A Multiclass-based Classification Strategy for Rhetorical Sentence Categorization from Scientific Papers

Dwi H. Widyantoro[1], Masayu L. Khodra[1], Bambang Riyanto[1]& E. Aminudin Aziz[2]

[1]School of Electrical Engineering and Informatics, Bandung Institute of Technology,
Jalan Ganesa No.10, Bandung 40132, Indonesia
[2]Faculty of Language and Arts Education, Indonesia University of Education,
Jalan Dr. Setiabudhi No. 229, Bandung 40154, Indonesia
Email: dwi@stei.itb.ac.id

Abstract. Rapid identification of content structures in a scientific paper is of great importance particularly for those who actively engage in frontier research. This paper presents a multi-classifier approach to identify such structures in terms of classification of rhetorical sentences in scientific papers. The idea behind this approach is based on an observation that no single classifier is the best performer for classifying all rhetorical categories of sentences. Therefore, our approach learns which classifiers are good at what categories, assign the classifiers for those categories and apply only the right classifier for classifying a given category. This paper employs *k*-fold cross validation over training data to obtain the category-classifier mapping and then re-learn the classification model of the corresponding classifier using full training data on that particular category. This approach has been evaluated for identifying sixteen different rhetorical categories on sentences collected from ACL-ARC paper collection. The experimental results show that the multi-classifier approach can significantly improve the classification performance over multi-label classifiers.

Keywords: *acl-arc; classification strategies; multiclass approach; multi-label classification; rhetorical sentence categorization; scientific papers.*

1 Introduction

Keeping abreast of the state-of-the-art of research topics is a must for researchers and reading new papers could be a daunting task with current proliferation of scientific publication. An alternative solution that can be considered more effective is to provide readers with structured information that is extracted from a scientific paper. This structured information is represented as Rhetorical Document Profile (RDP) [1]. RDP is a representation of information that readers want to know from a paper so that readers can identify the relevance of a paper just by reading its RDP. It is an instantiated template consisting of rhetorical slots. Each slot contains a collection of sentences with a specific rhetorical category. Rhetorical sentence classification is the most important and major step in creating an RDP. This process is also known as

Copyright © 2013 Published by ITB Journal Publisher, ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2013.7.3.5

argumentative zoning [1], section classification [2], information structure identification [3], or structural analysis [4].

The majority of works in this research area has been focusing on information structuring (i.e., defining various rhetorical categories on various domains) [1],[3],[5],[6] and features selection (i.e., what features to use for representing sentences) [1],[7],[8]. Various classification methods have also been used including Naïve Bayes Model [9], Maximum Entropy [7], Support Vector Machine [10], Hidden Markov Model [11] and Conditional Random Field [8]. However, the most suitable classifier for this task remains inconclusive. To our knowledge, little performance comparison among classification methods, if any, has been reported in the literature. Additionally, most prior works performed rhetorical structures categorization on scientific abstracts and only a few have investigated sentence classification from a full paper, which has more complex rhetorical structures. This paper attempts to address this gap and reports the impact of several classification strategies to the performance of existing classifiers on full scientific papers.

Our approach to rhetorical sentence classification is based on our observation that a classifier is only good at one or several rhetorical categories. Thus, if we rely only on a single classifier, it is difficult to further improve its classification performance. We address this technical problem by involving multiple classifiers. Unlike other similar methods that usually combine the classification models of various classifiers [12],[13], our approach simply learns the best classifier for a given rhetorical category and then uses only that classifier to determine if a new sentence belongs to that category. This approach is proven to be more effective.

The main contributions of our research work are three-fold: (1) developing a standard corpus based on ACL-ARC collection that has been annotated with sixteen rhetorical categories, (2) providing performance comparison among various classifiers and classification strategies on the standard corpus, and (3) improving the classification performance of rhetorical sentences by adopting multi-classifier approach.

The rest of the paper is organized as follows. The next section provides an overview of rhetorical categories used in this paper. Section 3 describes various strategies for classifying rhetorical structures. The setup and the results of our experiments are discussed in Section 4, followed by concluding remarks in Section 5.

2 Related Work

Works on sentence classification were usually specific to a particular domain, such as biomedical information [14], computational linguistics [1],[5], legal [15] and clinical notes [2], just to name a few. Accordingly, the categories of information to be identified could vary among these domains. As an extreme example, there is a need for identifying "Past Medical History" in clinical notes but not in legal domain. Nevertheles, the structure of information in scientific abstracts are generally similar across domains. Similarly, full scientific papers also contain many similar and richer categories such as in Teufel's argumentative zoning [1],[5].

Identifying sentence category shares common issues with typical text-classification problem in that the classification performance is affected by features and classification methods. Knight & Srinivasan [10] used bag of words & sentence location as the sentence features while Lin, *et al*. [11] employed bigram language model. Merity, *et al*. [7] employed a richer feature: n-gram, the first four words of a sentence, section counter, as well as sentence positions between two sections and within a paragraph. Hirohata and colleagues [8] showed that the performance of *n*-gram language model can be improved 5-10% by incorporating sentence position and its surrounding.

Various methods have also been employed for sentence classification in previous work. Ruch, *et al*. [9] claimed that Naïve Bayes classifiers with positional heuristics outperformed expert-driven approaches in argumentative classification. The performance of SVM was shown to be superior to that of Widrow-Huf for sentence type categorization in Medical abstracts [10]. Although not based on the same exact collection, Lin, *et al*. [11] demonstrated that Hidden Markov Model (HMM) was at least competitive with SVM from performance point of view. When the task is considered as sequence labeling, methods such HMM and Conditional Random Fields outperformed the baseline methods, which assume the classification of a section is independent of the other sections [2],[8],[11].

3 Rhetorical Structures

Rhetoric is the intention information that an author wants to convey to his/her readers. To date, a number of different schemes have been proposed to structure information from the coarse-grained to finer-grained. The former classifies sentences according to section names typically found in scientific abstracts. For example, abstracts are usually divided into objective, method, results and conclusion. The finer-grained scheme is based Teufel's argumentative zoning.

This scheme is first introduced with seven categories and recently is refined into 15 categories.

This research employs the refined version for structuring information from full scientific papers and adds TEXTUAL category from the 7-category scheme. Table 1 provides category abbreviations and a short description of each category. The refined scheme is considered to be more informative, better in recognizing the structure of problem solving, and subtler in describing a difference [5].

Table 1 Argumentative zoning with 15 categories [11] + TEXTUAL category.

Category	Description
AIM	Statement of specific research goal, or hypothesis of current paper
NOV_ADV	Novelty or advantage of own approach
CO_GRO	No knowledge claim is raised (or knowledge claim not significant for the paper)
OTHR	Knowledge claim (significant for paper) held by somebody else. Neutral description
PREV_OWN	Knowledge claim (significant) held by authors in a previous paper. Neutral description.
OWN_MTHD	New Knowledge claim, own work: methods
OWN_FAIL	A solution/method/experiment in the paper that did not work
OWN_RES	Measurable/objective outcome of own work
OWN_CONC	Findings, conclusions (non-measurable)of own work
CODI	Comparison, contrast, difference to other solution (neutral)
GAP_WEAK	Lack of solution in field, problem with other solutions
ANTISUPP	Clash with somebody else's results or theory; superiority of own work
SUPPORT	Other work supports current work or is supported by current work
USE	Other work is used in own work
FUT	Statements/suggestions about future work (own or general)
TEXTUAL	Indication of paper's textual structure.

4 Classification Strategies

Identifying rhetorical sentences is a multi-label classification problem. As the first classification strategy, it can be solved by naturally extending the binary classification technique for some supervised learning algorithms such as neural networks and SVM. In Neural Networks, in particular, the binary classification will have a single neuron in its output layer. It can be easily extended to address multi-label classification problem by adding the networks' output units for encoding multiple labels. The basic SVM also handles binary classification. Extension of SMV to multiclass is conducted by providing additional parameters and constraints to the optimization problem for supporting the separation of different classes. Naïve Bayes and SVM, however, can naturally handle binary or multi-label classification problem.

The second classification strategy is to decompose the multi-label classification problem into several binary classification tasks, i.e., the problem of classifying among *N* labels is reduced into *N* binary classification problems. This approach requires *N* binary classifiers where each of them is trained to discriminate a given label from the other (*N* – 1) labels. Given an unknown example, its label will be assigned to class label of classifier that produces the maximum output. This strategy belongs to the family of Ensemble classifiers.

In the context of constructing Rhetorical Document Profile (RDP), the process of identifying a rhetorical sentence and then inserting it into a rhetorical slot in RDP can also be considered as information extraction. In this problem setting, a binary classifier, like in ensemble classifiers, is trained to learn a specific rhetorical class. The difference is mainly in the classification process. When classifying a rhetorical category, this setting uses only binary classifier that has been trained on that category, ignoring other binary classifiers (unlike in Ensemble classifiers that use all the binary classifiers). Therefore, to fill in a specific rhetorical slot, the corresponding binary classifier will be run over a text document to identify all sentences belonging to that class. The process is repeated for the rest of binary classifiers in order to fill in all slots in the RDP.

While multi-label classification with single classifier and ensemble classifier require a single pass to identify all class labels, the classification of all class labels in the information extraction setting, however, will require *N* passes where *N* is the number of class labels. Nevertheless, the strategy involving a set of binary classifiers where each of them is trained and used exclusively for classifying a specific rhetorical category seems promising and little has been investigated in the literatures. This paper reports our exploration in adopting such an approach as the third classification strategy.

The main idea of our proposed strategy for classification of *N* rhetorical categories is to train *N* independent binary classifiers where each classifier is assigned to model the classification of a particular category. Given training data containing examples of all rhetorical sentences, each n^{th} binary classifier is trained with positive examples belonging to rhetorical category *k* and negative examples belonging to the other (*N* – 1) categories. This approach is similar to the one-against-all strategy except that the classification process of each binary classifier is independent of that of the other classifiers. In general, any classification algorithm can be employed as the base classifiers.

Two alternatives of this strategy can be further developed:

1. **Multi-HO** (multi-homogeneous classifier). In this alternative, all *N* binary classifiers use the same classification algorithm (base classifier).
2. **Multi-HE** (multi-heterogeneous classifier).In the second alternative, the base classifier assigned for a specific rhetorical category is the best classifier selected among various classification algorithms that are made available to the system. Hence, the base classifier for a particular rhetorical category could be different from the base classifiers assigned for other categories.

The best classification method for the n^{th} binary classifier in multi-heterogeneous classifier (multi-HE) is obtained by (1) performing *k*-fold cross validation for all base classifiers on training data and selecting the best performer, (2) re-training the best performer on the full training data, and (3) assigning the best performer as the base classifier for rhetorical category *n*.

5 Experimental Evaluation

5.1 Data

Since there was a lack of corpus annotated with Teufel's 15-rhetorical category, we constructed our own corpus from 75 ACL-ARC papers. Each paper was annotated by three independent annotators (graduate students who were knowledgeable in computational linguistics). Differences in annotations were resolved by discussion among the annotators until they reached an agreement.

The corpus contains 10877 annotated, distinct sentences in xml format. For experiments, it is split randomly into a training set and a test set. The training set consists of sentences from two third of the total number of papers in the corpus (50 papers), and sentences of the remaining papers are used as the test set. Table 2 shows detail descriptions of the training set and the test set while Table 3 depicts the distribution of data set on each rhetorical category for multiclass classification strategies. Note that the total number of data set (#sentences) in each rhetorical category (each row in Table 3) is the same as the total number of sentences in Table 2, i.e., each rhetorical category employs the same data set.

Table 2 Description of data set.

Description	Training-set	Test-set	Total
Number of papers	50	25	75
Number sentences	7239	3638	10877

Table 3 Data set distribution for each rhetorical category for multi-class experiments.

Rhetorical Category	Training Set (#sentences)		Test Set (#sentences)	
	Positive	Negative	Positive	Negative
AIM	136	7103	77	3561
NOV_ADV	179	7060	68	3570
CO_GRO	271	6968	113	3525
OTHR	528	6711	444	3194
PREV_OWN	471	6768	150	3488
OWN_MTHD	3608	3631	1717	1921
OWN_FAIL	46	7193	24	3614
OWN_RES	264	6975	155	3483
OWN_CONC	385	6854	193	3445
CODI	69	7170	42	3596
GAP_WEAK	241	6998	124	3514
ANTISUPP	36	7203	24	3614
SUPPORT	284	6955	109	3529
USE	244	6995	196	3442
FUT	113	7126	38	3600

5.2 Features

We combined Teufel's [1], Merity's [7] and our additional features as the sentence representations. There are eight types of Teufel's features: content, absolute location, explicit structure, sentence length, verb syntax, citations, formulaic expression, and agentivity [1]. Merity proposed different values of some features like straight counter for section, location, and paragraph. We added two additional features: abstract content and qualifying adjective incidence. Table 4 provides the complete list of feature set.

Table 4 Our feature pool based on Teufel's feature types [1].

Type	Name	Description	Values
Content	Cont-1	Significant terms incidence determined by tf.idf	0,1
	Cont-2	Incidence of title or headline words determined by tf.idf	0,1
	Cont-3**	Incidence of significant terms in abstract, determined by tf.idf	0,1
Absolute location	Loc	Sentence position within document relation to 10 segments	1-10
Explicit structure	Struct-1	Sentence position within section	1-7
	Struct-2	Sentence position within paragraf	1-3
	Struct-3	Headline type	0-16
	SectCount*	Section counter	1-10
	SectLoc*	Sentence position within section (straight counter)	1-10
	ParLoc*	Sentence position within paragraf (straight counter)	1-10
Sentence length	Length	Is the sentence longer than 15 words?	0,1
Syntax	Syn	Is the 1st finite verb modified by modal auxiliary ?	0.1
	Adj**	Inicidence of qualifying adjective	0,1
Citations	Cit-1	Citation or self citation incidence	0,1,2
	Cit-2	Citation location in sentence	0,1,2,3
Formulaic expression	Formu$_{1..21}$**	Incidence of each formulaic expression in sentence	0,1
Agentivity	Ag-1$_{1..16}$**	Incidence of each agent type	0,1
	Ag-2$_{1..9}$**	Incidence of each action type	0,1
	Negation	Incidence of negation in sentence	0,1

Content features are general features in sentence extraction for determining global sentence relevance. Teufel employed TF-IDF to identify concepts that are characteristic for the contents of the document, and the *n* top-scoring words are chosen as content words. Sentence scores are computed as a weighted count of the content words in a sentence, which are then normalized by sentence length. Since an abstract consists of important sentences that can be a part of important concepts of the paper, we also incorporate it as an additional feature.

Qualifying adjectives are used to state conclusion as author's opinion based on experiment facts. Its incidence is an important feature to identify a conclusion sentence. If there is a qualifying adjective, the sentence score is 1.

Formu$_{1..21}$, Ag-1$_{1..16}$, and Ag-2$_{1..9}$ are meta-discourse features extracted by using Teufel's defined patterns [1]. Teufel only used the first occurrence of a pattern in the sentence. Since a sentence can match no pattern, one pattern, or more than one pattern, we implemented each pattern incidence as one Boolean feature.

5.3 Base Classifiers

To test the various classification strategies, this paper employs the following algorithms as the base classifiers: Naive Bayes, Logistic Regression, Multi-layer Perceptron, *k*-Nearest Neighbours, PART, C4.5, Random Tree, Random Committee, and Support Vector Machines. We used Weka's implementation of these algorithms, except for SVM from LibSVM.

Naive Bayes (NB) provides a simple approach using probabilistic knowledge with two simplifying assumptions: conditional independence of features, and no hidden attributes influence the prediction [4]. The NB model contains: (1) each class c probability $P(c)$, and (2) conditional probability of each attribute value a given a class, i.e., $P(a|c)$. Classification uses the model to find a class with maximum probability given an instance, as follows:

Logistic Regression (LR) in this paper uses a multinomial logistic regression model, which assumes that the probability of each target class can be determined from a linear combination of observed features and some problem-specific parameters. Training data are utilized to determine the optimal value of the model parameters. This classifier employs Ridge estimator, a restricted maximum likelihood estimator. Ridge estimator is used to improve the parameter estimates and to reduce the error of predictions [16].

Multi-Layer Perceptron (MLP) is a multi-layer artificial neural networs that maps inputs into appropriate outputs. It consists of multiple layers of nodes and each layer is fully connected to the next one. The MLP in this paper employs backpropagation algorithm for training the network [17]. While single perceptron can learn only linearly separable data, the learning capability of MLP is more powerful in that it is also able to distinguish data that are nonlinearly separable.

***k*-Nearest Neighbour (*k*-NN)** is a lazy learning algorithm that only stores the verbatim training examples. There is no set of abstractions model derived from training examples [18]. In classification, it searches *k* closest members of the training data and the prediction is based on the majority class of those neighbours. Thus, this classifier constructs a different approximation of target function for each new instance. Despite its simplicity, it has the advantage for learning a very complext target function that can be descibed by a set of less complex local approximations.

PART generates rules by combining rules created from decision trees and the separate-and conquer rule-learning [19]. It learns one rule at a time from tree without performing global optimization on the produced rules. A single rule is generated from a pruned, partial decision tree by selecting a leaf with the greatest covereage. This learning strategy has been claimed to improve its efficiency over similar rule-larning methods but still maintains the accuracy of classification.

C4.5 produces decision tree by top-down induction derived from the divide-and-conquer algorithm. During the tree construction, each node in the tree is generated based on a data attribute that most effectively splits its samples into subsets enriched with one class. Information gain is employed as the splitting crition, i.e., attribute with the largest information gain (difference in entropy) will be selected. The splitting process continues recursively on smaller subsets of data.

Random Tree (RT) is included in the same package as C4.5. It constructs a tree whose nodes are randomly chosen attributes. The number of chosen attributes is a parameter of its technique [17].

Random Committee (**RC**) is a classifier that is an ensemble of randomizable base classifiers. Each base classifier is built using a different random number seed of the same training data [17]. The final prediction is calculated by averaging the predictions generated by the individual base classifiers.

Support Vector Machine (**SVM**) is a learning algorithm that constructs a hyper plane with maximal margin between classes (i.e., a clear gap that is as wide as possible) [20]. It finds some support vectors, which are the training data that constrain the margin width. Learning SVM can be considered as a quadratic optimization problem subject to linear constraints. Any non-linear problems must be converted into linear problem by applying kernel trick. The SVM effectiveness depends greatly on the kernel's selection, kernel's parameter and the soft margin parameter. This classifier has been well studied as among the best classifier to date.

5.4 Results

Table 5 provides the performance comparison among various classification strategies. The multi-label classifier is a single classifier that performs multi-label classification. The ensemble classifier makes prediction based on the class label of classifier with the highest output among other *n* classifiers (*n*=number of categories). The multi-HO/HE classifier performs classification on a given category according to the prediction given by binary classifier trained in that category. As described earlier, multi-HO refers to the classification strategy that trains a binary classifier for a specific class category but all employ the same classification algorithm (i.e. homogeneous classifiers). Similar to multi-HO in that it uses a classifier trained for a specific class category, the multi-HE (heterogeneous) selects the best performer among classifiers with various classification algorithms (i.e., each category may employ different classifier from other categories).

Table 5 The accuracy of various classification strategies.

Base Classifiers	**Accuracy (%)**			
	Multi-label	**Ensemble**	**Multi-HO**[1]	**Multi-HE**[2]
SVM	51.0	50.1	80.8	79.6
NB	47.8	48.7	79.5	
C4.5	46.8	47.3	80.6	
LR	51.1	52.4	82.4	
MLP	31.1	49.5	78.2	
1-NN	38.2	38.2	76.7	
PART	40.5	42.4	78.6	
RT	35.1	32.8	77.0	
RC	47.5	47.5	80.8	

[1] Multi-Homogeneous Classifier
[2] Multi-Heterogeneous Classifier

As shown in the table, the classification strategies provided by multi-HO/HE classifier can improve the accuracy by 84% on the average over the multi-label classifier while this average improvement is only 5% by ensemble classifier. Although the total numbers of predictions performed by multi-HO/HE classifiers (i.e., `#categories x #test_set`) are different from those of performed by ensemble and multi-label classifier (i.e., only `#test_set`), the values of accuracy are still comparable to one of another because these values are derived from the same dataset.

In addition to provide significantly better accuracy of prediction, the Multi-HO classification strategy produces much stable results regardless of the base classifier employed. In particular, the performances of multi-label and ensemble classification strategies vary greatly among different base classifiers, i.e., from 32% (Ensemble-RT) to 52% (Ensemble-Logistics). This is not the case for the Multi-HO where the performance differences provided by different base classifiers are much smaller (the largest difference is only about 5%) than that of in Ensemble (about 20%).

The average accuracy of Multi-HO (79.4%) is comparable to the accuracy of Multi-HE (79.6%). Base classifiers SVM, Logistics and RC are among the best performer in Multi-HO classification strategy with the performance of at least 80.6%, while the performances of the rest of them (NB, MLP, 1-NN & RT base classifiers in Multi-HO) are worse than Multi-HE. With this result, the chance for obtaining better performance by randomly picking a base classifier in Multi-HO is about fifty percent. Multi-HE classification strategy, however, provides the safe choice.

The performance of multi-HO/HE classifiers for each rhetorical category in terms of F-measure is depicted in Table 6. The left part of the table shows the performance of multi-homogeneous classification strategy under various base classifiers. The last column contains the best classification method for each category in the Multi-HE classification strategy. Therefore, its F-measure value is the same as the F-measure of Multi-HO classifier whose base classifier is the best method found in the multi-heterogeneous classifier on the same category. For instance, in the AIM category, the best method of the multi-HE classifier is 1NN, so its F-measure is 0.39 (the same as the performance value of multi-HO under 1NN base classifier on the AIM category).

Rhetorical sentence categorization is indeed a difficult problem in that many categories are hard to correctly predict. As indicated in Table 6, each base classifier in Multi-HO strategy suffers from (near) zero performance in one or several categories. Even with the Naïve Bayes classifier in Multi-HO strategy that is superior in this particular data set and experiments, it still suffers from zero F-measure on OWN_FAIL rhetorical category (i.e., none of its prediction is correct). In such cases, C4.5 performs the worst where it completely fails (zero F-measures) to correctly predict 11 out of 16 rhetorical categories. Despite the difficulty of a base classifier in predicting certain rhetorical categories, other base classifiers are always able to predict correctly at reasonably performance on that categories.

Table 6 empirically confirms the appropriateness of Multi-HE over Multi-HO classification strategy for rhetorical sentence categorization problem. The underlined values in the table indicate the best score on the given category over other classifiers. The multi-heterogeneous (multi-HE) classifier has the best F-measure values on 11 out of 16 rhetorical categories, contributed by NB, PART, 1-NN and Logistics base classifiers. In the multi-homogeneous (multi-HO) strategy, the largest number of categories with the best F-measure is provided by the one under Naïve Bayes base classifier (9 categories). Other base classifiers contribute only from 0 to 3 categories.

Table 6 also reveals that in some cases the multi-heterogeneous classifier missed to find the truly best methods. In AIM category, for example, the best method found based on *k*-fold cross validation on training data (in Multi-HE) is 1NN base classifier, but in the test set this is not the case (the best method in this category is under Logistic base classifier as shown under Multi-HO). Similar cases are also found in NOV_ADV, PREV_OWN, OWN_MTHD and ANTISUP categories.

Table 6 F-measure for each rhetorical category and classifier.

Rhetorical Category	**Multi-HO Classification Strategy**									**Multi-HE**	
	SVM	**NB**	**C4.5**	**LR**	**MLP**	**1NN**	**PART**	**RT**	**RC**		
AIM	0.48	0.40	0.42	**0.59**	0.48	0.39	0.43	0.44	0.38	0.39	1NN
NOV_ADV	0.00	**0.17**	0.00	0.00	0.00	0.10	0.04	0.01	0.05	0.01	RT
CO_GRO	0.30	**0.32**	0.00	0.27	0.17	0.22	0.28	0.17	0.28	**0.32**	NB
OTHR	0.01	**0.25**	0.00	0.03	0.00	0.15	0.17	0.16	0.06	**0.25**	NB
PREV_OWN	0.26	0.24	**0.29**	0.24	**0.29**	0.10	0.18	0.13	0.22	0.18	PART
OWN_MTHD	0.66	0.67	0.65	**0.70**	0.69	0.60	0.62	0.60	0.65	0.66	SVM
OWN_FAIL	0.00	0.00	0.00	0.00	0.00	**0.07**	0.00	0.05	0.00	**0.07**	1NN
OWN_RES	0.00	**0.17**	0.00	0.05	0.01	0.07	0.10	0.09	0.01	**0.17**	NB
OWN_CONC	0.01	**0.22**	0.00	0.09	0.04	0.10	0.17	0.17	0.08	**0.22**	NB
CODI	0.00	0.11	0.00	0.08	0.04	0.00	**0.12**	0.03	0.00	**0.12**	PART
GAP_WEAK	0.05	**0.25**	0.00	0.08	0.10	0.06	0.14	0.10	0.03	**0.25**	NB
ANTISUPP	0.08	0.12	0.00	0.00	**0.13**	0.06	0.00	0.00	0.00	0.06	1NN
SUPPORT	0.08	**0.32**	0.00	0.16	0.07	0.14	0.20	0.13	0.04	**0.32**	NB
USE	0.05	**0.21**	0.00	0.04	0.00	0.12	0.11	0.10	0.04	**0.21**	NB
FUT	0.40	0.26	0.30	**0.43**	0.33	0.21	0.24	0.15	0.29	**0.43**	LR
TEXTUAL	0.25	**0.31**	0.26	0.30	0.00	0.24	**0.31**	0.17	0.22	**0.31**	NB
Average	**0.16**	**0.25**	**0.12**	**0.19**	**0.15**	**0.16**	**0.19**	**0.16**	**0.15**	**0.25**	

6 Conclusions

In this paper we have described our multiclass-based classification strategy in Information Extraction setting to classify rhetorical sentences taken from full scientific papers. We provide several strategies for solving multi-label classification problem and conduct experiments to evaluate their effectiveness on a standard corpus. The experiment results reveal that the multi-classifier approach, which delegates the classification task to a specialist classifier, can significantly improve the classification accuracy over ensemble and multi-label classifiers. When this specialist classifier is selected from the best classification method, it boosts the number of best performer in each category.

For future work, we will investigate if more complex classification strategies can further improve the current performance, and how these strategies can be applied to ontological-based, concept-driven, hierarchical structures of documents.

Acknowledgements

This research was partially supported by ITB Research and Innovation Grant under contract# 0277/I1.C07/PL/2012 Any opinions, findings, or conclusions are those of the authors, and do not necessarily reflect the views of the funding organization.

References

[1] Teufel, S., *Argumentative Zoning: Information Extraction from Scientific Text*, PhD Dissertation, University of Edinburgh, 1999.

[2] Li, Y., Gorman, S. & Elhadad, N., *Section Classification in Clinical Notes Using Supervised Hidden Markov Model*, In Proceedings of the 1st ACM International Health Informatics Symposium, ACM, New York, USA, pp. 744-750, 2010.

[3] Guo, Y., Korhonen, A., Liakata, M., Silins, B., Sun, L. & Stenius, U., *Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes*, In Proceeding of the Workshop on Biomedical Natural Language Processing, pp. 99-107, 2010.

[4] John, G.H. & Langley, P., *Estimating Continuous Distributions in Bayesian Classifiers*, In Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338-345, 1995.

[5] Teufel, S., Siddhartan, A. & Batchelor, C., *Towards Discipline-Independent Argumentative zoning Evidence from Chemistry and Computational Linguistics*, InProceeding of Conference on Empirical Methods in NLP, **3**, pp. 1493-1502, 2009.

[6] Liakata, M., Teufel, S., Siddharthan, A. & Batchelor, C., *Corpora for the Conceptualisation and Zoning of Scientific Papers*, In Proceedings of the 7th International Conference on Language Resources and Evaluation, pp. 2054-2061, 2010.

[7] Merity, S., Murphy, T. & Curran, J., *Accurate Argumentative Zoning with Maximum Entropy Models*, In Proceedings of the ACL Workshop on Text and Citation Analysis for Scholarly Digital Library, pp. 19-26, 2009.

[8] Hirohata, K., Okazaki, N., Ananiadou, S., Ishizuka, M. & Biocentre, M., *Identifying Sections in Scientific Abstracts using Conditional Random Fields*, In Proceedings of IJCNLP, pp. 381-388, 2008.

[9] Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbuhler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D. & Lovis, C., *Using Argumentative to Extract Key Sentences from Biomedical Abstracts*, International Journal of Medical Informatics, **76**(2-3), pp. 195-200, 2007.

[10] McKnight, L & Srinivasan, P., *Categorization of Sentence Types in Medical Abstracts*, In Proceedings of the AMIA Annual Symposium, pp. 440-444, 2003.

[11] Lin, J. Karakos, D., Demmer-Fushman, D. & Khudanpur, S., *Generative Content Models for Structural Analysis of Medical Abstracts*, In Proceedings of the HLT-NA ACL BioNLP Workshop, pp. 65-72, 2006.

[12] Dietterich, T.G., *Ensemble Methods in Machine Learning*, *Multiple Classifier Systems: Lecture Notes in Computer Science*, Springer Verlag, **1857/2000**, pp. 1-15, 2000.

[13] Ranawana, R., *Multi-Classifier Systems-Review and a Roadmap for Developers*, International Journal of Hybrid Intelligent Systems, **3**(1), pp. 35-61, 2006.

[14] Corney, D.P.A., Buxton, B.F., Langdon, W.B. & Jones, D.T., *BioRAT: Extracting Biological Information from Full-Length Papers*, Bioinformatics, **20**(17), pp. 3206–3213, 2004.

[15] Wyner, A., Mochales-Palau, R., Moens, M.F. & Milward, D., *Approaches to Text Mining Arguments from Legal Cases*, Lecture Notes in Computer Science, **6036** , pp 60-79, 2010.

[16] Cessie, S. & van Houwelingen, J.C., *Ridge Estimators in Logistic Regression*, Applied Statistics, **41**(1), 191-201, 1992.

[17] Witten, I.H. & Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Ed., Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, 2005.

[18] Aha, D. & Kibler, D., *Instance-Based Learning Algorithms*, Machine Learning, **6**, pp. 37-66, 1991.

[19] Frank, E. & Witten, I.H., *Generating Accurate Rule Sets Without Global Optimization*, In: Fifteenth International Conference on Machine Learning, pp. 144-151, 1998.

[20] Platt, J.C., *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Advances in Kernel Methods-Support Vector Learning, The MIT Press, 1999.