relevant (2,221)	2,219	2
Non-Relevant (300,047)	61	299,986

Table 10. Performance (Naïve Bayes classifier)

Data Set	Precision	Recall	F-Measure
Training	98.56%	98.60%	98.58%
Test	97.32%	99.90%	98.60%

Table 11. Testing results (GS type classifier)

Articles (Total:2,000)	True	False
Relevant (1,906)	1,879	27
Non-Relevant (94)	90	4

Table 12. Performance (GS type classifier)

Data Set	Precision	Recall	F-Measure
Test	95.04%	98.58%	96.78%

5. Conclusions

This paper describes a system to infer the type of grant support (GS) underwriting the research reported in a medical article. It combines a Naïve Bayes classifier and a GS type classifier to automatically classify GS sentences and categorize GS types.

As a first step, the Naïve Bayes classifier, using frequently occurring words as features, identifies those sentences containing information on grant support. From these "GS sentences" the GS type classifier (using heuristic rules and a deleting procedure) then infers the type of grant support at a Precision, Recall and F-Measure exceeding 95%.

While overall performance is relatively high, improvements are still possible. For example, though the Naïve Bayes classifier accommodates problems such as typographical errors, it misclassifies sentences containing infrequently occurring word features. In future work we plan to combine this classifier with rules (possibly generated automatically by Random Forest and Decision Tree approaches) to compensate for such shortcomings. In addition, we will seek more robust rules for the GS type classifier to correctly infer support from organizations outside the government ("Non US Gov"). These steps are expected to improve performance.

6. Acknowledgment

The authors thank Dr. Glenn Pearson for comments and suggestions. This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications

7. References

[1] D. X. Le, L. Q. Tran, J. Chow, J. Kim, S. E. Hauser, C. W. Moon, and G. R. Thoma, "Automated Medical Citation Records Creation for Web-Based On-Line Journals," *14th IEEE Symposium on Computer-Based Medical Systems*, Bethesda, MD, pp. 315-320, July 2001.

[2] J. Kim, D. Le, and G. R. Thoma, "Automated Labeling of Biomedical Online Journal Articles," *Proc. 9th World Multiconference on Systemics, Cybernetics and Informatics*, July, Orlando, FL, Vol. 3, pp. 406-411, 2005.

[3] D. D. Lewis, "Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval," *ECML*, The Tenth European Conference on Machine Learning, pp.4-15, 1998.

[4] D. Madigan, "Statistics and the war on spam," *Statistics*: *A Guide to the Unknown*, 4th Ed. (R. Peck, G. Casella, G. Cobb, R. Hoerl, D. Nolan, R. Starbuck and H. Stern, eds.), Thomson Brooks/Cole, Belmont, CA, pp.135–147, 2005.

[5] A. McCallum and K. Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification," *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pp.577, 1998.

[6] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the Poor Assumptions of Naïve Bayes Text Classifiers," *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 616-623, 2003.

[7] L. Tababe and W. J. Wilbur, "Tagging Gene and Protein Names in Biomedical Text," *Bioinformatics*, Vol 18, No. 8, pp. 1124-1132, 2002.

[8] C. Nobata, N. Collier, J. Tsujii, "Automatic Term Identification and Classification in Biology Texts," *Proceedings of the Natural Language Pacific Rim Symposium*, pp. 369-374, 1999.

[9] E. F. Tjong, K. Sang and F. D. Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Indepeth Named Entity Recognition," *Proc. 7 Conf. Natural Language Learning (CoNLL-2003)*, pp. 142-147, 2003.

[10] NIH, *Activity Codes, Organization Codes, and Definitions Used in Extramural Programs,* July, 2007. (http://grants.nih.gov/grants/funding/ac.pdf).

[11] S. Sohn, W. Kim, D. C. Comeau, and W. J. Wilbur, "Optimal Training Sets for Bayesian Prediction of MeSH Assignment," *Journal of the American Medical Informatics Association*, Vo. 15, No. 4, pp.546-553, 2008.