Word Embedding for Rhetorical Sentence Categorization on Scientific Articles

Ghoziyah Haitan Rachman*, Masayu Leylia Khodra &
Dwi Hendratmo Widyantoro

School of Electrical Engineering and Informatics, Bandung Institute of Technology,
Jalan Ganesa No. 10, Bandung 40132, Indonesia
*E-mail: ghoziyahaitan@gmail.com

Abstract. A common task in summarizing scientific articles is employing the rhetorical structure of sentences. Determining rhetorical sentences itself passes through the process of text categorization. In order to get good performance, some works in text categorization have been done by employing word embedding. This paper presents rhetorical sentence categorization of scientific articles by using word embedding to capture semantically similar words. A comparison of employing Word2Vec and GloVe is shown. First, two experiments are evaluated using five classifiers, namely Naïve Bayes, Linear SVM, IBK, J48, and Maximum Entropy. Then, the best classifier from the first two experiments was employed. This research showed that Word2Vec CBOW performed better than Skip-Gram and GloVe. The best experimental result was from Word2Vec CBOW for 20,155 resource papers from ACL-ARC, features from Teufel and the previous label feature. In this experiment, Linear SVM produced the highest F-measure performance at 43.44%.

Keywords: *GloVe; rhetorical sentence categorization; scientific article; word embedding; Word2Vec.*

1 Introduction

Scientific studies are commonly reported in scientific articles to claim their novelty and contribution. Academics and researchers need these documents to collect relevant information and compare them with each other for their own research [1]. The abstract is the first section of a scientific article they read to find a summary of the information it contains [2]. It mostly contains a brief version of the study's objectives, methods, results and conclusion [3]. Nevertheless, readers cannot get all important information as needed only from the abstract because it does not reveal the correlation with other scientific articles. Due to this condition, readers prefer a summary of a collection of scientific articles in the form of an outline of certain points. These points can contain segments of text (i.e. sentences) with a rhetorical structure that contains a meaningful category in the body of each section [4]. In addition, classified rhetorical sentences are easier to structure into a summary specified by reader

Received January 7th, 2018, Revision August 3rd, 2018, Accepted for publication September 14th, 2018.
ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2018.12.2.5

needs [5]. Classifying rhetorical sentences passes through the process of text categorization. This process commonly produces a high-dimensional feature space, which represents the text in the document by employing bag of words [6]. Its vector representation can be formed into a distributional semantic model by capturing the meaning of each existing word [7].

Some works in rhetorical sentence categorization for scientific articles are argumentative zoning by classifying 7 rhetorical categories with Naïve Bayes [8], Maximum Entropy [9], Word2Vec [10], and 16 rhetorical categories adopted from [11] with a heterogeneous multi-classifier [3]. These works mostly adopted rhetorical sentence categorization from [8], which employs the meta-discourse feature. This is considered the most dependable indicator in determining rhetorical categories [8]. This feature is split into three parts, which are formulaic (*formu*), agentivity (*ag-1*), and action (*ag-2*). Each has several patterns that contain a list of defined words. For example, in the phrase 'we hope to improve our result', the word 'hope' is the reference for the meta-discourse *action* of the type *effect*. Since this phrase has a pattern with a word that is already in the vocabulary of special action effect, this sentence will receive the value 1. However, word references in this meta-discourse are not always applicable for all scientific articles. There are some important words in the corpus of scientific articles [3] that are not included in the meta-discourse vocabulary from [8]. Some of these words are 'align', 'annotate', 'argument', 'aspect', 'concept', 'context', 'data', 'direct', 'document', 'domain', 'represent', and so on. If the occurrence of words that are actually important is not covered by any feature, this may reduce the F-measure performance of rhetorical sentence categorization on scientific articles. This problem is related to out-of-vocabulary (OOV) words, i.e. words that do not exist in the vocabulary. The present research found that there are 118 important words (see Appendix) that do not exist in the meta-discourse vocabulary from [8]. The OOV percentage is about 12.45% of the total 948 words in the vocabulary.

To address this problem, the word representation feature was employed, which handles the semantics of words. Several works have shown that a semantic model can produce higher performance than a lexicon [12,13] because it captures the weight of similarities between words in a document. Therefore, this research conducted rhetorical sentence categorization by employing word embedding to detect the semantic meaning of words in scientific articles [7,14].

The rest of this paper is organized as follows. The following section provides an overview of related works. Section 3 describes the rhetorical categories and features used in this paper. The setup and results of our experiments are discussed in Sections 4 and 5, followed by the conclusion in Section 6.

2 Related Works

Some techniques of semantic word representation are Word2Vec and GloVe. Word2Vec, proposed by Mikolov [15], trains a neural network to predict the *n*-th word of a given set of words by computing a vector representation of the words and calculating the similarity between words using the cosine distance between their vectors. It provides two architectures, namely Continuous Bag-of-Words (CBOW) and Skip-Gram. CBOW predicts single words based on the context of the words, while Skip-Gram predicts the context of words based on a single given word [15]. In contrast to CBOW, the purpose of learning in the Skip-Gram model is to maximize the probability of $(w1, w2, ..., wc \mid wp)$, where $w1...c$ is the context of the words and wp is a given word. Besides that, GloVe first builds a co-occurrence matrix for the entire corpus and then factorizes it to yield the word and context vectors [16]. It considers the probability of a single given word occurring in the context of another given word.

Heffernan & Teufel [17] employed Word2Vec representation to identify problem statements in scientific texts. They used 18,753,472 sentences from a biomedical corpus consisting of all full-text Pubmed articles and then built a model from 200 words that are semantically similar to ‘problem’. The result showed that Word2Vec leads to a significantly performance increase because Word2Vec attributes had the greatest information gain compared with the other features. Putra & Khodra [12] showed that text representation using a semantic model has higher accuracy than using a lexicon model, which does not consider the semantic meaning of words. It reached the best accuracy by using ANN with Word2Vec CBOW at 82.94%. Naili, *et al.* [14] conducted a comparative study between LSA, Word2Vec, and GloVe for topic segmentation. They concluded that Word2Vec and GloVe performs better than LSA.

For rhetorical sentence categorization, Liu [10] employed Word2Vec for seven rhetorical categories from Teufel [8] (‘aim’, ‘textual’, ‘own’, ‘background’, ‘contrast’, ‘basis’, and ‘other’). In addition, Widyantoro, *et al.* [3] implemented 16 rhetorical categories and features adopted from [8] and [11] in combination with a heterogeneous multi-classifier. Its average F-measure result was about 25%. Actually, they employed a different corpus than Teufel [8]. This could affect the final rhetorical categorization model, because Teufel [8] employed the meta-discourse feature, in which the existing grammar depends on word patterns that always appear in sentences from the corpus.

3 Rhetorical Sentence Categorization

Rhetorical sentence categorization is the task of assigning a particular rhetorical status to every sentence in a document. In this research, we used the dataset of scientific articles from [5], which contains 75 papers from the ACL Anthology, and we added 50 new scientific articles. Every sentence in these papers was assigned to one of 16 rhetorical categories adopted from [3] and [11], as explained in Table 1.

Table 1 Description of 16 rhetorical categories [3,8,11].

Category	Description	Example
AIM	Specific objectives or hypotheses in current research	The aim of this paper is to examine the role that training plays in the tagging process . . .
NOV_ADV	Novelty or advantage of current approach	An important advantage of combining morphological analysis and error detection/ correction is . . .
CO_GRO	No insignificant knowledge claims for paper	It has often been stated that discourse is an inherently collaborative process . . .
OTHR	Significant knowledge claims by other research, neutral	But in Moortgat's mixed system all the different resource management modes of the different systems are left intact . . .
PREV_OWN	Significant knowledge claims by the author on the previous paper, neutral	Earlier work of the author (Feldweg 1993; Feldweg 1999a) within the framework of a project on corpus . . .
OWN_MTHD	New claims, methods in current research	In order for it to be useful for our purposes, the following extensions must be made: . . .
OWN_FAIL	Failure of solutions / methods / experiments in current research	When the ABL algorithms try to learn with two completely distinct sentences, nothing can be learned.
OWN_RES	Measurable results of current research	All the curves have a generally upward trend but always lie far below backoff (51% error rate).
OWN_CONC	Findings, the unmeasurable conclusions of current research	It appears that in fact the major problems do not lie in the area of grammar size, but in input length.
CODI	Comparison, contrast, differences with other solutions (neutral)	Unlike most research in pragmatics that focuses on certain types of presuppositions, we provide a global framework . . .
GAP_WEAK	Disadvantages / problems of previous solutions	This simple model leads to serious overestimates of system error rates.
ANTISUPP	Different with other research results or theories; current has better result	This result challenges the claims of recent discourse theories (Grosz and Sidner 1986, Reichman 1985) which argue . . .
SUPPORT	Other research that support current research	Work similar to that described here has been carried out by Merialdo (1994), with broadly similar conclusions.
USE	Another methods / framework used in the current research	We use the framework for the allocation and transfer of control of Whittaker and Stenton (1988).
FUT	Further research	An important area for future research is to develop principled methods . . .
TEXTUAL	Reference structure of text	Table 1 shows the main part of the pattern matchers.

Then, we used sentence features, i.e. *content*, *absolute location*, *explicit structure*, *sentence length*, *citation*, *formulaic*, *agentivity* and *sequential label*. These features, except *sequential label*, were adopted from [3] and [18]. A description of these categories is given in Table 2.

Table 2 Extraction features adopted from Teufel [3,18].

Type	Name	Description	Values
Content	Cont-1	Occurrence of 10 significant terms of document using TF-IDF	1, 0
	Cont-2a	Incidence of words occurring in document title	1, 0
	Cont-3	Occurrence of 10 significant terms of abstract using TF-IDF	1, 0
Absolute location	Loc	Sentence position in document to 10 segments	A-J
Explicit structure	Struct-1	Sentence position within section	7 values
	Struct-2	Sentence position within paragraph	Initial, Medial, Final
	Struct-3	Prototypical type of section title	15 section or Non-Prototypical
Length	Length	Sentence has longer than 15 words or not	1, 0
Syntax	Syn	Occurrence of 1st finite verb and auxiliary modal	1, 0
	Adj	Occurrence of adjective	1, 0
Citations	Cit-1	Occurrence of citation or self-citation	Citation, Self-Citation, None
	Cit-2	Citation location in sentence	Beginning, Middle, End, or None
Formulaic	Formu	Occurrence of formulaic expression	1, 0
Agentivity	Ag-1	Occurrence of agent type	1, 0
	Ag-2	Occurrence of action type	1, 0
	Negation	Occurrence negation in sentence	1, 0
Sequential Label	PrevLabel	Previous label	Previous category or 'start' for first sentence in document

4 Experimental Setup

The methodology used in this research can be divided into three main processes. The first is pre-processing, followed by constructing the model of word vector representations by employing Word2Vec. The second process is extracting all features. The last process is testing to know the F-measure details of rhetorical category categorization by using different classifiers, namely Naïve Bayes, Linear SVM, IBK, J48, and Maximum Entropy. Pre-processing is conducted to clear the dataset for training and testing; the processes involved are case folding, stemming, and stop word removal. For building the Word2Vec model, each sentence in the collection of scientific articles was constructed into a word representation with a string-to-word vector. Then, feature extraction adopted from [3] and [18] was applied. For the experiments on sequence labeling, the previous label feature was added. The value of this feature is the previous rhetorical category, or 'start' for the first sentence in the document. The process of training and testing the classification model used four classifiers, namely Naïve Bayes [19], Linear SVM [20], IBk [21], J48 (C4.5) [22], and Maximum Entropy. These classifiers were only employed in the first experiment in order to find the best classifier for this rhetorical sentence categorization. After that, only the best classifier was employed in the next experiment. The tools we used for this research were Weka [23], LibSVM, SMILE library, and DeepLearning4J. The general process is shown in Figure 1.

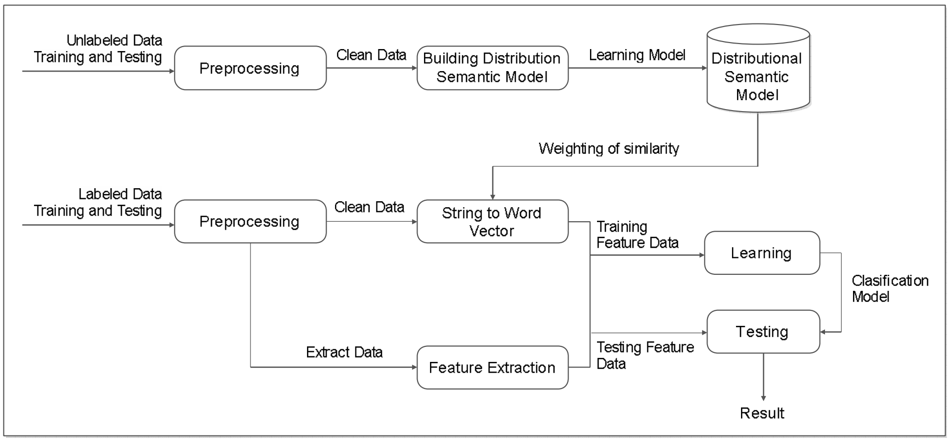


Figure 1 General process of research methodology.

The semantic model was gradually built using 34,741 scientific articles from the ACL Anthology. Its model employs Word2Vec and GloVe with various parameters using DeepLearning4J. For the experiment on comparing semantic word representation, we also used a Word2Vec model from Google News and a GloVe model from Stanford.

The data used in this experiment are of two types, namely data for building the semantic model and annotated data for rhetorical sentence categorization. For categorization, this research used 75 papers [5] and 50 new scientific articles from the ACL Anthology. The total number of sentences was 16,046. The training data consisted of 100 papers and the testing data consisted of 25 papers. The ratio of the two was 4:1. The number of testing data was 3452 while the number of training data was 12,594. The number of sentences in every category can be seen in Table 3.

Table 3 Number of sentences in 16 rhetorical categories.

No	Rhetorical Category	Data			No	Rhetorical Category	Data		
		All	Training	Testing			All	Training	Testing
1	Aim	415	350	65	9	Own_conc	1091	941	150
2	Antisupp	125	109	16	10	Own_fail	142	109	33
3	Codi	244	201	43	11	Own_mthd	6429	4844	1585
4	Co_gro	706	602	104	12	Own_res	585	389	196
5	Fut	205	156	49	13	Prev_own	699	383	316
6	Gap_weak	560	445	115	14	Support	745	634	111
7	Nov_adv	342	278	64	15	Textual	794	594	200
8	Othr	1854	1605	249	16	Use	1110	954	156

The experiment created a classification model of the feature extraction results from the labeled sentences and then tested the model with the specified test data. The classification employed Naïve Bayes, Linear SVM, IBk, J48, and Maximum Entropy for the baseline experiment. The tools used in this classification were Weka and SMILE. Scenarios of experiments to be performed, namely:

1. *Baseline experiment.* In this experiment, the rhetorical sentence classification model was built from the extraction of adaptation features [5] as shown in Table 2 without the previous label feature. It employed all classifiers, namely Naïve Bayes, Linear SVM, IBk, J48, and Maximum Entropy.
2. *Baseline + sequence labeling experiment*. In this experiment, the training and testing of the rhetorical sentence classification model was built from the extraction of adaptation features [5] as shown in Table 2, added with sequence labeling. The added feature was previous label (previous category). It employed all classifiers, namely Naïve Bayes, Linear SVM, IBk, J48, and Maximum Entropy.
3. *Word embedding experiment*. In this experiment, the rhetorical sentence classification model was built from word semantic representation. It employed Word2Vec and GloVe. The weight of each sentence against a word is the result of the average calculation of each word in a sentence that weighs the resemblance in the Word2Vec and the GloVe model. This scenario employed only the best classifier from the baseline experiment. For the experiment on comparing semantic word representation, it also used the pre-trained Word2Vec model from Google News and the pre-trained GloVe model from Stanford. Then the number of papers was split to gradually build semantic model as follows: 5,000; 10,000; 15,000; 20,155; 25,000; 30,000; and 34,741.
4. *Baseline + word embedding experiment*. In this experiment, the rhetorical sentence classification model was built from the extraction of adaptation features [5] as shown in Table 2, added with the semantic models from

Word2Vec and GloVe. In this scenario, we also tried to add similar words to 10 significant TF-IDF terms in the content feature. This scenario only employed the best classifier from the baseline experiment.

5. *Previous label + baseline + word embedding experiment.* In this experiment, the rhetorical sentence classification model was built from the extraction of adaptation features [5] as shown in Table 1, added with the previous label feature and the semantic models from Word2Vec and GloVe. This scenario only employed the best classifier from the baseline experiment.

5 Results and Analysis

First, we used 100 scientific articles for training and 25 for testing. We extracted all features regarding the scenario we planned. The baseline experiment was evaluated by five classifiers, namely Naïve Bayes, Linear SVM, IBK, J48, and Maximum Entropy. The result of the first experiment is shown in Table 4.

Table 4 F-measure performance of experiments 1 and 2.

Category	Scenario 1					Scenario 2				
	NB	SVM Lin	IBk	J48	Max. Entr	NB	SVM Lin	IBk	J48	Max. Entr
aim	46.00	50.50	27.10	48.30	0.00	52.00	55.20	32.50	37.80	5.33
nov_adv	15.50	0.00	5.90	7.90	0.00	21.10	3.00	9.20	11.00	0.00
co_gro	30.90	33.60	19.50	25.50	9.22	45.30	49.30	29.90	46.80	16.48
othr	13.40	6.50	14.50	14.90	3.62	44.70	48.00	30.60	47.00	37.43
prev_own	16.00	22.00	7.10	16.80	0.00	61.70	74.50	35.40	76.60	23.10
own_mthd	68.10	68.30	61.30	68.40	64.30	74.60	76.00	70.40	76.90	69.20
own_res	10.90	0.00	7.90	10.00	0.00	24.10	33.90	10.70	27.40	3.85
own_conc	25.30	25.90	15.00	24.10	14.10	34.30	35.40	19.30	24.30	25.60
own_fail	4.50	0.00	0.00	0.00	0.00	12.20	0.00	0.00	15.20	0.00
codi	0.00	0.00	0.00	3.80	0.00	9.70	19.20	3.50	6.20	0.00
gap_weak	23.40	21.20	7.50	12.40	6.06	32.40	33.80	9.50	21.00	27.55
antisupp	0.00	0.00	7.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00
support	21.40	17.80	10.50	9.00	4.44	33.80	37.20	12.90	34.90	34.73
use	16.80	16.10	7.40	14.10	0.00	22.30	50.10	20.80	34.40	28.57
fut	32.30	35.30	11.60	37.40	11.43	40.40	43.90	30.10	42.20	43.30
textual	17.90	1.00	13.10	8.70	0.00	29.10	5.70	22.50	25.10	6.19
Average	**21.40**	**18.64**	**13.51**	**18.83**	**7.07**	**33.61**	**35.33**	**21.08**	**32.93**	**20.08**

Table 5 Average F-measure score of experiments 1 and 2.

Experiment	Classifier				
	NB	SVM Lin	IBk	J48	Max. Entr
Scenario 1	**21.40**	18.64	13.51	18.83	7.07
Scenario 2	33.61	**35.33**	21.08	32.93	20.08

Table 5 shows that in the baseline experiment of this study (Scenario 1), the best performing classification method was Naïve Bayes with an F-measure result of 21.40%. The second position was obtained by J48 tree with a difference of 2.57%. The classification with maximum entropy had the F-measure worst result, which was 7.07%. If we look in more detail at each class in Table 4, there are some categories that have very low F-measure scores, even up to 0%. For example, for Naïve Bayes, the categories 'codi' and 'antisupp' are simply not recognized. This can be because both belong to a minority class. The cause for the low F-measure of the 'codi' category is that more instances of this category were classified as 'own_mthd'. This also applies to other classification methods.

Table 4 indicates that the use of sequence labeling can improve the rhetorical classification performance of the baseline experiment and the results will exceed the F-measure in [3], which was only about 25%. The results of this experiment showed that the highest F-measure score was achieved by employing Linear SVM at 35.33%. Again, the Maximum Entropy method had the worst result, as in the baseline experiment, with 20.08%. The difference between the F-measure of Scenario 1 and Scenario 2 is large. It can be seen from Table 5 that for Linear SVM, the increase of the F-measure results from the previous experiment reached 16.69% after using the previous label feature. A significant increase of F-measure performance also occurred with other classification methods. However, in this experiment, Linear SVM was the best performing method, while in previous experiments naïve Bayes performed best.

Unfortunately, in Table 4 for the second scenario, the F-measure results for the 'antisupp' category for all classification methods were 0%, which means that no method recognized this category. This could be because there was too much variance in the previous label for this category, so no previous category could characterize this category. Also, it is a minority class. In contrast, the category 'start', the most previous label, appeared in 'aim' and 'co_gro'. This means that the use of 'start' can be more helpful in determining 'aim' and 'co_gro'. This is proven by Table 4, which shows that the F-measure of 'aim' and 'co_gro' in Scenario 2 increased significantly after using the previous label feature. Besides that, some sections in scientific articles can have more bias towards previous label patterns between sentences. Therefore, the section pattern and the previous label pattern can be interdependent and help to classify rhetorical categories of sentences.

Table 6 shows that the previous label pattern in sentences from scientific articles is more likely to occur in several sections (*Section*). The rhetorical sequence 'Aim – Own_mthd' occurs most in the *Abstract* section with a total of 70, then in the *Introduction* section with a total of 45. Meanwhile, the rhetorical pattern ‘Co_gro – Gap_weak’ occurs most in the *Introduction* section with a total of 42. By looking at the distribution of the unbalanced rhetorical category patterns in Table 6, there are indications that the rhetorical patterns have a role in determining the targets of the previous rhetorical category. For example, in the *Introduction* section in this example, the ‘aim’ category in the rhetorical pattern *Sequence* has more influence on determining the next ‘own_mthd’ category. Finally, these rhetorical patterns are related to each other.

Table 6 Distribution of previous label pattern examples in training dataset.

Section	**Examples of Previous Label Pattern in Training dataset**		
	Aim – Own_mthd	**Co_gro – Gap_weak**	**Prev_own – Own_mthd**
Abstract	70	10	2
Conclusion	20	1	3
Data	4	0	1
Discussion	2	1	3
Evaluation	0	0	2
Experiment	6	1	4
Introduction	45	42	13
Method	5	2	14
Non-prototypical	11	4	31
Related Work	5	3	8
Result	1	0	2
Solution	0	1	1

For the experiment using Scenario 3 we used the pre-trained Word2Vec model from Google News and the pre-trained GloVe model from Stanford. We also used 34,741 scientific articles from the ACL Anthology to build a semantic word representation model. We tried to modify the parameters of the architecture (CBOW/Skip-Gram), negative sampling (yes or no), hierarchical softmax (yes or no), and dimension (300 or 500). Then we gradually split the number of papers to build the semantic model, as follows: 5,000; 10,000; 15,000; 20,155; 25,000; 30,000; and 34,741 to know what the effect is of increasing the number of resources for building the semantic model.

Tables 7 and 8 show the results of the experiment employing different parameters and a larger number of resources (scientific articles) to build the Word2Vec model from pre-trained Google News. It indicates that the added resources did not have a significant impact on F-measure performance. This could be because the 5,000 to 34,741 papers have many words and patterns that are not too different from each other. It can even be stated that 20,155 papers are sufficient to represent 34,741 papers because the F-measure score of rhetorical categorization with the Word2Vec model of the 20,155 papers is almost the same as that of the 34,741 papers (better even). Besides that, it also shows that the best performing parameters in most cases occurred when implementing CBOW with negative sampling. This research used a value of 10 for negative sampling. The Word2Vec model from pre-trained Google News was still not better than the resources of scientific articles because it is related to a different domain from our research so the words from Google News cannot cover all words that exist in the scientific articles.

Table 7 F-measure score for scenario 3 (Word2Vec) for first eight categories.

Source	Arch.	Neg. Sam.	Hier. Soft.	Dim.	othr	own_ mthd	nov_ adv	gap_ weak	aim	textual	support	use
Scientific articles												
20,155	CBOW	Yes		300	2.70	67.00	0.00	8.80	30.90	65.00	7.50	12.50
20,155	CBOW		Yes	300	3.40	66.90	0.00	11.70	36.70	66.80	5.90	11.00
20,155	CBOW	Yes		500	6.80	66.80	0.00	16.90	37.60	64.70	11.80	13.50
20,155	CBOW		Yes	500	2.10	66.40	0.00	12.70	35.40	63.80	8.80	13.00
20,155	Skip-Gram	Yes		500	8.70	66.70	0.00	8.50	30.50	62.10	7.40	10.30
20,155	Skip-Gram		Yes	500	3.90	66.60	0.00	12.00	28.80	64.60	7.10	12.90
34,741	CBOW	Yes		300	4.80	66.80	0.00	10.50	31.90	65.80	10.10	12.90
34,741	CBOW		Yes	300	2.70	66.80	0.00	16.90	28.00	64.30	8.80	11.80
34,741	CBOW	Yes		500	5.60	67.20	0.00	13.70	35.30	65.50	8.60	14.80
34,741	CBOW		Yes	500	4.90	66.40	0.00	14.30	32.80	65.30	7.80	13.30
5,000	CBOW	Yes		500	1.40	66.10	0.00	8.40	33.30	63.60	6.20	13.20
10,000	CBOW	Yes		500	4.20	66.30	0.00	7.50	33.10	63.10	13.00	11.90
15,000	CBOW	Yes		500	2.70	66.10	0.00	8.50	33.30	66.20	9.10	10.20
25,000	CBOW	Yes		500	4.00	66.90	0.00	14.90	37.20	63.10	10.40	13.30
30,000	CBOW	Yes		500	3.40	66.50	0.00	14.90	33.30	64.70	4.60	11.60
Pre-trained Google News												
		Yes		300	0.00	66.00	0.00	12.80	35.30	59.80	3.30	12.30

Table 8 F-measure Score for scenario 3 (Word2Vec) for second eight categories.

Source	Arch.	Neg. Sam.	Hier. Soft.	Dim.	fut	own_ conc	co_ gro	codi	own_ res	own _fail	prev_ own	antisupp	Ave.
Scientific articles													
20,155	CBOW	Yes		300	16.70	30.30	20.40	0.00	22.60	0.00	5.30	0.00	18.11
20,155	CBOW		Yes	300	10.20	26.90	22.40	0.00	22.70	0.00	1.90	0.00	17.91
20,155	CBOW	Yes		500	13.60	27.70	25.50	0.00	24.20	0.00	7.50	0.00	19.79
20,155	CBOW		Yes	500	18.50	26.90	24.10	0.00	28.50	0.00	4.80	0.00	19.06
20,155	Skip-Gram	Yes		500	13.30	28.40	16.90	0.00	21.10	0.00	**10.10**	0.00	17.75
20,155	Skip-Gram		Yes	500	10.20	29.60	18.10	0.00	20.60	0.00	**13.90**	0.00	18.02
34,741	CBOW	Yes		300	19.40	32.20	26.40	0.00	19.50	0.00	5.30	0.00	19.10
34,741	CBOW		Yes	300	19.00	27.80	24.00	0.00	21.90	0.00	5.80	0.00	18.61
34,741	CBOW	Yes		500	25.40	29.00	24.50	0.00	22.50	0.00	3.60	0.00	19.73
34,741	CBOW		Yes	500	22.20	25.70	18.20	0.00	25.90	0.00	4.80	0.00	18.85
5,000	CBOW	Yes		500	18.80	27.80	20.40	0.00	19.20	0.00	7.60	0.00	17.88
10,000	CBOW	Yes		500	18.80	28.40	22.90	0.00	21.60	0.00	6.50	0.00	18.58
15,000	CBOW	Yes		500	16.40	27.80	23.30	0.00	20.60	0.00	4.80	0.00	18.06
25,000	CBOW	Yes		500	20.00	29.00	23.60	0.00	22.40	0.00	9.70	0.00	19.66
30,000	CBOW	Yes		500	16.10	25.90	29.10	0.00	19.50	0.00	6.60	0.00	18.51
Pre-trained Google News													
		Yes		300	10.90	27.50	12.90	0.00	19.20	0.00	0.60	0.00	16.29

The Word2Vec architecture of CBOW gives a better F-measure performance than Skip-Gram when there are many words that occur frequently [14]. In scientific articles there are many words that are repeated in several particular patterns. For example, the synonyms of the word ‘aim’ from the Word2Vec model are ‘propose’, ‘present’, ‘purpose’, and so on. These words often occur in scientific articles, which do not contain many variant words. But for the category 'prev_own', the F-measure score by employing Skip-Gram was higher than by employing CBOW. This indicates that many sentences in this category contain infrequent words. As explained in Table 2 (description of 16 rhetorical categories), sentences with the 'prev_own' category contain significant knowledge claims by the author in a previous research. This category does not have words that always occur because its sentences are structured by the author's perspective to describe what he or she has done before. Thus, there is large variance in the words contained in the sentences. In contrast to Skip-Gram, CBOW predicts single words based on the context of words [15]. This characteristic of CBOW makes that infrequent words are rarely selected as predicted due to their low probability.

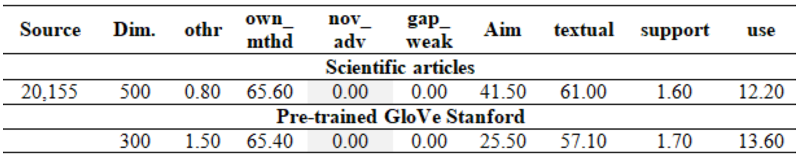
Besides that, the F-measure of some categories is still 0.00%, i.e. 'nov_adv', 'codi', 'own_fail', and 'antisupp'. This can be because these categories have a small amount of sentences and contain many infrequent words. Actually, the category 'fut' has a small amount of sentences too, but they contain words that appear very frequently. Some examples from this category are:

1. "They have more complex surface forms and should be investigated further." (5th testing paper)
2. "This remains for future research." (7th testing paper)
3. "In future work, we will analyze the difference of the expression of the titles composed with and without using the wizard, and investigate what sort expression is effective to lay readers." (10th testing paper)

From three sentences of category 'fut' above, it can be seen that there are words that commonly appear in future sentences. This makes that the F-measure of 'fut' does not become 0.00%, although its F-measure is still not high.

Tables 9 and 10 show the result of using GloVe for building a semantic model based on the scientific articles and the pre-trained data from Stanford. Unfortunately, both results were not better than by employing Word2Vec. Actually, GloVe depends on how many iterations are set for training the data. The higher the number of iterations, the higher its F-measure performance will be. Because the result of employing GloVe is not better than employing Word2Vec, so we only employed Word2Vec CBOW in the next experiment.

Table 9 F-measure score of scenario 3 (GloVe) for the first 8 categories.



Source	Dim.	othr	own_ mthd	nov_ adv	gap_ weak	Aim	textual	support	use
				Scientific articles					
20,155	500	0.80	65.60	0.00	0.00	41.50	61.00	1.60	12.20
				Pre-trained GloVe Stanford					
	300	1.50	65.40	0.00	0.00	25.50	57.10	1.70	13.60

Table 10 F-measure score of scenario 3 (GloVe) for the second 8 categories.

Source	Dim.	fut	own_ conc	co_ gro	codi	own_ res	own _fail	prev_o wn	antisupp	Ave.
Scientific articles										
20,155	500	7.30	17.80	1.90	0.00	17.70	0.00	0.60	0.00	14.25
Pre-trained GloVe Stanford										
	300	7.40	27.40	8.30	0.00	12.80	0.00	0.60	0.00	13.83

The results of Scenario 4 are shown in Tables 11 and 12. For this scenario we only used the architecture of CBOW from Word2Vec, added with features from Teufel [18]. In this experiment, we tried to modify the content feature by adding similar words to 10 significant words from Cont1, Cont2a, and Cont3. For using 20,155 instances and CBOW, it was shown that the result of employing the base of content feature was still better than adding similar words. The F-measure score was 28.01%. But this condition was different when using 34,741 instances. Although its increase was very small, using 10 similar words for the content feature did affect the F-measure performance of rhetorical categorization; the F-measure score went from 27.79 to 27.94%. This could be because similar words become more accurate when using a CBOW model from 34,741 papers. From Table 9 and 10, the best result was still when employing the base of content feature and 20,155 papers for the CBOW model.

Table 11 F-measure score of scenario 4 for the first 8 categories.

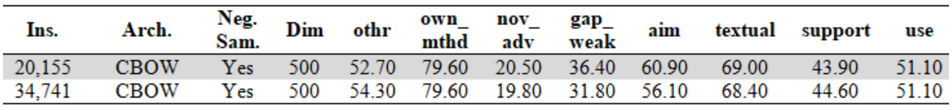
Ins.	Arch.	Neg. Sam.	Content Feature	Dim.	othr	own_ mthd	nov_ adv	gap_ weak	aim	textual	support	use
20,155	CBOW	Yes	Base	500	15.30	72.10	10.40	29.30	62.10	66.70	22.60	19.20
20,155	CBOW	Yes	10 similar	500	15.50	72.00	10.10	29.10	59.00	68.40	22.90	18.80
20,155	CBOW	Yes	5 similar	500	14.70	72.20	10.10	28.00	58.80	68.10	22.10	18.80
34,741	CBOW	Yes	Base	500	15.40	71.60	7.90	29.90	60.20	66.30	22.50	21.10
34,741	CBOW	Yes	10 similar	500	15.40	71.40	7.80	27.40	61.50	66.30	24.50	21.10

Table 12 F-measure score of scenario 4 for second 8 categories.

Ins.	Arch.	Neg. Sam.	Content Feature	Dim	fut	own_ conc	co_ gro	codi	own_ res	own _fail	prev_ own	antisupp	Ave.
20,155	CBOW	Yes	Base	500	38.00	34.10	37.70	0.00	18.70	0.00	22.00	0.00	28.01
20,155	CBOW	Yes	10 similar	500	35.00	32.20	36.50	0.00	12.20	0.00	12.70	0.00	26.53
20,155	CBOW	Yes	5 similar	500	37.50	32.80	36.50	0.00	19.60	0.00	21.50	0.00	27.54
34,741	CBOW	Yes	Base	500	38.50	32.50	39.60	0.00	17.20	0.00	22.00	0.00	27.79
34,741	CBOW	Yes	10 similar	500	40.00	32.30	38.30	0.00	19.10	0.00	22.00	0.00	27.94

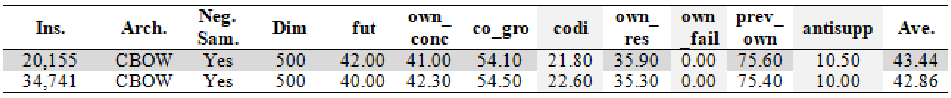
The final experiment is shown in Tables 13 and 14. From all the conducted experiments, the best combination was when employing Word2Vec CBOW, the features from Teufel [18], and the previous label feature. Unfortunately, the F-measure score of the category 'own_fail' was still 0.00%. This could be because the amount of 'own_fail' sentences was not insufficient to build a usual context of words existing in scientific articles. Thus CBOW, which is more suitable for frequent words, cannot handle this issue well.

Table 13 F-measure score for scenario 5 for the first 8 categories.



Ins.	Arch.	Neg. Sam.	Dim	othr	own_ mthd	nov_ adv	gap_ weak	aim	textual	support	use
20,155	CBOW	Yes	500	52.70	79.60	20.50	36.40	60.90	69.00	43.90	51.10
34,741	CBOW	Yes	500	54.30	79.60	19.80	31.80	56.10	68.40	44.60	51.10

Table 14 F-measure score of scenario 5 for the second 8 categories.



Ins.	Arch.	Neg. Sam.	Dim	fut	own_ conc	co_gro	codi	own_ res	own _fail	prev_ own	antisupp	Ave.
20,155	CBOW	Yes	500	42.00	41.00	54.10	21.80	35.90	0.00	75.60	10.50	43.44
34,741	CBOW	Yes	500	40.00	42.30	54.50	22.60	35.30	0.00	75.40	10.00	42.86

From the two tables above, using 20,155 scientific articles for the CBOW model still improved the F-measure performance compared to using 34,741 articles. It can be concluded that if the amount of resources is quite high, for example 20,000 documents, and there are frequently occurring words and grammars in these documents, the addition of this amount will not have a significant effect. Thus is because the existing words have low variance so 20,000 documents is enough to build the semantic model. Thus, in the final experiment, the best F-measure score so far was 43.44% by employing 20,155 papers for the CBOW model and the rest of the features.

6 Conclusion

In our research, it was found that the F-measure performance of rhetorical categorization on scientific articles could be improved by using sequence labeling (previous label pattern) and semantic word representation by Word2Vec. First, our research showed that the rhetorical pattern of the previous label feature in sentences of scientific articles is more likely to appear in several sections (*Section*). The distribution of the unbalanced rhetorical category patterns in sections of scientific articles indicates that rhetorical patterns have a role in determining the targets of the previous rhetorical category. For example, in the *Introduction* section, the 'aim' category in the rhetorical pattern sequence has more influence on determining the next 'own_mthd' category. These rhetorical patterns are related to each other and interdependent on the sections of scientific articles. Adding the previous label feature can improve the F-measure performance of rhetorical sentence categorization. Comparing the four classification methods used, Linear SVM reached the highest F-measure score at 35.33%.

Secondly, our research showed that Word2Vec CBOW performed better than Skip-Gram and GloVe. This is because CBOW is suitable to catch frequent words while Skip-Gram is suitable to catch infrequent words, as stated in [14]. Sentences in scientific articles contain words and grammars that frequently occur. For example, synonyms of the word ‘aim’ from the Word2Vec model are ‘propose’, ‘present’, ‘purpose’, and so on. But for the categories ‘prev_own’ and ‘othr’, these sentences do not contain any words that frequently occur because they are structured according to the author’s perspective to describe what he or she has done previously. Thus, there is more variance in words. The condition of frequent words for the rest of the rhetorical categories makes the CBOW model perform better for most rhetorical categories. The results of all experiments showed that the highest F-measure was obtained when employing Word2Vec CBOW with 20,155 resource papers, the features from Teufel [8] and [18], and the previous label feature, at 43.44%. Finally, Linear SVM achieved the highest F-measure performance, indicating that this classifier is suitable for high-dimensional feature spaces. In addition, this result was also better than that of the rhetorical sentence categorization in [3], with a significant improvement from about 25% to 43.44%. After comparing the prediction from [3] and our research, the probability of random values with the same result was 0.00. This is lower than alpha (0.05), so that the difference between the actual mean values is probably significant; this was calculated with RapidMiner.

After analyzing the result of rhetorical classification, we found that most of the ‘own_fail’ sentences were categorized as ‘own_mthd’ (15 sentences) and ‘own_conc’ (8 sentences). This needs to be further investigated. Another suggestion is to define what kind of pattern for ‘own_fail’ is different from ‘own_conc’, instead of adding another dataset of scientific articles. In addition, we can decrease the amount of ‘own_mthd’ to avoid overfitting.

Acknowledgement

This research was partially supported by the Master’s Program toward a Doctoral Degree for Excellent Graduates (Program Pendidikan Magister Menuju Doktor untuk Sarjana Unggul/PMDSU) from Kemenristekdikti Indonesia within the research project Automatic Guided Summarization.

References

[1] Schwegler, R.A. & Shamoon, L.K., *The Aims and Process of the Research Paper*, College English, **44**(8), pp. 817-824, 1982.

[2] Luhn, H.P., *The Automatic Creation of Literature Abstracts*, IBM Journal, **2**(2), pp. 159-165, 1958.

[3] Widyantoro, D.H., Khodra, M.L., Trilaksono, B.R. & Aziz, E.A., *A Multiclass-based Classification Strategy Sentence Categorization from*

Scientific Papers, Journal of ICT Research and Applications, **7**(3), pp. 235-249, 2013.

[4] Taboada, M. & Mann, W.C., *Rhetorical Structure Theory: Looking Back and Moving Ahead*, Discourse studies, **8**(3), pp. 423-459, 2006.

[5] Khodra, M.L., Widyantoro, D.H., Aziz, E.A. & Trilaksono, B.R., *Automatic Tailored Multi-Paper Summarization Based on Rhetorical Document Profile and Summary Specification*, ITB Journal of Information and Communication Technology, **6**(3), pp. 220-239, 2012.

[6] Yang, Y. & Pedersen, J.O., *A Comparative Study on Feature Selection in Text Categorization*, Proceedings of the 14th International Conference on Machine Learning, pp. 412-420, 1997.

[7] Rong, X., *word2vec Parameter Learning Explained*, Cornell University Library, https://arxiv.org/abs/1411.2738, (5 June 2016).

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

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

[10] Liu, H., *Automatic Argumentative-Zoning Using Word2vec*, Cornell University Library, https://arxiv.org/abs/1703.10152, (29 March 2017).

[11] Teufel, S., Siddharthan, A. & Batchelor, C., *Towards Discipline-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics*, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1493-1502, 2009.

[12] Putra, Y.A. & Khodra, M.L., *Deep Learning and Distributional Semantic Model for Indonesian Tweet Categorization*, Proceedings of the 2016 International Conference on Data and Software Engineering (ICoDSE), pp. 1-6, 2016.

[13] Rahmawati, D. & Khodra, M.L., *Word2vec Semantic Representation in Multilabel Classification for Indonesian News Article*, Proceedings of the 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), pp. 1-6, 2016.

[14] Naili, M., Chaibi, A.H. & Ghezala, H.H.B., *Comparative Study of Word Embedding Methods in Topic Segmentation*, Procedia Computer Science, **112**, pp. 340-349, 2017.

[15] Mikolov, T., Chen, K., Corrado, G. & Dean, J., *Efficient Estimation of Word Representations in Vector Space*, Cornell University Library, https://arxiv.org/abs/1301.3781, (7 September 2013).

[16] Pennington, J., Socher, R. & Manning, C., *Glove: Global Vectors for Word Representation*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, 2014.

[17] Heffernan, K. & Teufel, S., *Identifying Problem Statements in Scientific Text*, Workshop on Foundations of the Language of Argumentation (in conjunction with COMMA), 2016.

[18] Teufel, S. & Moens, M., *Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status*, Computational Linguistics, **28**(4), pp. 409-445, 2002.

[19] Rish, I., *An Empirical Study of the Naive Bayes Classifier*, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, pp. 41-46, 2001.

[20] Chang, C.C. & Lin, C.J., *LIBSVM: A Library for Support Vector Machines*, ACM Transactions on Intelligent Systems and Technology (TIST), **2**(3), Article No.27, 2011.

[21] Jiang, L., Cai, Z., Wang, D. & Jiang, S., *Survey of Improving K-Nearest-Neighbor for Classification*, Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 679-683, 2007.

[22] Bhargava, N., Sharma, G., Bhargava, R. & Mathuria, M., *Decision Tree Analysis on J48 Algorithm for Data Mining*, International Journal of Advanced Research in Computer Science and Software Engineering, **3**(6), pp. 1114-1119, 2013.

[23] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H., *The WEKA Data Mining Software: An Update*, ACM SIGKDD Explorations Newsletter, **11**(1), pp. 10-18, 2009.

Appendix

List of out of vocabulary words for Teufel's meta-discourse [8].

align
annotate
argument
aspect
Chinese
concept
consist
context
data
definite
depend
differ
direct
distribute
document
domain
embed

English
entity
event
noun
object
order
output
pair
parse
part
pattern
phrase
probable
produce
query
recognition
refer

relate
represent
rule
score
segment
sentence
word
action
agent
candidate
cluster
dictionary
element
frame
group
Japanese
lexicon

list
node
operate
parameter
string
value
column
detail
example
express
extract
feature
function
generate
grammar
include
inform

input
knowledge
label
language
learn
level
lexicon
make
map
name
natural
sequence
set
sign
source
syntactic
speaker

speech
tag
term
text
token
train
translate
tree
type
user
utter
vector
verb
dataset
length
neural

network
weight
plan
addition
effect
found
high
lower
rate
significant
human
interest
recent
understand
average
reduce