Hybrid approach combining contextual and statistical information for identifying MEDLINE citation terms

In Cheol Kim *, Daniel X. Le, George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894

ABSTRACT

There is a strong demand for developing automated tools for extracting pertinent information from the biomedical literature that is a rich, complex, and dramatically growing resource, and is increasingly accessed via the web. This paper presents a hybrid method based on contextual and statistical information to automatically identify two MEDLINE citation terms: NIH grant numbers and databank accession numbers from HTML-formatted online biomedical documents. Their detection is challenging due to many variations and inconsistencies in their format (although recommended formats exist), and also because of their similarity to other technical or biological terms. Our proposed method first extracts potential candidates for these terms using a rule-based method. These are scored and the final candidates are submitted to a human operator for verification. The confidence score for each term is calculated using statistical information, and morphological and contextual information. Experiments conducted on more than ten thousand HTML-formatted online biomedical documents show that most NIH grant numbers and databank accession numbers can be successfully identified by the proposed method, with recall rates of 99.8% and 99.6%, respectively. However, owing to the high false alarm rate, the proposed method yields F-measure rates of 86.6% and 87.9% for NIH grants and databanks, respectively.

Keywords: NIH grant numbers, databank accession numbers, online biomedical documents, hybrid approach, confidence score, contextual and statistical information

1. INTRODUCTION

MEDLINE® is the premier bibliographic online database of the National Library of Medicine (NLM), containing approximately 15 million citations and abstracts for articles from over 4,800 biomedical journals published in the United States and 80 other countries. The biomedical literature, increasingly accessed via the web, is a rich, complex, and dramatically growing resource: between 2,000 and 4,000 completed references are added to MEDLINE each day, amounting to over 623,000 new references in 2006. Thus, there is a strong incentive for automated tools for extracting this bibliographic data to minimize human labor and improve timeliness and accuracy.

Such a system has been developed by the Lister Hill National Center for Biomedical Communications (LHNCBC), a research and development division of NLM. This automated system, called the Web-based Medical Article Records System (WebMARS), analyzes and extracts bibliographic information such as title, author, affiliation, abstract, etc. from online biomedical journal articles to create citations for MEDLINE [1][2]. This paper presents a hybrid method based on contextual and statistical information incorporated in WebMARS to automatically identify two additional citation items: NIH grant numbers and databank accession numbers (referred to as “NIH grants” and “databanks”, for short).

While these items have established formats, their detection is not trivial due to the following problems: First, authors often ignore the predefined formats resulting in variations and inconsistencies (shortened, abbreviated, and slightly altered forms). Secondly, other technical terms that have a similar format such as protein name, non-NIH grants, and even ZIP code often appear together. Moreover, new types of NIH grants or databanks are added periodically. These new types may have significantly different formats such as a newly created organization code in NIH grants or new prefix in databanks, and/or different number of digits in the serial number.

Thus conventional detection methods employing hand-crafted rules based on heuristics and domain-specific word/pattern dictionaries cannot easily solve this problem due to the lack of generalization capability [3][4]. To overcome the limitation of such rule-based methods, statistical approaches based on word distributions have been developed [5][6]. However, these statistical methods often yield unreliable results in the analysis of biomedical text if appropriate training datasets are not provided, the number of words included in a given test sentence or abstract is inadequate, or certain technical words closely related to a specific term appear infrequently. While it may be possible to use machine learning techniques such as support vector machine, hidden Markov model, etc that have been reported to show a good performance in the biomedical named entity recognition research field, the time-consuming task of building a large annotated training corpus is essential [7][8].

Our proposed method first broadly extracts potential candidates for NIH grants and databanks using a rule-based method, and then calculates the confidence score of each candidate based on the relative frequency of occurrence of all individual words found in the sentence in which the candidate term appears (called "grant sentence" or "databank sentence"). Such statistical information is based on sentence-level word frequency, i.e., how frequently a given word appears in the grant or databank sentence, and is estimated using a training dataset obtained from MEDLINE. In addition, to offset statistical errors the confidence score is weighted by contextual information such as specific keywords or phrases strongly indicating the existence of NIH grants or databanks.

The remainder of this paper is organized as follows. In the next section, we introduce the features and formats of NIH grants and databanks. In addition, several real examples of these terms showing variations and inconsistencies in their format are given to illustrate the difficulties in identifying them. In Section 3, we provide a detailed description of the proposed method for identifying NIH grants and databanks based on the combination of contextual and statistical information. Section 4 provides the results of recognition experiments on HTML-formatted online biomedical documents and error analysis. Final conclusions and issues related to future research are presented in Section 5.

2. FEATURES OF NIH GRANT NUMBERS AND DATABANK ACCESSION NUMBERS

2.1 NIH grant numbers in MEDLINE citations

Journal articles often include the source of funding and support for the reported work along with the grant or contract number(s) within their body text. The extraction of these numbers is important because MEDLINE citations include them to identify funding from U.S. Public Health Service (PHS) agencies, such as the National Institutes of Health (NIH), the Centers for Disease Control and Prevention (CDC), and the Food and Drug Administration (FDA).

The NIH grant consists of five parts, each having a distinct meaning as shown in Table 1: 1) single-digit code identifying the type of application, 2) three-digit code denoting a specific category of extramural activity, 3) two-letter code designating the administering organization, 4) a five or six-digit serial number, and 5) suffixes including two-digit grant year, and others. A more detailed description of NIH grants can be found in [9].

Table 1. Format of NIH grant number

Application Type	**Activity Code**	**Administering Organization**	**Serial Number**	**Suffixes**	
				Grant Year	Other
3	R01	CA	12921(9)	04	S1A1

While a particular format for the grant number is recommended, authors and publishers present them in a variety of ways. As shown in Table 2, even in a single sentence, these numbers may be expressed in different ways. In the first example, the first grant (1-P50-CA108786-01) has all components of the required format, except for a hyphen inserted between components. The second and third grants (NS20023 and CA11898) include only two components: the administering organization and serial number. The last grant (MO1 RR 30) has some missing components and the letter 'O' incorrectly used instead of zero. In examples 2 and 3, NIH grants and those from other organizations appear together. The last example shows NIH grant (CA97022) and a US Zip code (CA 92037) having the same prefix and same number of digits. The NIH grants in these examples can hardly be detected with a simple rule-based method.

Table 2. Examples of NIH grant numbers

No	Example texts
1	Supported by National Institutes of Health Grants No. **1-P50-CA108786-01**, **NS20023** and **CA11898** and by Grant No. **MO1 RR 30** through the General Clinical Research Centers Program, National Center for Research Resources, National Institutes of Health.
2	Supported by National Institutes of Health (NIH) grant no. **CA78657**, Department of Defense grant no. **BC010002**, Aging and Alzheimer Research Center grants (A.F.), and NIH grant no. **1CA76274** (R.B.).
3	This research was supported in part by DGAPA/UNAM (DirecciÃ³n General del Personal AcadÃ©mico/Universidad Nacional AutÃ³noma de MÃ©xico) **IN207503-3**, **IN206503-3** and **IX217404**, CONACyT (Consejo Nacional de Ciencia y TecnologÃ-a) **36505-N**, USDA (United States Department of Agriculture) **2002-35302-12539** and NIH (National Institutes of Health, U.S.A.) **1R01 AI066014-01**.
4	Supported by National Institutes of Health Grants **CA97022** and **GM68487**. To whom correspondence should be addressed: The Scripps Research Institute, Dept. of Immunology, SP231, 10550 N. Torrey Pines Road, La Jolla, **CA 92037**. Tel.: 858-784-7750; Fax: 858-784-7785; E-mail: klemke@scripps.edu.

2.2 Databank accession numbers in MEDLINE citations

A databank represents a set of molecular sequences registered in a particular database. At present, as shown in Table 3, NLM indexes eleven types of databanks. These are included in the Secondary Source ID (SI) field of a MEDLINE citation. Additional information on databanks and MEDLINE SI field is available at: http://www.ncbi.nlm.nih.gov.

Each type of databank has a unique format (e.g., RefSeq consists of a two-letter prefix, followed by an underscore and a six or nine digit number; GEO has one of four prefixes, GDS, GSE, GPL, GSM followed by one or more numbers). However, many variants of these formats can also be found in the literature. The first example in Table 4 shows one standard format (NT_011786) and two slight variants (XM 658485 and NM177427) of RefSeq databank accession numbers. In addition, the presence of other technical or biological terms significantly increases the complexity of identifying databanks, as shown in the second example. Moreover, new types databanks are added periodically; for instance in 2006 and 2007, ISRCTN and PubChem were newly added. In a rule-based extraction method, this would require a continuous update of rules and word/pattern dictionaries, a manual effort.

Table 3. Eleven types of databanks indexed by MEDLINE

Type of databanks	**Description**
ClinicalTrials.gov	ClinicalTrials.gov identifier number
ISRCTN	International Standard Randomised Controlled Trial Number
GDB	Johns Hopkins University Genome Data Bank
GENBANK	GenBank Nucleic Acid Sequence Database
GEO	Gene Expression Omnibus
OMIM	Mendelian Inheritance of Man
PDB	Protein Data Bank
PIR	Protein Identification Resource
PubChem	Information on the biological activities of small molecules
RefSeq	Reference Sequence
SWISSPROT	Protein Sequence Database

Table 4. Examples of databanks, their variants and other technical or biological terms

No	**Example texts**
1	To identify new isoforms of AIF we first analyzed, by an in silico approach, human AIF (NCBI Gene Data Base accession number **NT_011786** [GenBank] , gene ID 9131). The nucleotide sequence of the pkcB mRNA were previously deposited as “Aspergillus nidulans FGSC A4 hypothetical protein” (AN5973.2; REFSEQ accession number. **XM 658485**). We named this newly discovered variant as P2X7-j because previous studies identified splice variants isoforms designated P2X7-b-P2X7-h (Ref. 25, accession numbers AY847 (298-304)), and a truncated P2X7 variant 2 (149 residues) (Ref. 26, accession number **NM177427**).
2	Sequences from this study have been deposited in GenBank under accession numbers **CY003847** to **CY006042**. This work was supported by the American Lebanese Syrian Associated Charities, a Cancer Center Support Grant (**CA 21765**), the U.S. Public Health Service (grant **AI95357**), and the Hartwell Foundation. Identical sequences were found for five strains: **MDA2833**, **MDA0990**, **HUMC1166**, **CCUG38963**, and **CCUG50611**. Queries through GenBank BLAST showed that the organisms most closely matched N. meningitidis (GenBank accession no. **AL162758** and many others) at 95.7% (1,410 of 1,473 bp).

3. HYBRID METHOD

We propose a hybrid approach combining rules as well as contextual and statistical information to identify NIH grants and databanks in HTML-formatted online biomedical documents. Figure 1 illustrates an overview of the proposed method. First, the HTML-formatted body text of an article is segmented into smaller text zones, and the zones that contain clues indicating the existence of NIH grants and databanks are located and labeled as “grant zone” and “databank zone” by other modules in WebMARS [1][10]. Next, a rule-based method is applied to extract candidate NIH grants and databanks. For each candidate, the corresponding confidence score is calculated using contextual and statistical information. The candidates with confidence scores exceeding a predefined threshold are submitted to a human operator for final verification. The detailed procedure for extracting candidates and estimating their confidence scores is described below.

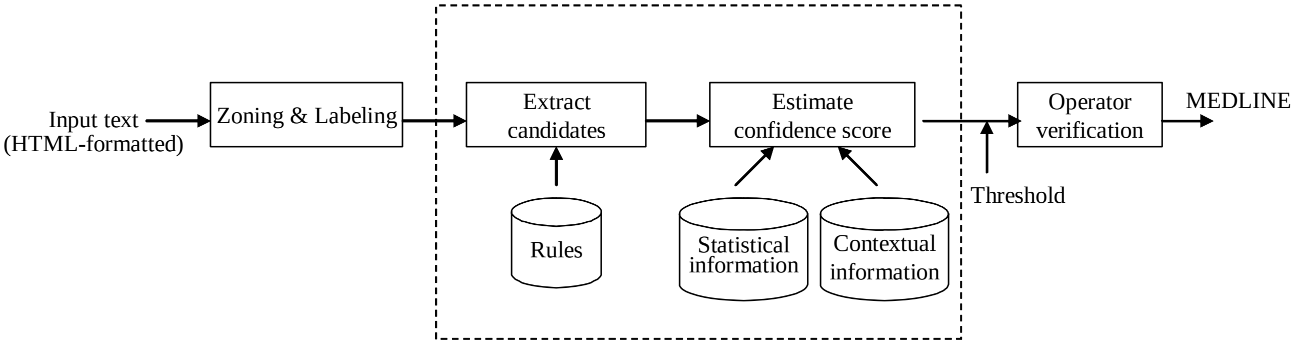

Fig. 1. An overview of the automated system proposed to identify NIH grants and databanks

3.1 Extraction of candidates for NIH grants and databanks

In this study, our goal is to minimize false rejection errors or false negative errors (i.e., real NIH grants or databanks are missed), since false alarm errors or false positive errors (other terms are misrecognized as NIH grants or databanks) can be reduced at the verification stage. To achieve our goal, we first extract any potential candidates from the text zones labeled as grant zone using the following simple rules reflecting the aforementioned characteristics of NIH grants and their variations;

1) word or word string consisting of capital letter(s) and two or more consecutive numerals

2) word or word string consisting of five or more consecutive numerals with/without lowercase letters

3) SPACE, ‘-’, and ‘/’ accepted as legitimate components

4) word or word string at least four characters long

5) Three or more numerals in word or word string.

The rules 4) and 5) are applied to the candidates that satisfy the rule 1).

The candidates for databanks are extracted by applying similar rules to the databank zones.

3.2 Estimation of confidence score

Once all possible candidates for NIH grants or databanks are extracted, their confidence scores are calculated based on the statistical information, sentence-level frequency of the words contained in the grant (or databank) sentence. To estimate this sentence-level word frequency reliably, we first created a large volume of training data consisting of online biomedical journal articles that were indexed by MEDLINE in 2006 and found to have NIH grants or databanks in their body text. Next, we extract the text zones containing NIH grants (databanks) from the body text of each article in training dataset. Usually, such text zones consist of several sentences: at least one grant (databank) sentence, in addition to others. Finally, we build a dictionary consisting of words appearing in grant (databank) sentence, and estimate their frequency of occurrence in the grant sentences and other sentences.

Let $N_n(c_g, w_x)$ and $N_n(c_o, w_x)$ be the number of occurrences of word w_x in grant sentence class c_g and other sentence class c_o, respectively. The probability that w_x has occurred in class c_g is then estimated by the conditional relative frequency score:

$$P(w_x \mid c_g) = \frac{N_n(c_g, w_x)}{N_n(c_g)} \tag{1}$$

where, $N_n(c_g) = \sum_i N_n(c_g, w_i)$. Similarly, $P(w_x \mid c_o)$, the conditional probability that w_x has occurred in class c_o is also estimated. The conditional probabilities for a sentence, S_x consisting of a set of words, $\{w_1, w_2, \ldots w_k\}$ in class c_g and c_o are equal to the product of the conditional probabilities of individual words by making the naïve Bayes assumption that all words in the sentence are conditionally independent of each other.

$$\begin{aligned} P(S_x \mid c_g) &= P(w_1, w_2, \ldots w_k \mid c_g) \\ &= \prod_{x=1}^{k} P(w_x \mid c_g) \end{aligned} \tag{2}$$

Assuming that the words closely related to NIH grants have a high relative frequency score in the grant sentence class, we estimate the confidence score of a given NIH grant candidate, gr_x found in the sentence, S_x based on the difference between $P(S_x \mid c_g)$ and $P(S_x \mid c_o)$. Finally, the confidence score of the candidate, gr_x normalized into the range from 0 to 1 is obtained by simply applying the sigmoid normalization described in [11].

$$Conf(gr_x) = \frac{1}{1 + \exp\{-\alpha(l_x(c_g) - l_x(c_o))\}} \tag{3}$$

Where, $l_x(c_g) = \log P(S_x \mid c_g)$ and $l_x(c_o) = \log P(S_x \mid c_o)$ are employed to avoid a floating-point underflow resulting from the product of conditional probabilities of individual words in sentence, S_x, which are between 0 and 1. α is the steepness parameter whose value is empirically determined.

Such a confidence score based on word frequency can often be unreliable when the number of words in the sentence is inadequate, or when certain technical words closely related to NIH grants or databanks occur infrequently, thereby resulting in false rejection errors.

To avoid such a problem, the confidence score is positively or negatively weighted depending on morphological and contextual information embedded in grant and databank zones. The morphological cue is based on the standard format and prefix of NIH grants and databanks mentioned in the previous section. Contextual information depends on specific keywords and phrases strongly suggesting the existence of NIH grants and databanks in a sentence. Examples of these are: “National Institutes of Health”, “supported by”, “accession number”, and name of databanks, etc., which were collected by analyzing a large number of articles that have NIH grants or databanks. So a candidate would be positively weighted if it has the correct format and prefix, or its surrounding sentence has the word or phrase listed in the lookup table. It would be considered most likely to be a NIH grant or databank. The highest scoring candidates, the ones exceeding a set threshold are presented to a human operator for final verification.

4. EXPERIMENTAL RESULTS

As mentioned earlier, this study focuses on minimizing false rejection errors because the false alarm errors can be sorted out by an operator at the final verification stage. Thus we evaluate the performance of our proposed method in terms of recall rate. A test dataset consisting of 10,237 HTML-formatted online articles from over 52 different biomedical journal titles is created for evaluating our proposed method. 8982 articles from this set are used for extracting NIH grants, and the remaining 1255 articles for the eleven types of databanks. All experiments were carried out on a Pentium IV based PC running Windows XP.

Our experiments show that most NIH grants and databanks can be successfully identified by the proposed method, with recall rates of 99.8% and 99.6%, respectively, when the threshold is set to 5. However, owing to the high false alarm rate, the proposed method yields F-measure rates of 86.6% and 87.9% for NIH grants and databanks, respectively.

An analysis of the experimental results shows that a poorly formatted NIH grant or databank is often not recognized by the proposed method. In Table 5 (a), “290-02-0024” was not recognized as NIH grant correctly because its sentence has no morphological cue and no contextual information such as specific keywords or phrase indicating the existence of NIH grant, and consists largely of words that are also commonly found in other parts of the body text of many biomedical documents, or describe other technical and biological terms, thereby generating a low confidence score (=2). Conversely, other terms such as “NC_100692” and “1057” in Table 5 (b) and (c) can be misrecognized when they follow the accepted formats, and when real NIH grants or databanks are also found within the same sentence.

Table 5. Examples of (a) false rejection, and (b) and (c) false alarm errors

Sentence	This study was conducted by the Oregon Evidence-Based Practice Center under contract to the Agency for Healthcare Research and Quality (Rockville, MD) contract **290-02-0024**, Task Order 2.
Candidates (**Confidence**)	220-02-0024 (**2**)

(a)

Sentence	This study was supported by National Heart, Lung, and Blood Institute Grants **HL-10337** and **HL-75360** (to M. L. Lindsey), **HL-65273** (to R. T. Lee), **HL-65662** (to A. J. Sinusas), **HL-45024**, **HL-97012**, **P01-48788**, and a Veterans Administration Career Development Award (to F. G. Spinale). **NC-100692** was provided through a grant from GE Healthcare (to A. J. Sinusas).
Candidates (**Confidence**)	HL-10337 (**10**), HL-75360 (**10**), HL-65273 (**10**), HL-65662 (**10**), HL-45024 (**10**), HL-97012 (**10**), P01-48788 (**8**), NC-100692 (**10**)

(b)

Sentence	We have previously calculated a quasi-atomic resolution model of the echovirus (EV) type 12Â·receptor complex based on cryo-negative stain transmission electron microscopy and image reconstruction of EV12 bound to a fragment of DAF comprising SCR3 and SCR4 (DAF34) (EM Data Bank code **1057** and Protein Data Bank code **1UPN** [PDB]) (21).
Candidates (**Confidence**)	PDB/1057 (**10**), PDB/1UPN (**10**)

(c)

5. CONCLUSIONS

In this paper, we have introduced a hybrid method based on contextual and statistical information to automatically identify two MEDLINE citation terms, NIH grant numbers and databank accession numbers from HTML-formatted online biomedical documents. These citation terms have many variations and inconsistencies in their format, similarities to other technical or biological terms, and new types added periodically. Thus simple hand-crafted rules and domain-specific word/pattern dictionaries based methods encounter substantial difficulties in effectively recognizing them from the documents.

Basically, our proposed method consists of two steps: 1) extracting potential candidates for these MEDLINE citation terms using a rule-based method and 2) calculating a confidence score for each candidate based on the statistical information, sentence-level relative frequency of occurrences of all individual words contained in the grant (databank) sentence. This confidence score is positively or negatively weighted depending on morphological and contextual information, to offset statistical errors. The candidates with confidence scores exceeding a predefined threshold are submitted to a human operator for final verification.

Our experiments conducted on more than ten thousand HTML-formatted online biomedical documents show that most NIH grant numbers and databank accession numbers can be successfully identified by the proposed method, with recall rates of 99.8% and 99.6%, respectively. However, owing to the high false alarm rate, the proposed method yields F-measure rates of 86.6% and 87.9% for NIH grants and databanks, respectively.

Future work is planned to employ a machine learning method such as support vector machine or hidden Markov models to reduce the false alarm errors, thereby improving overall performance and further reducing human labor required for correction.

ACKNOWLEDGMENT

This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

REFERENCES

[1] J. Kim, D. X. Le, and G. R. Thoma, "Automated labeling of bibliographic data extracted from biomedical online journals," *Proc. SPIE conf, Document Recognition and Retrieval*, **5010**, 47-56, San Jose, Jan. (2003).

[2] I. Kim, D. X. Le, and G. R. Thoma, "Identification of "comment-on sentences" in online biomedical documents using support vector machines," *Proc. SPIE conf, Document Recognition and Retrieval*, **6500**, 65000O (1-8), San Jose, Jan. (2007).

[3] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, "Toward information extraction: identifying protein names from biological papers," *Proc. of the Pacific Symposium on Biocomputing'98*, 705-716, Jan. (1998).

[4] S. Mukherjea, L. V. Subramaniam, G. Chanda, R. Kothari, V. Batra, D. Bhardwaj, and B. Srivastava, "Enhancing a biomedical information extraction system with dictionary mining and context disambiguation," IBM Journal of Research and Development, **48(56)**, 693-701 (2004).

[5] C. Nobata, N. Collier, and J. Tsujii, "Automatic term identification and classification in biology texts," *Proc. 5th Natural Language Processing Pacific Rim Symposium*, 369-374 (1999).

[6] M. A. Andrade and A. Valencia, "Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families," Bioinformatics, **14(7)**, 600-607 (1998).

[7] N. Collier, C. Nobata, and J. Tsujii, "Extracting the names of Genes and Gene products with a hidden Markov model," *Proc. The 18th Int'l Conf. Computational Linguistics*, 201-207, Saarbrucken, Germany (2000).

[8] J. Kazama, T. Makino, Y. Ohta, J. Tsujii, "Tuning support vector machine for biomedical named entity recognition," *Proc. Workshop on NLP in the biomedical domain*, 1-8 (2002).

[9] NIH, *Activity Codes, Organization Codes, and Definitions Used in Extramural Programs* (available at: http://grants.nih.gov/grants/funding/ac.pdf) (2002).

[10] J. Zou, D. X. Le, and G. R. Thoma, "Online medical journal article layout analysis," *Proc. SPIE-IS&T Electronic Imaging 2007*, **6500**, 65000V (1-12), San Jose, Jan. (2007).

[11] C. Sanderson and K. K. Paliwal, "Noise compensation in s person verification system using face and multiple features," Pattern Recognition, **36(2)**, 293-302, (2003).