Move Classification in Scientific Abstracts using Linguistic Features

Tasanawan Soonklang

Department of Computing, Faculty of Science
Silpakorn University, Nakorn Pathom, Thailand
soonklang_t@su.ac.th

Abstract. Move structure is a framework for analyzing the rhetorical structure in research papers. This framework is very useful for assisting in the reading and writing of research article. In this paper, we present a computational method for sentence classification of move structure in the abstract of research articles. We propose two machine learning approaches: naive Bayes classifier and decision tree classifier. Both methods were trained on two groups of linguistic features: lexical features and grammatical features. These two approaches and linguistic features were evaluated with a small set of abstracts in the field of biomedical engineering using 10-fold cross validation. The experimental results indicate the benefit of lexical features and suggest that decision tree is a promising approach for move classification.

Keywords: Move Classification, Sentence Classification, Machine Learning, Abstract, Discourse Analysis, Research Article

1 Introduction

The study of Swales' [1] theoretical framework is used to draw up move categories. This framework is commonly applied for examining the rhetorical structure of research articles in various disciplines. Moreover, it help many non-native scientists to understand and construct a scientific paper by identifying the structure of writing. Later, Santos [2] proposed his five-move pattern and widely used for abstract analysis. A few study in computational linguistics focused on the task of automatic move tagging [3, 4]. Those study proposed the move tagging algorithms and exploited them in their pedagogical applications for computer-assisted acedemic writing. The best performance of the previous study is around 80%. Despite the success of the previous work, there are still opportunities for further enhancements.

In this paper, we propose move classification algorithms based on two machine learning approaches: naive Bayes (NB), and decision tree (DT). NB classifier is a simple statistical algorithm for sentence classification, performed as a baseline. One of the main attractions of using NB is that it was reported as a successful method when applied in the automatic identification of move structure in abstract [3]. DT is a symbolic method that classification decisions are easily interpretable as a rule set [5]. Additionally, DT is suitable for language processing tasks with symbolic textual data [6]. These classical methods can used for extracting knowledge about which features that are the most informative for sentence classification.

The important key for both machine learning method is the selection of appropriate features. In order to investigate this issue, we designed the features based on expert knowledge-guidance. The first group of features is which are "bag of word/phrases" extracting from a training examples and expert. The second group of features is grammatical features that experts used to identify move structure such as tense, voice, pronoun, modal, and preposition. The additional feature is a position feature, which is boost grammatical features by giving information about the location of the sentence in the abstract.

The paper is organized as follows: in Section 2 we briefly describe move structure and its related work. Section 3 briefly presents the move classification and our features. In Section 4 we report and discuss the experimental results. Section 5 concludes the paper.

2 Move Structure

Text structures in introduction of research article firstly analysed as a series of "moves" by Swales [1]. Since then the move structure has successfully utilized by many linguistic researchers in various disciplines such as engineering [7, 8], sociology [9], computer science [10], and biochemistry [11]. Subsequently, Santos [2] suggested that an abstracts in research article should be composed of five different moves, including background (B) introduces the current research background, purpose (P) presents its purpose or objectives, method (M) describes the methodology, result (R) states the results or summerizes the findings and discussion (D) draws conclusions or discussions.

This framework is broadly applied for abstract analysis [12–16]. During the last decade, a few digital learning tools exploited move structure for assisting novice or non-native speaker in academic reading and writing. The earliest work in computational analysis of move structure is the Mover [3], a machine learning tool for move classification using NB approach. Later, the CARE online learning system was proposed a HMM model for automatic analysing move structure in research abstract [4]. Recently, the Mover tool was applied for the sentence classification task in biomedical domain [17], focusing on the information about effect and patients. Their results were compared with other machine learning methods.

3 Move Classification

Two supervised classifiers were chosen from different learning techniques for sentence classication, including NB classifier from statistical method and DT classifier from symbolic methods. We divided the features into two categories:

grammatical features and lexical features, based on linguistic knowledge from acedemic writing experts.

3.1 Preprocessing

All sentences were preprocessed as follows. First, the sentences were segmented into list of words. Then all symbols and stopwords were removed from the list. After that, all words were POS-tagged, lemmatized and used to construct frequency corpus and feature sets.

3.2 Lexical Features

The frequency corpus consists of three collections of files, each containing five files for each move, built from training data. The first collection provides words and their frequencies, the second provides words with pos tags and their frequencies, and the last one provides bigram and their frequencies. The words or bigrams whose values appeared lower than 3 were not considered.

Table 1. Lexical features

Feature	Description
Word	word lexicon
n-grams	n-gram lexicon
phrase	phrase lexicon
collocation	collocation lexicon
freq-word	frequency of a word
freq- tag	frequency of a word and pos tag
freq- bigram	frequency of bigram

Additionally, there are four expert-created lexicons used for lexical features, including words, n-grams, phrase and collocation. All lexicons were created from a collection of words or phrases usually appeared in each move. Thus, each lexicon comprises of five files for each move.

The total 7 features includes three features from frequency corpus and four features from lexicons as shown in Table 1.

To extract features from frequency corpus, we summed the frequencies of all words that found in the corpus according to their moves and returned the move that have a maximum value. To extract features from each lexicon, we counted all words that appeared in the lexicon according to their moves, and returned the move that have a maximum value.

3.3 Grammatical Features

All grammatical features in this group are suggested by expert. The 13 features for identifying the structure of abstract are shown in Table 2.

Table 2. Grammatical features

Feature	Description
Tense	past, present, or future
Voice	active or passive
Pronoun	occurrence of 'we', 'our', 'this', or 'those'
Preposition	occurrence of preposition
Modal	occurrence of modal
To infinitive	occurrence of 'to' + verb
Whether	occurrence of 'whether'
By+ Gerund	occurrence of 'by' + verb
Article	occurrence of 'a' or 'a'
Determiner	occurrence of 'the'
Extraposition	occurrence of 'it' + verb to be + 'that'
Nominalization	occurrence of 'the' + noun + 'of'
Position	first, last, or ignored

The position feature is another feature adding to grammatical features. If the sentence is the first or second second sentence of an abstract, the value of position feature is first. If the sentence is one of the last two sentences, the value is last. The rest of the sentence is ignored.

4 Experimental Results and Discussion

Two categories of features were used for evaluation each algorithm, generating four variations for the experiment. The objective of testing is to determine which combination of variations performs the best out of all of the possible combinations.

4.1 Data Collections

The research abstracts were manually collected from top 5 journals with high impact factors in the field of biomedical engineering published in 2006. The selected journals are as follows: *IEEE Transactions on Medical Imaging, Journal of Biomedical Materials Research, IEEE Transactions on Biomedical Engineering, Artificial Organs, and IEEE Transactions on Neutral Systems and Rehabilitation Engineering.* Finally, a total of 60 research articles from the work of Kanoksilapatham [18] were carefully chosen to build the abstract corpus. The total of 528 sentences were labelled by the specialist. The occurrences of each move are listed in Table 3.

The accuracy of classification was evaluated using a 10-fold cross validation due to a small set of abstracts. In terms of 10-fold cross validation, all sentences in 60 abstracts were randomly divided into 10 folds. Each fold was removed in turn from the corpus and used as a testing data; the remaining 9 folds were used as a training data to build a learning model for classification.

Table 3. Distribuition of each move in abstract corpus

Move	Occurrence	Percentage
Background	97	18.37
Purpose	52	9.85
Methods	148	28.03
Result	161	30.49
Discussion	70	13.26
Total	528	100

The accuracies of classifiers were reported in terms of sentences correct reflecting the the number of sentences in which all move of the output exactly match those of the corresponding move in each article.

4.2 Results of All Moves

In this section, we presented the overall performance of sentence classification. The accuracies obtained from all models using 10-fold cross validation is given in Figure 1.

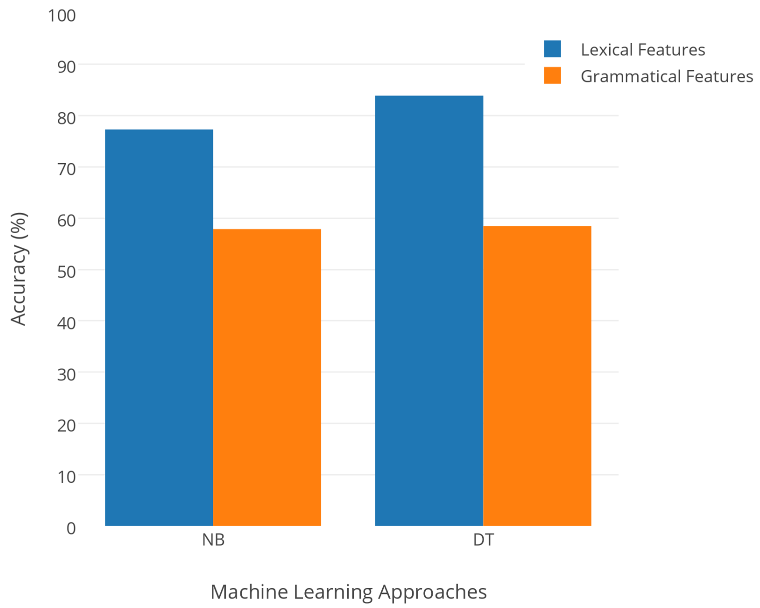

Fig. 1. The Overall Accuracy of Move Classification using All Four Models

The results show that in general the models with lexical features perform better than those of grammatical features. The best result was achieved with DT algorithm with lexical features, performing consistently across the different data sets with an average accuracy of 82.69%. The NB model with lexical features performed far higher than both models with grammatical features. The results of a DT model with grammatical features was given almost the same accuracy as a NB model with grammatical features, approximately 58%. These results also confirmed that DT algorithm performed slightly better than NB algorithm.

4.3 Results of Each Move

We further investigated the results of move classification algorithms by comparing the results of each move from four models. We presented the results of applying lexical and grammatical features with DT and NB models to identify each move in Figures 2.

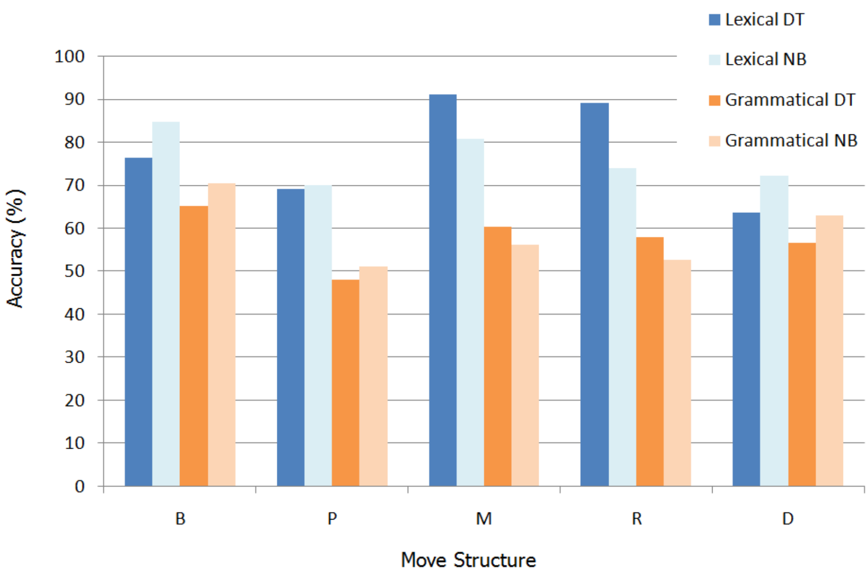

Fig. 2. Accuracies of all four Models on Different Move Structure

We achieved a high accuracy around 90% on move 'M' and move 'R' from DT method. These results are reasonable because the distribution of these two moves are much higher than the other moves. For move 'P', both models gave nearly the same results. Surprisingly, the accuracies of NB models were greater than those of DT models for the rest of move structure.

The accuracies of grammatical features in all moves conform to the performance characteristic of DT and NB models for move analysis. However, the overall accuracies are much lower than those of lexical features. As a result, we can conclude that the lexical features are more suitable for move analysis than grammatical features.

It is noteworthy that the two classification models with lexical features are dependent on the number sentences in training data, but this is not true for the models with grammatical features. Moverover, the accuracy of DT model with lexical features on each move varied with a wide range.

Table 4. Confusion matrix for all abstracts in the corpus using DT with Lexical Features

Actual vs. Predict	B	P	M	R	D
B	**86**	1	4	7	2
P	1	**43**	4	4	1
M	6	3	**134**	11	3
R	4	4	6	**139**	6
D	0	1	0	0	**58**
Accuracy	88.7%	82.7%	90.5%	86.3%	82.9%

Table 5. Confusion matrix for all abstracts in the corpus using NB with Lexical Features

Actual vs. Predict	B	P	M	R	D
B	**77**	9	15	15	10
P	6	**38**	3	10	1
M	8	4	**113**	12	3
R	5	1	7	**123**	8
D	1	0	10	1	**48**
Accuracy	79.4%	73.1%	76.4%	76.4%	68.6%

The induction rules from DT with lexical features suggested that bigram is an important key for identifying move. A knowledge extraction from DT with grammatical features suggested that position feature plays important role than the other features. Furthermore, DT can provide the characteristics of sentences in each move. For example, if the position is one of the last two sentences in an abstract and neither pronoun nor modal appeared, it should be classified as move 'R'. If the position is in the middle of an abstract and is active voice, containing any pronoun, it should be classified as move 'P'.

Table 4 and Table 5 showed the confusion matrix for classification tree and naive Bayes, tesing on all abstracts in the corpus using their best models with lexical features. The overall accuracies of DT and NB are 83.90% and 75.57%, respectively. Both confusion matrices showed the similar performance on move 'P' and 'D', which yields the two lowest accuracies.

By using the models obtained from the best results of testing set in 10-fold cross validation, the results in confusion matrix still corresponded to those of the average 10-fold-cross-validated accuracy. The classification accuracies obtained from both confusion matrix showed that the misclassification are as follows. Move 'B' often misclassiflied as all other moves, since background sentences possibly mention about the previous work in all aspects. Additionally, move 'M', 'R', and 'D' often misclassifield among themselves. In a preliminary error analysis, we found that move 'D' obtained the lowest result because some sentences composed of two clauses with two different moves.

5 Conclusion and Future Work

In this paper, we demonstrate the use of two feature sets, lexical and grammatical, and two supervised learning algorithms, Naive Bayes and Decision Trees, in the sentence classification of abstracts, where five move categories are identified. The sentence classification approaches are evaluated over a coprus of 60 biomedical engineering paper abstracts. We compare the performance of four models obtained from these two feature sets and two classifiers. The best results achieved from decision trees method using lexical features. The overall accuracy can reach 82.69%, which was higher than the previous work. Thus, our best model in this study shows potential use as a move analyzing tool for abstract in research article.

An error analysis of classification tree models will be explored thoroughly in the near future, since the accuracy of each move in 10-fold cross validation varied with a wide range from 63% to 91%. Moreover, the other machine learning method will be investigated for improving the performance. The evaluation will be conducted with a large number of abstracts from a variety of research areas.

Acknowledgments

I would like to address special thanks to Prof. Budsaba Kanoksilapatham for her invaluable help in manually collecting the research abstracts, classifying each move, and suggesting very useful features for machine learning techniques. I would also like to thank my lovely advisees, Pik, Nice and Jane, for serving as my research assistants.

References

1. Swales, J.: Aspects of Article Introductions. Aston ESP research reports. Language Studies Unit, University of Aston in Birmingham (1981)
2. Santos, M.D.: The textual organization of research paper abstracts in applied linguistics. Text **16** (1996) 481–499

3. Anthony, L., Lashkia, G.: Mover: a machine learning tool to assist in the reading and writing of technical papers. Professional Communication, IEEE Transactions on **46** (2003) 185–193
4. Wu, J.C., Chang, Y.C., Liou, H.C., Chang, J.S.: Computational analysis of move structures in academic abstracts. In: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia, Association for Computational Linguistics (2006) 41–44
5. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York, NY (1997)
6. Zimmermann, H.J., Tselentis, G., van Someren, M., Dounias, G.: Advances in Computational Intelligence and Learning: Methods and Applications. Springer Science & Business Media, New York, NY (2002)
7. Kanoksilapatham, B.: Structure of research article introductions in three engineering subdisciplines. IEEE Transactions on Professional Communication **55** (2012) 294–309
8. Sayako Maswana, T.K., Tajino, A.: Move analysis of research articles across five engineering fields: What they share and what they do not. Amersand **2** (1994) 1–11
9. Brett, P.: A genre analysis of the result section of sociology articles. English for Specific Purposes **13** (1994) 47–59
10. Posteguillo, S.: A genre-based approach to the teaching of reading and writing abstracts in computer science. In Pique, J., Andreu-Beso, J.V., Viera, D.J., eds.: English in Specific Settings. NAU Llibres Valencia, Spain (1996) 47–57
11. Kanoksilapatham, B.: Rhetorical structure of biochemistry research articles. English for Specific Purposes **24** (2005) 269–292
12. Graetz, N.: Teaching efl students to extract structural information from abstracts. In Ulign, J.M., Pugh, A.K., eds.: Reading for professional purposes: Methods and materials in teaching languages. Acco, Leuven (1985) 123–135
13. Bhatia, V.K.: Analysis genre: Language use in Professional Settings. Longman, London, UK (1993)
14. Cross, C., Oppenheim, C.: A genre analysis of scientific abstracts. Journal of Documentation **62** (2006) 428–446
15. Pho, P.D.: Research article abstracts in applied linguistics and educational technology: A study of linguistic realizations of rhetorical structure and authorial stance. Discourse Studies **10** (2008) 231–250
16. Abarghooeinezhad, M., Simin, S.: A structural move analysis of abstract in electronic engineering articles. International Journal of Research Studies in Language Learning **4** (2015) 69–80
17. Matos, P.F., Lombardi, L.O., Pardo, T.A.S., Ciferri, C.D.A., Vieira, M.T.P., Ciferri, R.R.: An environment for data analysis in biomedical domain: Information extraction for decision support systems. In Garca-Pedrajas, N., Herrera, F., Fyfe, C., Ali, J.M.B.M., eds.: Trends in Applied Intelligent Systems, volume 6096 of the series Lecture Notes in Computer Science. Springer Berlin Heidelberg (2010) 306–316
18. Kanoksilapatham, B.: Components of the discussion section in biomedical engineering research articles and their linguistic characterization. Journal of Science and Technology, Mahasarakham University **32** (2013) 92–97