

A Text-Mining Technique for Literature Profiling and Information Extraction from Biomedical Literature

Hisham Al-Mubaid

Abstract—Massive amounts of biomedical literature are readily available online in many forms. Huge amounts of valuable knowledge and relationships are embedded in these resources and need to be properly extracted, discovered, and utilized. Recognizing and classifying biomedical entity names and terms are important steps for developing efficient knowledge/information extraction techniques from these repositories. This research investigates and develops effective computational methods for literature profiling for the biomedical field. Specifically, this paper presents new techniques for biomedical term identification and classification. We utilize the advances in feature selection techniques (e.g., *MI*, X^2) in IR in this task to select the key features for term identification and classification. We evaluated the method using *Genia 3.0* corpus with about 3,000 to more than 34,000 biomedical terms and entity names. The outcome of this project can be applied in various fields including the *Aerospace* domain. In the aerospace field, there is a great interest in discovering the relations between certain changes in the body of astronauts and changes in structure at the levels of genes, proteins, and bindings.

MASSIVE AMOUNTS OF BIO-medical literature are readily available online to researchers in many forms: text abstracts (*PubMed* contains over 14 million biomedical abstracts¹), full text research articles, databases of protein interactions, dictionaries of gene and protein names, and much more. Huge amounts of valuable knowledge and useful information are embedded in these resources waiting to be properly extracted, discovered, and utilized. There is great need for computational techniques to utilize and extract the useful knowledge from these resources. A number of systems and software tools have been developed to utilize these overwhelming resources.²⁻⁶ Biomedical research has shown that *text mining* can be effective in this field, making text mining increasingly important and necessary for biology and medicine.

The purpose of this research is to investigate and design an effective computational method for literature profiling to extract and organize important information and relationships from biomedical literature. For that, we implemented new methods to identify and classify technical terms and entity names in biomedical texts. The methods are based on machine learning and



Dr. Hisham Al-Mubaid

can be viewed as a word classification task. We utilized feature extraction techniques like *MI* (mutual information) and X^2 (Chi-square) to select the key features in the contexts of the terms of interest. The methods were evaluated extensively with a large number of experiments. The outcome of this project can be applied in various fields including the *Aerospace* domain. In the field of aerospace, there is a great interest in discovering the relations between certain changes in the body of the astronauts (due to radiation, reduced gravity, and isolation) and the structural changes at the levels of genes, proteins, and bindings. Moreover, in aerospace, certain symptoms have need of being explained at the levels of gene or protein, so that the consequences and future complications can be known and treated in a timely manner.

Related Work

In the biomedical domain, the majority of term identification and recognition techniques target certain specific entities and terms (mostly gene and protein names); this way term identification and term classification are integrated as one task.⁷ A number of machine learning and statistically-based approaches have been proposed for term identification and classification in the past.⁷⁻⁹ For example, Morgan et al.⁸ used HMMs based on local context and simple orthographic and case variations and reported F-measure of 75% for the recognition of *Drosophila* gene names. Moreover, Shen et al.¹⁰ used POS tags and noun heads as features and achieved F-scores of 16.7% to 80%, depending on

Table 1. Results of the JNLPBA-2004 competition of Bio-Entity recognition: (recall/precision/F-score) results of each one of the participating systems and the baseline (BL), taken from Kim et al. (2004)⁹

	1978–1989 set	1990–1999 set	2000–2001 set	S/1998–2001 set	Total
[Zho04]	75.3/69.5/72.3	77.1/69.2/72.9	75.6/71.3/73.8	75.8/69.5/72.5	76.0/69.4/72.6
[Fin04]	66.9/70.4/68.6	73.8/69.4/71.5	72.6/69.3/70.9	71.8/67.5/69.6	71.6/68.6/70.1
[Set04]	63.6/71.4/67.3	72.2/68.7/70.4	71.3/69.6/70.5	71.3/68.8/70.1	70.3/69.3/69.8
[Son04]	60.3/66.2/63.1	71.2/65.6/68.2	69.5/65.8/67.6	68.3/64.0/66.1	67.8/64.8/66.3
[Zha04]	63.2/60.4/61.8	72.5/62.6/67.2	69.1/60.2/64.7	69.2/60.3/64.4	69.1/61.0/64.8
[Rös04]	59.2/60.3/59.8	70.3/61.8/65.8	68.4/61.5/64.8	68.3/60.4/64.1	67.4/61.0/64.0
[Par04]	62.8/55.9/59.2	70.3/61.4/65.6	65.1/60.4/62.7	65.9/59.7/62.7	66.5/59.8/63.0
[Lee04]	42.5/42.0/42.2	52.5/49.1/50.8	53.8/50.9/52.3	52.3/48.1/50.1	50.8/47.6/49.1
BL	47.1/33.9/39.4	56.8/45.5/50.5	51.7/46.3/48.8	52.6/46.0/49.1	52.6/43.6/47.7

the class, and reported that POS tags proved to be among the most useful features. A number of approaches employed SVM for term identification and recognition. For example, Kazama et al.¹¹ used SVMs for multi-class classification. They annotated the training data class label with B, I, and O labels to indicate that a term is *beginning*, *inside*, or *outside* the term.¹¹ The JNLPBA-2004 competition¹² included eight systems for the Bio-Entity recognition task.⁹ The competition was an open challenge, and the participants were allowed to use whatever techniques and data resources they preferred. However, the systems were evaluated using a common evaluation methodology and a common dataset. Four types of classification models were used: SVM, HMM, MEMM, and CRFs. The overall results (Table 1) showed the *recall* ranges from 50.8% to 76.0%, *precision* from 43.6% to 69.4%, and *F-score* from 47.7% to 72.6%.^{9,12}

The Techniques

A number of previous related methods utilized the words in terms of interest as features for term identification or classification.^{13,14} We also use word features to represent the biomedical terms, but the words in the context of the term are not used directly as features. Instead we select, as features, only those words having high ‘discriminating’ capabilities between the various classes of terms. These *word* features are used to represent each instance (example) of the terms in the training and testing. The method then employs machine learning (SVMs) to train classifiers with labeled (*training*) examples. So, some already labeled terms (*annotated with class labels*) are used as training examples. The classifiers will then be used to classify unseen and unlabeled examples (term instances) in the testing (*classification*) phase. One of the main contributions of this work is the way we select features for learning and classification.

Feature Selection

Assume that we have two classes, C_1 and C_2 , of labeled examples extracted from biomedical texts. Let C_1 be examples of biomedical term instances and their contexts from one category (C_1), whereas C_2 includes examples with their contexts from another category (C_2). We want to classify terms from C_1 and C_2 into their correct classes. The *term*, which belongs to either C_1 or C_2 , is what is to be classified in this case, and the words preced-

ing and following the term are its *context words*. Consequently each example in the set C_1 or C_2 can be represented as:

$$p_n \dots p_3 p_2 p_1 \langle term \rangle f_1 f_2 f_3 \dots f_n$$

where the words $p_1, p_2, p_3, \dots, p_n$ and $f_1, f_2, f_3, \dots, f_n$ are the preceding and following words (context words) surrounding the *term*, and n is called the *window size* (w). We extract all the context words $W = \{w_1, w_1, \dots, w_m\}$ from the examples in the sets C_1 and C_2 . Now, each such context word $w_i \in W$ may occur in contexts from either C_1 or C_2 or both with different frequency distributions. We want to determine that if we see a context word w_i in an ambiguous example the extent to which this occurrence of w_i suggests that this example belongs to C_1 or C_2 . Thus, we select those words w_i from W which are highly associated with either C_1 or C_2 (the highly discriminating words) as features. We utilize feature selection techniques like *mutual information* (MI) and *chi-square* (X^2)^{14,15} to select the highly discriminating words from W . We now explain how we implement and use MI and X^2 . Let us first define the notions of a , b , c , and d : From the training examples, we calculate a , b , c , and d for each context word $w_i \in W$ as follows:

- a = number of occurrences of w_i in C_1
- b = number of occurrences of w_i in C_2
- c = number of examples of C_1 that do not contain w_i
- d = number of examples of C_2 that do not contain w_i .

Then, the mutual information (MI) is defined as:

$$MI = \frac{N \cdot a}{(a + b) \cdot (a + c)},$$

where N is the total number of examples in C_1 and C_2 . And Chi-Square (X^2) is computed as:

$$X^2 = \frac{N \cdot (ad - cb)_2}{(a + c) \cdot (b + d) \cdot (a + b) \cdot (c + d)}. \quad (2)$$

When using the MI technique for feature selection, we calculate MI values for each $w_i \in W$; then we choose the top v words

$w_i \in W$ with the highest MI values as features in this term's *feature vectors*. In our experiments, we tested on v values of 10, 20, 30, 50, and 100. For example, if $v = 10$, then each training example is represented by a vector size of 10 entries (thus, v : *vector size*) such that the first entry represents the word with the highest MI value, the second entry represents the word with the second highest MI value, and so on. Then for a given training example, the feature vector entry is set to "1" if the corresponding feature word occurs in that training example and set to "0" otherwise.

Learning and Classification

We generate feature vectors from the training examples using the top words selected using MI or X^2 . Then, we use a well-established learning technique *Support Vector Machines (SVM)*¹⁶ to train classifiers with the training vectors. SVM is an inductive learning technique for two-class classification. Significant theoretical and empirical justifications exist in the literature to support *SVM*.¹⁶ We construct, for each class, one feature vector for each training example. Then we take two classes at a time and apply *SVM* to train, and the classifier (*model*) is produced. The classifier will then be used in the testing/classification phase to classify testing instances. We use *SVM^{light}* (<<http://svmlight.joachims.org>>) with the default parameters except that we adjust the cost factor (j parameter) by which training errors on positive examples outweigh errors on negative examples (*default* $j = 1$).

Evaluation and Discussion

The proposed techniques have been evaluated with a large variety of experiments using data from the *Genia 3.0* corpus. In this section, we describe the datasets, the experimental design, and then we discuss the results.

Table 2. 36 Terminal Classes of *Genia 3.0*

Class Names		
amino_acid_monomer	DNA_domain_or_region	atom
peptide	DNA_family_or_group	carbohydrate
protein_N/A	DNA_molecule	lipid
protein_complex	DNA_substructure	virus
protein_domain_or_region	RNA_N/A	mono_cell
protein_family_or_group	RNA_domain_or_region	multi_cell
protein_molecule	RNA_family_or_group	body_part
protein_substructure	RNA_molecule	tissue
protein_subunit	RNA_substructure	cell_type
nucleotide	other_organic_compound	cell_component
polynucleotide	organic	cell_line
DNA_N/A	inorganic	other_name

Dataset

Data for training and testing are taken from the *Genia* corpus version 3.0.¹⁷ This corpus is used as benchmark in most of the biomedical term/entity name related problems.^{9,12} The *Genia* corpus was developed at the University of Tokyo and constructed from *Medline*¹ by querying the terms '*human*,' '*blood cells*,' and '*transcription factors*.' From this search process, 2,000 abstracts were selected for the corpus. The identified terms in these selected documents were hand annotated with 36 classes/types, these classes are shown in Table 2. The corpus contains a total of 75,108 term occurrences.

Experimental design

We selected for testing 30 pairs of classes from the classes in Table 2. For space constraints, Table 3 contains only part of these selected classes. We used five-fold cross validation, such that we divided the data into five equal folds and repeated each experiment five times. Each time we leave one fold (20%) out for testing and use the remaining four folds (80%) for training. In the text preprocessing step, the training and testing texts were preprocessed as follows: (1) We changed all the letters into lower case; (2) Word stemming: all words converted to their

Table 3. Main Selected Class Pairs for Our Evaluations

C1	C1 Instances	C2	C2 Instances	C1 + C2 Instances	Training 80%	Testing 20%
amino_acid_monomer	780	protein_domain_or_region	990	1770	1418	352
lipid	2357	virus	2117	4474	3580	894
lipid	2357	multi_cell	1745	4102	3283	819
peptide	518	peptide	557	1075	861	214
DNA_substructure	106	protein_substructure	127	233	187	46
multi_cell	1745	virus	2117	3862	3091	771
protein_family_or_group	8247	virus	2117	10364	8292	2072
protein_family_or_group	8247	tissue	678	8925	7141	1784
protein_family_or_group	8247	lipid	2357	10604	8484	2120
cell_line	3846	lipid	2357	6203	4963	1240
.....
	142,638		30481	173119	138525	34594

Table 4. Results of the first set of experiments using different feature selection (f.s.) techniques, window size (w), and vector sizes (v)

Experiment			A	P	R	F1
f.s.	w	v				
MI	3	10	54.63	26.39	41.37	32.22
	3	20	55.86	56.84	44.08	49.65
	3	30	57.23	81.48	47.12	59.71
	3	30	57.23	81.48	47.12	59.71
X^2	3	10	69.07	70.79	69.83	70.31
	3	20	71.93	73.71	70.36	72.00
	3	30	75.17	76.09	75.01	75.55
MI	5	10	54.59	25.34	44.38	32.26
	5	20	54.78	41.5	44.82	43.10
	5	30	55.48	66.62	43.23	52.43
X^2	5	10	58.91	64.27	51.13	56.95
MI	10	10	54.66	28.12	44.55	34.48
	10	20	55.02	55.55	45.37	49.95
	10	30	55.09	60.54	45.52	51.97
X^2	10	10	57.66	66.08	46.63	54.68
	10	20	52.57	67.26	9.15	16.11
	10	30	53.96	72.74	11.72	20.19

stems using *Porter's* stemming algorithm.¹⁸ (3) *Stopword* removal: we removed all the function words (*stopwords*) like 'the', 'of', 'in', 'for', 'on',...etc. For performance metrics, we use *accuracy*, *precision*, *Recall*, and *F1-score*.

Results

First, we conducted a variety of experiments using feature selection techniques *MI* and X^2 to compare their performance. In these experiments, we changed window size w (number of neighboring context words) with varying vector size v as well. Consider the first class pair in Table 3 [*amino_acid_monomer*, *protein_domain_or_region*]. The first class (*amino_acid_monomer*) includes 780 annotated terms from this class, whereas the second class contains 990 annotated terms, and the total is 1,770 terms. Of these 1,770 instances, 80% (1,418 instances) were used for training, and the remaining 20% (354 instances) are used for testing. This step is repeated five times by changing the training/testing folds. We record accuracy, precision, and recall for each round, and then we take the average accuracy, precision, and recall of the five rounds.

Finally, we take the *microaverage* of accuracy, precision, and recall for all of the 30 testing pairs. Table 4 shows the results of the first set of experiments in which we changed the window size w and vector size v with the two feature selection techniques *MI* and X^2 . In this table, we notice that using windows size $w = 3$ and vector of size $v = 30$ with X^2 for feature selection produces the highest *accuracy* (75.17%) and F_1 (75.55%) results, while the best precision (81.48%) was produced with *MI* when $w = 3$ and $v = 30$. In the second set of experiments, we examined the performance after preprocessing steps. The results are in Table 6 when we applied the preprocessing steps one at a time. Table 7 contains the results when combinations of preprocessing steps were applied.

Table 5. Results of the main, larger datasets, using $w = 5, 10, v = 20, 30$, and feature selection (f.s.) is X^2

Experiment			A	P	R	F1
f.s.	w	v				
X^2	5	20	84.53	85.20	94.25	89.50
	5	30	85.07	85.67	95.33	90.24
	10	20	82.87	84.28	92.56	88.23
	10	30	84.30	85.54	93.08	89.15

Table 6. Results of the main larger datasets, $w = 5, v = 20, 30$, and feature selection (f.s.) is X^2 with preprocessing steps

Experiment			A	P	R	F1
Preprocessing	w	v				
Stemming (Porter's)	5	20	84.14	84.17	95.86	89.63
	5	30	84.38	84.71	95.16	89.63
Stopword removed	5	20	83.85	85.22	92.42	88.67
	5	30	84.20	85.74	92.42	88.96
Convert to lowercase	5	20	84.07	87.37	90.37	88.85
	5	30	84.56	74.94	78.25	76.56

These results clearly demonstrate that our technique produces impressive performance results proven by a large number and variety of experiments. Moreover, we notice that the strength of the proposed method lies mostly in the feature selection techniques and the learning/classification process. We have seen that the preprocessing steps (Table 6 and Table 7) did not improve the performance results of Table 5. Furthermore, we conclude that the best performance can be achieved when the X^2 feature selection is used.

Conclusion

Interest in bioinformatics and biotechnology is rising for many reasons, among which are the massive amounts of biomedical information and data and the significant knowledge embedded in them. The objective of this research is to devise effective computational techniques to extract and discover useful and significant knowledge from the existing repositories of biomedical literature. This paper presents new techniques for biomedical terms and entity names identification and classification to constitute an important component in an effective computational system. Experimental results showed that the method is effective in dealing with ambiguous biomedical terms using few surrounding context words as features. The strength of the method lies in the way we select these context features. We borrowed from the IR and TC domains two successful feature selection techniques (viz. *mutual information* and *Chi-square*) and proved with a variety of experiments the effectiveness of the approach. The outcome of this research can be applied in various fields. For example, in the *Aerospace* domain, certain symptoms in the body of the astronauts need explanations at the level of gene or protein of the body, so that the consequences and future complications can be known and treated in a timely manner.

Table 7. Results of the third set of experiments using combinations of preprocessing steps. In these experiments, window size $w = 5$, vector size $v = 20, 30$, and f.s. is χ^2

Experiment		A	P	R	F1
Preprocessing	v				
lowercase + stopword removed	20	84.34	86.27	93.18	89.59
	30	84.65	86.39	93.58	89.84
lowercase + word stemming	20	84.17	84.19	95.83	89.63
	30	84.33	84.66	95.16	89.60
stopword removed + word stemming	20	84.21	84.68	95.36	89.71
	30	84.36	84.89	95.20	89.75
lower case + stopword removed + word stemming	20	84.19	84.82	95.09	89.66
	30	84.38	84.91	95.12	89.72
Average		84.33	85.10	94.81	89.69

References

- ¹Medline: accessed using Entrez PubMed Interface <<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>>
- ²D. Chaussabel and A. Sher, "Mining Microarray Expression Data by Literature Profiling," *Genome Biology* 3.10 (2002): research0055.1–0055.16.
- ³C. Creighton and S. Hanash, "Mining Gene Expression Databases for Association Rules," *Bioinformatics* 19.1 (2003): 79-86.
- ⁴E. M. Marcotte, I. Xenarios, and D. Eisenberg, "Mining Literature for Protein-Protein Interactions," *Bioinformatics* 17.4 (2001): 359-63.
- ⁵Medminer, Genomics and Bioinformatics Group and SRA International Inc. <<http://discover.nci.nih.gov/textmining/filters.html>>
- ⁶T. Ono, H. Hishgaki, A. Tanigami, and T. Takagi, "Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature," *Bioinformatics* 17.2 (2001): 155-61.
- ⁷M. Krauthammer and G. Nenadic, "Term Identification in the Biomedical Literature," *J. Biomed. Info.* 37.6 (2004): 512-26.
- ⁸A. Morgan, A. Yeh, L. Hirschman, and M. Colosimo, "Gene Name Extraction Using FlyBase Resources," *Proc.*, of NLP in Biomedicine, ACL 2003, Sapporo, Japan, 2003. 1-8.
- ⁹J.-D. Kim, O. Tomoko, T. Yoshimasa, Y. Tateisi, and N. Collier, "Introduction to the Bio-Entity Recognition Task at JNLPBA," *Proc.*, Intl. Workshop on Natural Language Processing in Biomedicine and its Applications, 2004.
- ¹⁰D. Shen, J. Zhang, G. Zhou, J. Su, and C. Tan., "Effective Adaptation of Hidden Markov Modelbased Named Entity Recognizer for Biomedical Domain," *Proc.*, NLP in Biomedicine, ACL 2003, Sapporo, Japan, (2003): 49-56.
- ¹¹J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, "Tuning Support Vector Machines for Biomedical Named Entity Recognition," *Proc.*, Workshop on NLP in the Biomedical Domain, ACL 2002.
- ¹²JNLPBA-04 Workshop: <<http://www.genisis.ch/~natlang/JNLPBA04/>>. Shared task homepage: <<http://research.nii.ac.jp/~collier/workshops/JNLPBA04st.htm>>

nii.ac.jp/~collier/workshops/JNLPBA04st.htm>

¹³F. Ginter, J. Boberg, J. Jarvinen, and T. Salakoski, "New Techniques for Disambiguation in Natural Language and Their Application to Biological Text," *JMLR* 5 (2004): 605-21.

¹⁴L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization," *Proc.*, 4th European Conf. on Research and Advanced Technology for Digital Libraries, 2000.

¹⁵Y. Yang and J. P. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in *The 4th Intl Conf. on Machine Learning*, Ed. Jr. D. H. Fisher. San Francisco: Morgan Kaufman Pub., 1997. 412-20.

¹⁶V. Vapnik, *The Nature of Statistical Learning Theory*, N.Y.: Springer, 1995.

¹⁷J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA Corpus—A Semantically Annotated Corpus for Bio-Textmining," *Bioinformatics* 19 Suppl. 1 (2003): i180-82.

¹⁸M. F. Porter, "An Algorithm for Suffix Stripping," *Program* 14 (1980): 130-7.

¹⁹H. Al-Mubaid and M. Siddiqui, "Automatic Text Categorization with Learning Logic," *Proc.*, 16th Intl. Conf. for Computer Applications in Industry and Engineering, Las Vegas, NV, Nov. 11–13, 2003.

²⁰G. Boetticher, H. Al-Mubaid, and K. Frasier-Scott, "Automated Hybridization of Machine Learners for Recursive Spot Identification, Optimization, and Gel Matching of 2-Dimensional Gel Electrophoresis," *International Space Systems Annual Report*, 2003.

Publications

Al-Mubaid, H. "Context-Based Technique for Biomedical Term Classification," IEEE GrC-06. (*Submitted paper, 2006.*)

Al-Mubaid, H. and N. Ghaffari. "A New Gene Selection Technique Using Feature Selection Methodology Gene Selection," CATA-2006. (*Accepted paper.*)

Presentations

See "Natural Language Interface Models for Fast Responsiveness Applications"