Identifying Sections in Scientific Abstracts using Conditional Random Fields

Kenji Hirohata[†]
hirohata@nii.ac.jp

Naoaki Okazaki[†]
okazaki@is.s.u-tokyo.ac.jp

Sophia Ananiadou[‡]
sophia.ananiadou@manchester.ac.uk

Mitsuru Ishizuka[†]
ishizuka@i.u-tokyo.ac.jp

[†]Graduate School of Information
Science and Technology,
University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan

[‡]School of Computer Science,
University of Manchester
National Centre for Text Mining (NaCTeM)
Manchester Interdisciplinary Biocentre,
131 Princess Street, Manchester M1 7DN, UK

Abstract

OBJECTIVE: The prior knowledge about the rhetorical structure of scientific abstracts is useful for various text-mining tasks such as information extraction, information retrieval, and automatic summarization. This paper presents a novel approach to categorize sentences in scientific abstracts into four sections, *objective*, *methods*, *results*, and *conclusions*. **METHOD**: Formalizing the categorization task as a sequential labeling problem, we employ Conditional Random Fields (CRFs) to annotate section labels into abstract sentences. The training corpus is acquired automatically from Medline abstracts. **RESULTS**: The proposed method outperformed the previous approaches, achieving 95.5% per-sentence accuracy and 68.8% per-abstract accuracy. **CONCLUSION**: The experimental results showed that CRFs could model the rhetorical structure of abstracts more suitably.

1 Introduction

Scientific abstracts are prone to share a similar rhetorical structure. For example, an abstract usually begins with the description of background information, and is followed by the target problem, solution to the problem, evaluation of the solution, and conclusion of the paper. Previous studies observed the typical move of rhetorical roles in scientific abstracts: *problem*, *solution*, *evaluation*, and *conclusion* (Graetz, 1985; Salanger-Meyer, 1990; Swales, 1990; Orăsan, 2001). The American National Standard Institute (ANSI) recommends authors and editors of abstracts to state the *purpose*, *methods*, *results*, and *conclusions* presented in the documents (ANSI, 1979).

The prior knowledge about the rhetorical structure of abstracts is useful to improve the performance of various text-mining tasks. Marcu (1999) proposed an extraction method for summarization that captured the flow of text, based on Rhetorical Structure Theory (RST). Some extraction methods make use of cue phrases (e.g., “in conclusion”, “our investigation has shown that ...”), which suggest that the rhetorical role of sentences is to identify important sentences (Edmundson, 1969; Paice, 1981). We can survey the problems, purposes, motivations, and previous approaches of a research field by reading texts in background sections of scientific papers. Tbahriti (2006) improved the performance of their information retrieval engine, giving more weight to sentences referring to *purpose* and *conclusion*.

In this paper, we present a supervised machine-learning approach that categorizes sentences in scientific abstracts into four sections, *objective*, *methods*, *results*, and *conclusions*. Figure 1 illustrates the task of this study. Given an unstructured abstract *without* section labels indicated by boldface type, the proposed method annotates section labels of each sentence. Assuming that this task is well formalized as a sequential labeling problem, we use Conditional Random Fields (CRFs) (Lafferty et al., 2001) to identify rhetorical roles in scientific abstracts.The proposed method outperforms previous approaches to this problem, achieving 95.5% per-

OBJECTIVE: This study assessed the role of adrenergic signal transmission in the control of renal erythropoietin (EPO) production in humans. **METHODS**: Forty-six healthy male volunteers underwent a hemorrhage of 750 ml. After phlebotomy, they received (intravenously for 6 hours in a parallel, randomized, placebo-controlled and single-blind design) either placebo (0.9% sodium chloride), or the beta 2-adrenergic receptor agonist fenoterol (1.5 microgram/min), or the beta 1-adrenergic receptor agonist dobutamine (5 micrograms/kg/min), or the nonselective beta-adrenergic receptor antagonist propranolol (loading dose of 0.14 mg/kg over 20 minutes, followed by 0.63 micrograms/kg/min). **RESULTS**: The AUCEPO(0-48 hr)fenoterol was 37% higher (p ¡ 0.03) than AUCEPO(0-48 hr)placebo, whereas AUCEPO(0-48 hr)dobutamine and AUCEPO(0-48 hr)propranolol were comparable with placebo. Creatinine clearance was significantly increased during dobutamine treatment. Urinary cyclic adenosine monophosphate excretion was increased only by fenoterol treatment, whereas serum potassium levels were decreased. Plasma renin activity was significantly increased during dobutamine and fenoterol infusion. **CONCLUSIONS**: This study shows in a model of controlled, physiologic stimulation of renal erythropoietin production that the beta 2-adrenergic receptor agonist fenoterol but not the beta 1-adrenergic receptor agonist dobutamine is able to increase erythropoietin levels in humans. The result can be interpreted as a hint that signals for the control of erythropoietin production may be mediated by beta 2-adrenergic receptors rather than by beta 1-adrenergic receptors. It appears to be unlikely that an increase of renin concentrations or glomerular filtration rate is causally linked to the control of erythropoietin production in this experimental setting.

Figure 1: An abstract with section labels indicated by boldface type (Gleiter et al., 1997).

sentence accuracy and 68.8% per-abstract accuracy.

This paper is organized as follows. Section 2 describes previous approaches to this task. Formalizing the task as a sequential-labeling problem, Section 3 designs a sentence classifier using CRFs. Training corpora for the classifier are acquired automatically from the Medline abstracts. Section 4 reports considerable improvements in the proposed method over the baseline method using Support Vector Machine (SVM) (Cortes and Vapnik, 1995). We conclude this paper in Section 5.

2 Related Work

The previous studies regarded the task of identifying section names as a text-classification problem that determines a label (section name) for each sentence. Various classifiers for text categorization, Naïve Bayesian Model (NBM) (Teufel and Moens, 2002; Ruch et al., 2007), Hidden Markov Model (HMM) (Wu et al., 2006; Lin et al., 2006), and Support Vector Machines (SVM) (McKnight and Arinivasan, 2003; Shimbo et al., 2003; Ito et al., 2004; Yamamoto and Takagi, 2005) were applied.

Table 1 summarizes these approaches and performances. All studies target scientific abstracts except for Teufel and Moens (2002) who target scientific full papers. Field *classes* show the set of section names that each study assumes: background (B), objective/aim/purpose (O), method (M), result (R), conclusion (C), and introduction (I) that combines the background and objective. Although we should not compare directly the performances of these studies, which use a different set of classification labels and evaluation corpora, SVM classifiers appear to yield better results for this task. The rest of this section elaborates on the previous studies with SVMs.

Shimbo et al. (2003) presented an advanced text retrieval system for Medline that can focus on a specific section in abstracts specified by a user. The system classifies sentences in each Medline abstract into four sections, *objective*, *method*, *results*, and *conclusion*. Each sentence is represented by words, word bigrams, and contextual information of the sentence (e.g., class of the previous sentence, relative location of the current sentence). They reported 91.9% accuracy (per-sentence basis) and 51.2% accuracy (per-abstract basis[1]) for the classification with the best feature set for quadratic SVM. Ito et al. (2004) extended the work with a semi-supervised learning technique using transductive SVM (TSVM).

Yamamoto and Takagi (2005) developed a system to classify abstract sentences into five sections, *background*, *purpose*, *method*, *result*, and *conclusion*. They trained a linear-SVM classifier with features such as unigram, subject-verb, verb tense, relative sentence location, and sentence score (average TF*IDF score of constituent words). Their method achieved 68.9%, 63.0%, 83.6%, 87.2%, 89.8% F-scores for classifying *background*, *purpose*, *method*, *result*, and *conclusion* sentences respectively. They also reported the classification performance of *introduction* sentences, which combines *background* and *purpose* sentences, with 91.3% F-score.

[1] An abstract is considered correct if all constituent sentences are correctly labeled.

Methods	Model	Classes	Performance (reported in papers)
Teufel and Moens (2002)	NBM	(7 classes)	44% precision and 65% recall for *aim* sentences
Ruch et al. (2007)	NBM	O M R C	85% F-score for *conclusion* sentences
Wu et al. (2006)	HMM	B O M R C	80.54% precision
Lin et al. (2006)	HMM	I M R C	88.5%, 84.3%, 89.8%, 89.7% F-scores
McKnight and Srinivasan (2003)	SVM	I M R C	89.2%, 82.0%, 82.1%, 89.5% F-scores
Shimbo et al. (2003)	SVM	B O M R C	91.9% accuracy
Ito et al. (2004)	TSVM	B O M R C	66.0%, 51.0%, 49.3%, 72.9%, 67.7% F-scores
Yamamoto and Takagi (2005)	SVM	I (B O) M R C	91.3% (68.9%, 63.0%), 83.6%, 87.2%, 89.8% F-scores

Table 1: Approaches and performances of previous studies on section identification

3 Proposed method

3.1 Section identification as a sequence labeling problem

The previous work saw the task of labeling as a text categorization that determines the class label y_i for each sentence x_i. Even though some work includes features of the surrounding sentences for x_i, e.g. “class label of x_{i-1} sentence,” “class label of x_{i+1} sentence,” and “unigram in x_{i-1} sentence,” the classifier determines the class label y_i for each sentence x_i independently. It has been an assumption for text classification tasks to decide a class label independently of other class labels.

However, as described in Section 1, scientific abstracts have typical moves of rhetorical roles: it would be very peculiar if *result* sentences appearing before *method* sentences were described in an abstract. Moreover, we would like to model the structure of abstract sentences rather than modeling just the section label for each sentence. Thus, the task is more suitably formalized as a sequence labeling problem: given an abstract with sentences $\boldsymbol{x} = (x_1, ..., x_n)$, determine the optimal sequence of section names $\boldsymbol{y} = (y_1, ..., y_n)$ of all possible sequences.

Conditional Random Fields (CRFs) have been successfully applied to various NLP tasks including part-of-speech tagging (Lafferty et al., 2001) and shallow parsing (Sha and Pereira, 2003). CRFs define a conditional probability distribution $p(\boldsymbol{y}|\boldsymbol{x})$ for output and input sequences, $\boldsymbol{y}$ and $\boldsymbol{x}$,

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{\lambda}}(\boldsymbol{x})} \exp\left\{\boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})\right\}. \qquad (1)$$

Therein: function $\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})$ denotes a global feature vector for input sequence $\boldsymbol{x}$ and output sequence $\boldsymbol{y}$,

$$\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x}) = \sum_i \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}, i), \qquad (2)$$

i ranges over the input sequence, function $\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}, i)$ is a feature vector for input sequence $\boldsymbol{x}$ and output sequence $\boldsymbol{y}$ at position i (based on state features and transition features), $\boldsymbol{\lambda}$ is a vector where an element $\boldsymbol{\lambda}_k$ represents the weight of feature $\boldsymbol{F}_k(\boldsymbol{y}, \boldsymbol{x})$, and $Z_{\boldsymbol{\lambda}}(\boldsymbol{x})$ is a normalization factor,

$$Z_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp\left\{\boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})\right\}. \qquad (3)$$

The optimal output sequence $\hat{\boldsymbol{y}}$ for an input sequence $\boldsymbol{x}$,

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}), \qquad (4)$$

is obtained efficiently by the Viterbi algorithm. The optimal set of parameters $\boldsymbol{\lambda}$ is determined efficiently by the Generalized Iterative Scaling (GIS) (Darroch and Ratcliff, 1972) or Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) (Nocedal and Wright, 1999) method.

3.2 Features

We design three kinds of features to represent each abstract sentence for CRFs. The contributions of these features will be evaluated later in Section 4.

Content (n-gram) This feature examines the existence of expressions that characterize a specific section, e.g. “to determine ...,” and “aim at ...” for stating the *objective* of a study. We use features for sentence contents represented by: i) words, ii) word bigrams, and iii) mixture of words and word bigrams. Words are normalized into their base forms by the GENIA tagger (Tsuruoka and Tsujii, 2005), which is a part-of-speech tagger trained for the biomedical

Rank	OBJECTIVE	METHOD	RESULTS	CONCLUSIONS
1	# to	be measure	%)	suggest that
2	be to	be perform	(p	may be
3	to determine	n =	p <	# these
4	study be	be compare	) .	should be
5	this study	be determine	% .	these result

Table 2: Bigram features with high χ^2 values ('#' stands for a beginning of a sentence).

domain. We measure the co-occurrence strength (χ^2 value) between each feature and section label. If a feature appears selectively in a specific section, the χ^2 value is expected to be high. Thus, we extract the top 200,000 features[2] that have high χ^2 values to reduce the total number of features. Table 3.2 shows examples of the top five bigrams that have high χ^2 values.

Relative sentence location An abstract is likely to state *objective* of the study at the beginning and its *conclusion* at the end. The position of a sentence may be a good clue for determining its section label. Thus, we design five binary features to indicate relative position of sentences in five scales.

Features from previous/next w sentences This reproduces features from previous and following w sentences to the current sentence ($w = \{0, 1, 2\}$), so that a classifier can make use of the content of the surrounding sentences. Duplicated features have prefixes (e.g. `PREV_` and `NEXT_`) to distinguish their origins.

3.3 Section labels

It would require much effort and time to prepare a large amount of abstracts annotated with section labels. Fortunately, some Medline abstracts have section labels stated explicitly by its authors. We examined section labels in 7,811,582 abstracts in the whole Medline[3], using the regular-expression pattern:

```
^[A-Z]+([ ][A-Z]+){0,3}:[ ]
```

A sentence is qualified to have a section name if it begins with up to 4 uppercase token(s) followed by a colon ':'. This pattern identified 683,207 (ca. 9%) abstracts with structured sections.

Table 3 shows typical moves of sections in Medline abstracts. The majority of sequences in this table consists of four sections compatible with the ANSI standard, *purpose*, *methods*, *results*, and *conclusions*. Moreover, the most frequent sequence is "OBJECTIVE → METHOD(S) → RESULTS → CONCLUSION(S)," supposing that AIM and PURPOSE are equivalent to OBJECTIVE. Hence, this study assumes four sections, OBJECTIVE, METHOD, RESULTS, and CONCLUSIONS.

Meanwhile, it is common for NP chunking tasks to represent a chunk (e.g., NP) with two labels, the *begin* (e.g., B-NP) and *inside* (e.g., I-NP) of a chunk (Ramshaw and Marcus, 1995). Although none of the previous studies employed this representation, attaching B- and I- prefixes to section labels may improve a classifier by associating clue phrases (e.g., "to determine") with the starts of sections (e.g., B-OBJECTIVE). We will compare classification performances on two sets of label representations: namely, we will compare four section labels and eight labels with BI prefixes attached to section names.

4 Evaluation

4.1 Experiment

We constructed two sets of corpora ('pure' and 'expanded'), each of which contains 51,000 abstracts sampled from the abstracts with structured sections. The 'pure' corpus consists of abstracts that have the exact four section labels. In other words, this corpus does not include AIM or PURPOSE sentences even though they are equivalent to OBJECTIVE sentences. The 'pure' corpus is useful to compare the performance of this study with the previous work.

[2] We chose the number of features based on exploratory experiments.

[3] The Medline database was up-to-date on March 2006.

Rank	# abstracts	(%)	Section sequence
1	111,617	(17.6)	OBJECTIVE → METHOD(S) → RESULT(S) → CONCLUSION(S)
2	107,124	(16.9)	BACKGROUND(S) → METHOD(S) → RESULT(S) → CONCLUSION(S)
3	40,083	(6.3)	PURPOSE → METHOD(S) → RESULT(S) → CONCLUSION(S)
4	20,519	(3.2)	PURPOSE → MATERIAL AND METHOD(S) → RESULT(S) → CONCLUSION(S)
5	16,705	(2.6)	AIM(S) → METHOD(S) → RESULT(S) → CONCLUSION(S)
6	16,400	(2.6)	BACKGROUND → OBJECTIVE → METHOD(S) → RESULT(S) → CONCLUSION(S)
7	12,227	(1.9)	OBJECTIVE → STUDY DESIGN → RESULT(S) → CONCLUSION(S)
8	11,483	(1.8)	BACKGROUND →METHOD(S) AND RESULT(S) → CONCLUSION(S)
9	8,866	(1.4)	OBJECTIVE → MATERIAL AND METHOD(S) → RESULT(S) → CONCLUSION(S)
10	8,537	(1.3)	PURPOSE → PATIENT AND METHOD(S) → RESULT(S) → CONCLUSION(S)
..	...	...	...
Total	683,207	(100.0)	

Table 3: Typical sequences of sections in Medline abstracts

Representative	Equivalent section labels
OBJECTIVE	AIM, AIM OF THE STUDY, AIMS, BACKGROUND/AIMS, BACKGROUND/PURPOSE, BACKGROUND, BACKGROUND AND AIMS, BACKGROUND AND OBJECTIVE, BACKGROUND AND OBJECTIVES, BACKGROUND AND PURPOSE, CONTEXT, INTRODUCTION, OBJECT, OBJECTIVE, OBJECTIVES, PROBLEM, PURPOSE, STUDY OBJECTIVE, STUDY OBJECTIVES, SUMMARY OF BACKGROUND DATA
METHOD	ANIMALS, DESIGN, DESIGN AND METHODS, DESIGN AND SETTING, EXPERIMENTAL DESIGN,INTERVENTION, INTERVENTION(S), INTERVENTIONS, MATERIAL AND METHODS, MATERIALS AND METHODS, MEASUREMENTS, METHOD, METHODOLOGY, METHODS, METHODS AND MATERIALS, PARTICIPANTS, PATIENT(S), PATIENTS, PATIENTS AND METHODS, PROCEDURE, RESEARCH DESIGN AND METHODS, SETTING, STUDY DESIGN, STUDY DESIGN AND METHODS, SUBJECTS, SUBJECTS AND METHODS
RESULTS	FINDINGS, MAIN RESULTS, RESULT, RESULT(S), RESULTS
CONCLUSIONS	CONCLUSION, CONCLUSION(S), CONCLUSIONS, CONCLUSIONS AND CLINICAL RELEVANCE, DISCUSSION, IMPLICATIONS, INTERPRETATION, INTERPRETATION AND CONCLUSIONS

Table 4: Representative section names and their expanded sections

In contrast, the 'expanded' corpus includes sentences in equivalent sections: AIM and PURPOSE sentences are mapped to the OBJECTIVE. Table 4 shows the sets of equivalent sections for representative sections. We created this mapping table manually by analyzing the top 100 frequent section labels found in the Medline. The 'expanded' corpus is close to the real situation in which the proposed method annotates unstructured abstracts.

We utilized FlexCRFs[4] implementation to build a classifier with linear-chain CRFs. As a baseline method, we also prepared an SVM classifier[5] with the same features.

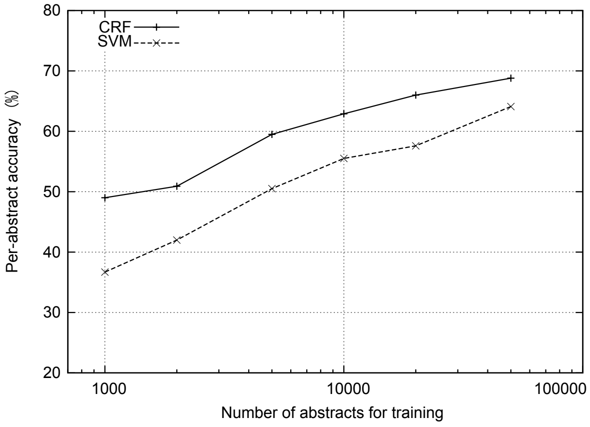

Figure 2: Training curve

4.2 Results

Given the number of abstracts for training n, we randomly sampled n abstracts from a corpus for training and 1,000 abtracts for testing. Content (n-gram) features were generated for each trainig set. We

[4] Flexible Conditional Random Field Toolkit (FlexCRFs): http://flexcrfs.sourceforge.net/

[5] We used SVM^{light} implementation with the linear kernel, which achieved the best accuracy through this experiment: http://svmlight.joachims.org/

Section labels	With B- and I- prefixes		Without B- and I- prefixes	
Features	CRF	SVM	CRF	SVM
n-gram	88.7 (42.4)	81.5 (19.1)	85.7 (33.0)	83.3 (23.4)
n-gram + position	93.4 (59.7)	88.2 (35.5)	92.4 (55.4)	89.6 (39.4)
n-gram + surrounding ($w = 1$)	93.3 (60.4)	89.9 (42.2)	92.1 (52.8)	90.0 (42.0)
n-gram + surrounding ($w = 2$)	93.7 (61.1)	91.8 (49.4)	92.8 (54.3)	91.8 (47.0)
Full	94.3 (62.9)	93.3 (55.5)	93.3 (56.1)	92.9 (52.2)

Table 5: Classification performance (accuracy) on 'pure' corpus ($n = 10,000$)

Section labels	With B- and I- prefixes		Without B- and I- prefixes	
Features	CRF	SVM	CRF	SVM
n-gram	87.7 (35.6)	78.5 (14.5)	81.9 (21.0)	80.0 (16.2)
n-gram + position	92.6 (54.3)	87.1 (31.2)	91.4 (48.7)	88.1 (31.2)
n-gram + surrounding ($w = 1$)	92.3 (52.0)	88.5 (37.6)	89.9 (44.0)	88.4 (37.1)
n-gram + surrounding ($w = 2$)	92.4 (52.5)	90.1 (41.1)	91.2 (46.6)	90.4 (41.6)
Full	93.0 (55.0)	92.0 (47.3)	92.5 (50.9)	91.7 (44.0)

Table 6: Classification performance (accuracy) on 'expanded' corpus ($n = 10,000$)

measured the classification accuracy of sentences (per-sentence accuracy) and abstracts (per-abstract accuracy). In per-abstract accuracy, an abstract is considered correct if all constituent sentences are correctly labeled.

Trained with $n = 50,000$ abstracts from 'pure' corpus, the proposed method achieved 95.5% per-sentence accuracy and 68.8% per-abstract accuracy. The F-score for each section label was 98.7% (O), 95.8% (M), 95.0% (R), and 94.2% (C). The proposed method performed this task better than the previous studies by a great margin. Figure 2 shows the training curve for the 'pure' corpus with all features presented in this paper. CRF and SVM methods performed better with more abstracts used for training. This training curve demonstrated that, with less than half the number of training corpus, the proposed method could achieve the same accuracy as the baseline method.

Tables 5 and 6 report the performance of the proposed and baseline methods on 'pure' and 'expanded' corpora respectively ($n = 10,000$). These tables show per-sentence accuracy followed by per-abstract accuracy in parentheses with different configurations of features (row) and label representations (column). For example, the proposed method obtained 94.3% per-sentence accuracy and 62.9% per-abstract accuracy with 10,000 training abstracts from 'pure' corpus, all features, and BI prefixes for class labels.

The proposed method outperformed the baseline method in all experimental configurations. This suggests that CRFs are more suitable for modeling moves of rhetorical roles in scientific abstracts. It is noteworthy that the CRF classifier gained higher per-abstract accuracy than the SVM. For example, both the CRF classifier with features from surrounding sentences ($w = 1$), and SVM classifier with full features, obtained 93.3% per-sentence accuracy in Table 5. Nevertheless, the per-abstract accuracies of the former and latter were 60.4% and 55.5% respectively: the CRF classifier had roughly 5% advantage on per-abstract accuracy over SVM. This analysis reflects the capability of CRFs to determine the optimal sequence of section names.

Additional features such as sentence position and surrounding sentences improved the performance by ca. 5–10%. The proposed method achieved the best results with all features. Another interesting discussion arises with regard to the representations of section labels. The BI representation always boosted the per-abstract accuracy of CRF classifiers by ca. 4–14%. In contrast, the SVM classifier could not leverage the BI representation, and in some configurations, even degraded the accuracy.

5 Conclusion

This paper presented a novel approach to identifying rhetorical roles in scientific abstracts using CRFs. The proposed method achieved more successful results than any other previous reports. The CRF classifier had roughly 5% advantage on per-abstract accuracy over SVM. The BI representation of section names also boosted the classification accuracy by 5%. In total, the proposed method gained more than 10% improvement on per-abstract accuracy.

We have evaluated the proposed method only on medical literatures. In addition to improving the classification performance, a future direction for this study would be to examine the adaptability of the proposed method to include other types of texts. We are planning to construct a summarization system using the proposed method.

References

ANSI. 1979. American national standard for writing abstracts. Z39.14-1979, American National Standards Institute (ANSI).

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

John N. Darroch and Douglas Ratcliff. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480.

Harold P. Edmundson. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

Christoph H. Gleiter, Tilmann Becker, Katharina H. Schreeb, Stefan Freudenthaler, and Ursula Gundert-Remy. 1997. Fenoterol but not dobutamine increases erythropoietin production in humans. *Clinical Pharmacology & Therapeutics*, 61(6):669–676.

Naomi Graetz. 1985. Teaching EFL students to extract structural information from abstracts. In Jan M. Ulijn and Anthony K. Pugh, editors, *Reading for Professional Purposed: Methods and Materials in Teaching Languages*, pages 123–135. Acco, Leuven, Belgium.

Takahiko Ito, Masashi Simbo, Takahiro Yamasaki, and Yuji Matsumoto. 2004. Semi-supervised sentence classification for medline documents. In *IPSJ SIG Technical Report*, volume 2004-ICS-138, pages 141–146.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, pages 282–289.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing (BioNLP'06)*, pages 65–72, New York City, USA.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press.

Larry McKnight and Padmini Arinivasan. 2003. Categorization of sentence types in medical abstracts. In *AMIA 2003 Symposium Proceedings*, pages 440–444.

Jorge Nocedal and Stephen J. Wright. 1999. *Numerical Optimization*. Springer-Verlag, New York, USA.

Constantin Orăsan. 2001. Patterns in scientific abstracts. In *Proceedings of Corpus Linguistics 2001 Conference*, pages 433 – 443, Lancaster University, Lancaster, UK.

Chris D. Paice. 1981. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *SIGIR '80: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 172–191, Kent, UK. Butterworth & Co.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the ACL 3rd Workshop on Very Large Corpora*, pages 82–94.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, and Anne-Lise Veuthey. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2–3):195–200.

Françoise Salanger-Meyer. 1990. Discoursal flaws in medical english abstracts: A genre analysis per research- and text-type. *Text*, 10(4):365–384.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 134–141, Edmonton, Canada.

Masashi Shimbo, Takahiro Yamasaki, and Yuji Matsumoto. 2003. Using sectioning information for text retrieval: a case study with the medline abstracts. In *Proceedings of Second International Workshop on Active Mining (AM'03)*, pages 32–41.

John M. Swales, 1990. *Genre Analysis: English in academic and research settings*, chapter 6. Cambridge University Press, UK.

Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. 2006. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *International Journal OF Medical Informatics*, 75(6):488–495.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 467–474, Vancouver, British Columbia, Canada.

Jien-Chen Wu, Yu-Chia Chang, Hsien-Chin Liou, and Jason S. Chang. 2006. Computational analysis of move structures in academic abstracts. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 41–44, Sydney, Australia.

Yasunori Yamamoto and Toshihisa Takagi. 2005. A sentence classification system for multi-document summarization in the biomedical domain. In *Proceedings of the International Workshop on Biomedical Data Engineering (BMDE2005)*, pages 90–95.