Rhetorical Sentence Categorization for Scientific Paper Using Word2Vec Semantic Representation

G H Rachman[1], M L Khodra[2], and D H Widyantoro[3]

[1, 2, 3]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia

Email: [1]ghoziyahaitan@gmail.com, [2]masayu@stei.itb.ac.id, [3]dwi@stei.itb.ac.id

Abstract. One of some ways to summarize scientific papers is by employing rhetorical structure of sentences. Determining rhetorical sentence itself passes through the process of text categorization. In order to get good performance, some works in text categorization have been done by employing semantic similarity words. Therefore, this paper aims to present the rhetorical sentence categorization from scientific paper by using selected features, added previous label, and Word2Vec to capture semantic similarity words. Then, this paper shows the result of employing resampling for balancing the existing instances per class and combining resampling and Word2Vec representation itself. Every experiment is tested in two classifiers, namely IBk and J48 tree. It shows that the use of previous label, Word2Vec (Skip-Gram), and resampling improves performance. After doing all the experiments in the 10-fold cross-validation, the highest performance of F-measure is achieved 84.97% by combining Word2Vec (Skip-Gram), all features, and resampling.

Keywords. Rhetorical categorization; scientific paper; word2vec; feature extraction.

1. Introduction

Scientific papers become one kind of document that continues to increase in number along with more research conducted by academics and researchers themselves. They initially use, at least one, scientific paper to collect relevant information for their own argument in the research [1]. Abstracts of that papers can be read in advance to find out the summarized meaningful information [2]. However, abstracts does not always give all important information as needed by readers and does not expose the correlation of scientific papers at each other. Due to this condition, readers prefer summary of a collection of scientific papers in the form of outline on certain points. A scientific paper has rhetorical structure, in which every segment of a text (i.e. sentence) has a meaningful category in the body of each section [3]. The classified rhetorical sentences are easier to be structured into a summary as needed by readers [4].

Determining rhetorical sentence passes through the process of sentence categorization. This process produces the high dimensionality of feature space which represents text in the document by employing string to word vector [5]. Its vector representation can be formed into distributional semantic model by capturing the meaning of each existing words [6]. For example, because the word of "aim" and "purpose" has the similar meaning, these two words should also have high word similarity. Therefore, the more words with similar meaning on the scientific paper exist, the higher effect against performance of classification can be acquired. In this paper, we will conduct rhetorical sentence categorization on scientific paper by employing Word2Vec as a tool for computing word vector representation of semantic similarity words [6]. Since rhetorical dataset is an imbalanced dataset, this paper will also involve handling of imbalanced data set.

The rest of this paper is organized as follows. The next section provides the related works on rhetorical sentence categorization from scientific paper and implementation of word2vec semantic representation. Our method is explained in section 3. We also define various rhetorical categories and feature sets for classifying sentences. We describe our experiments in section 4. The results are analyzed in section 5. Finally, the conclusion and further work are described in section 6.

2. Related Works

Text categorization is conducted to assign label of a text based on its text features. Rhetorical sentence categorization is commonly used for scientific papers. Teufel and Moens [7] conducted paper summarization by first assign one of seven rhetorical categories to every sentence. These seven are "aim", "textual", "own", "background", "contrast", "basis", and "other". They [8] employed Naïve Bayes classifier and got the F-measure result from 26% to 86%. Then, its result has risen up until 92.93% after employing Maximum Entropy [9]. These seven categories next become 15 rhetorical categories, namely "aim", "nov_adv", "co_gro", "othr", "prev_own", "own_mthd", "own_fail", "own_res", "own_conc", "codi", "gap_weak", "antisupp", "support", "use", and "fut". Widyantoro, et al. [10] implemented these 15 categories added with "textual" and employed multi-heterogeneous classifier. Its average of F-measure result is about 25%. It still does not significantly show the high performance for this 15 rhetorical sentence categorization model. The main problem is imbalanced dataset and word variation in sentences.

Word2Vec is used to group words which have similar meaning into vector representation. This model is proposed by Mikolov [6] and released by Google in 2013. There are two architectures of Word2Vec, namely the continuous bag-of-words (CBOW) and skip-gram. The CBOW predicts the current word based on the context and the skip-gram predicts surrounding words given the current word [11].

Heffernan and Teufel [12] employed Word2Vec representation to identifying problem statement in scientific text. They used 18,753,472 sentences from a biomedical corpus based on all full-text Pubmed articles and then built model from the 200 semantically similar words to only "problem". Its result showed that Word2Vec model caused significant performance increase, because Word2Vec attributes have the greatest information gain compared the other features.

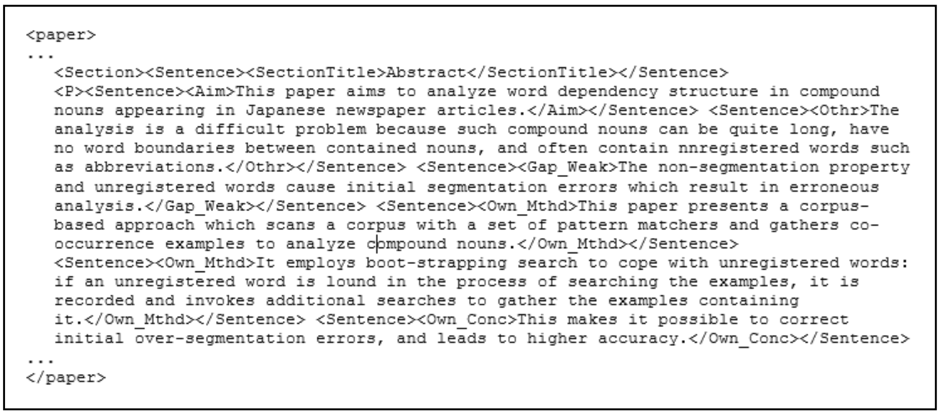
Rahmawati and Khodra [13] have employed Word2Vec representation in multi-label classification for Indonesian news article. They did experiment by using CBOW (once) and Skip-Gram in the vector length from 200 to 500. Its result showed the testing F-measure value from 76.73% to 81.63%. It is proven that Word2Vec has better performance than TF-IDF in their previous research. Putra and Khodra [14] also showed that text representation using semantic model has higher accuracy than lexicon model which does not consider about semantic meaning of words. They obtained the best accuracy by using ANN with word2vec CBOW at 82.94%.

3. Methodology

In this research, we will use the data set of scientific paper from [10]. Every sentence in these papers had been annotated into one of 16 rhetorical categories from Teufel, namely "aim", "nov_adv", "co_gro", "othr", "prev_own", "own_mthd", "own_fail", "own_res", "own_conc", "codi", "gap_weak", "antisupp", "support", "use", "fut", and "textual". The description of these categories is explained in [10]. The example of an annotated paper is shown in Figure 1.

We use sentence features i.e. content, absolute location, explicit structure, sentence length, citation and sequential label. These features, except previous label, are adapted from [7] and [10]. Its explanation can be seen in Table 1. Then this paper employs Word2Vec representation algorithm from Medallia[1] library. We use 75 scientific papers still from [10] to build the word vector itself. Actually that data set was taken from ACL Anthology Reference Corpus (ACL-ARC).

[1] https://github.com/medallia/Word2VecJava



```
<paper>
...
   <Section><Sentence><SectionTitle>Abstract</SectionTitle></Sentence>
   <P><Sentence><Aim>This paper aims to analyze word dependency structure in compound
   nouns appearing in Japanese newspaper articles.</Aim></Sentence> <Sentence><Othr>The
   analysis is a difficult problem because such compound nouns can be quite long, have
   no word boundaries between contained nouns, and often contain nnregistered words such
   as abbreviations.</Othr></Sentence> <Sentence><Gap_Weak>The non-segmentation property
   and unregistered words cause initial segmentation errors which result in erroneous
   analysis.</Gap_Weak></Sentence> <Sentence><Own_Mthd>This paper presents a corpus-
   based approach which scans a corpus with a set of pattern matchers and gathers co-
   occurrence examples to analyze compound nouns.</Own_Mthd></Sentence>
   <Sentence><Own_Mthd>It employs boot-strapping search to cope with unregistered words:
   if an unregistered word is lound in the process of searching the examples, it is
   recorded and invokes additional searches to gather the examples containing
   it.</Own_Mthd></Sentence> <Sentence><Own_Conc>This makes it possible to correct
   initial over-segmentation errors, and leads to higher accuracy.</Own_Conc></Sentence>
...
</paper>
```

Figure 1. Example of annotated paper

Table 1. Features and its description

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
Sentence length	Length	Whether sentence has longer than 15 words or not	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
Sequential Label	PrevLabel	Previous label	Previous rhetoric category or “start” for first sentence in the document

In addition, resampling would be included in the experiment process for solving the imbalanced data [15]. Resampling is conducted by using WEKA. It will do oversampling for categories in smaller number and under-sampling for categories in higher number. So that every instances of categories is balanced in number.

Actually, our method is generally divided into four main processes. The first is preprocessing, and then constructing the vector representations of the vocabularies of data set by using Word2Vec. The second is producing all the features. The third is implementing resampling before building classification model. The last is we do 10-fold cross validation to know the F-measure detail of rhetorical category by using some classifiers, namely IBk and J48 in Weka.

We consider four preprocessing that consist of case folding, tokenization, stemming, and stop word elimination. These are done using Apache Lucene library and Weka. All of them were fixtures that would always be applied in all experiments. Then we use our annotated preprocessed sentences to output a model containing word vectors through Word2Vec representation (CBOW and Skip-Gram). We set its layer size into the vector length of 200. In the end, all classification models are produced from the last process.

4. Experiment

We conducted the rhetorical sentence categorization using the data set of scientific paper from [10]. This data has 75 annotated scientific papers that have been split into 10880 sentences. The number of sentences in every category before and after doing resampling can be seen in Table 2 and 3.

Table 2. The number of sentences in first eight rhetorical categories

Resampling	Aim	Nov_adv	Co_gro	Othr	Prev_own	Own_mthd	Own_fail	Own_res
Before	214	248	386	967	625	5325	70	419
After	663	631	689	724	691	673	718	678

Table 3. The number of sentences in second eight rhetorical categories

Resampling	Own_conc	Codi	Gap_weak	Antisupp	Support	Use	Fut	Textual
Before	581	111	367	60	392	426	154	535
After	673	632	692	709	649	734	667	657

In addition, there are five experiments that would be conducted. It is related to existence of previous label as sequential feature, resampling, and Word2Vec of CBOW and Skip-Gram. These are as follows:

1. *Baseline*. The types of feature to use are content, absolute location, explicit structure, sentence length and citation, without sequential label (previous label);
2. *Scenario 1*. In this experiment, all the features are used, including sequential label (previous label), to know whether previous label increases performance or not;
3. *Scenario 2*. In this experiment, scenario 1 will be conducted and completed by doing resampling. This method is used because [16] showed that resampling is the best method for solving imbalanced data set;
4. *Scenario 3*. This experiment will employ all the features and involve Word2Vec representation by using CBOW and Skip-Gram algorithm. For this scenario, resampling is not used, because this scenario is only to know the effect of CBOW and Skip-Gram against performance; and
5. *Scenario 4*. After knowing the best one of Word2Vec algorithm from comparing the result of scenario 3, this experiment will combine all the features, the best one of Word2Vec (Skip-Gram), and resampling.

For the last scenario, only Skip-Gram is used, because this kind of Word2Vec has better F-measure result than CBOW if compared with the result of first scenario. In addition, every scenario would be evaluated in two classifiers, namely IBk and J48 tree. Actually, determining classifiers for this research is based on [16] showing that IBk and J48 are the first two classifiers that have the highest performance in its last experiment. Then from all the experiments, we would know how some features have effect on the accuracy of F-measure in every rhetorical categorization.

5. Result

First, we built the word vectors from the data set which consist of 75 scientific papers that have been split into 10880 sentences. We extracted all the features without previous label. This was evaluated by two classifiers, namely IBk and J48 tree.

Table 4 shows the result of experiments. We can see that previous label and method of resampling has significantly increased the performance. In the first experiment (baseline), the F-measure average from using IBk classifier achieved 17.39% and J48 achieved 18.79%. Then these two performance have increased becoming 30.07% for IBk and 39.12% for J48 tree in the scenario 1 that involves previous label. Its increase is until about 18% from initial performance in baseline. It is proven that previous label has a large effect to raise the performance of classification as concluded in [16]. It is because previous label as sequential feature shows some patterns which always appear in rhetorical sentence categorization that are correlated with the target classes [17].

In fact, the data set we got has the imbalanced instances per category, so we have to make it balanced. Then, the third experiment (scenario 2) exists to know how much the effect of handling imbalanced data set, by using resampling in Weka, against the performance. Its result shows that the F-measure average from using IBk classifier achieved 79.91% and J48 achieved 72.1%. These two increases are significant from the result of scenario 1. However, there is a difference in this scenario

Table 4. The result of experiment

Category	Baseline		Scenario 1 (Baseline + Previous Label)		Scenario 2 (Resample)		Scenario 3 (CBOW or Skip-Gram)				Scenario 4 (Skip-Gram + Resample)	
	IBk	J48	IBk	J48	IBk	J48	IBk (CBOW)	J48 (CBOW)	IBk (Skip-Gram)	J48 (Skip-Gram)	IBk	J48
Othr	24.7	24.2	47	57.6	69	60.1	39.6	48.9	38.6	44.2	72.8	64.6
Own_mthd	69.3	71.7	75.9	79.6	47.1	42.3	71	77.3	71.7	76.1	53	49.9
Nov_adv	6	6.4	11.4	18.7	81.4	73.8	12.5	16.4	16.9	13.5	90.3	84.5
Gap_weak	12.3	7	23.5	28.5	79.7	63.7	21.9	19.6	23	20.4	83.8	75.7
Aim	18.7	18	28.3	34.6	85.2	77.5	28.2	27.5	34.7	23.9	92	85.4
Textual	14.3	16.7	22.3	26	68	55.5	19.1	32.3	28.3	29.4	79.5	72.1
Support	19	28.4	37	48	85	76	33.6	42.4	30.2	40.9	85.4	80.7
Use	6.9	6.9	27.3	45	81.5	73.7	26.2	37.1	28.5	35.2	84.9	80.6
Fut	29.9	36.4	36.7	42.2	89.3	85.5	28.2	32.6	30.6	31.3	93.9	91.4
Own_conc	18.8	16.9	30.4	40.6	73.3	60.5	27.9	34.2	36.8	29.2	80.8	71.6
Co_gro	24.3	25	42.6	52.5	80.2	71.8	39.6	45.1	38.9	40.8	87.5	79.3
Codi	2.9	9.9	11.8	18.3	89.7	86.6	7.1	14	8.4	9.9	95.7	92.5
Own_res	14.7	6.7	23.8	39.5	79.2	65.3	26.2	30.4	35.3	29.2	84.7	78.7
Own_fail	0	2.6	4.3	23.9	93.3	92.4	6.2	14.1	8.4	16.1	96.1	95.1
Prev_own	16.4	26.7	56.3	70.9	81.9	76.4	48.1	65.7	47.8	61.7	80.8	82.1
Antisupp	0	0	2.6	0	94.7	92.5	5.9	2.7	14.6	0	98.4	96.5
Average	**17.39**	**18.97**	**30.07**	**39.12**	**79.91**	**72.1**	**27.58**	**33.77**	**30.79**	**31.36**	**84.97**	**80.04**

result compared with the two before. In the first two experiment, J48 always has the higher performance, but in the third, IBk becomes the highest. This result is the same as concluded in [16] that resampling is more suitable for IBk classifier than J48 tree.

The next is we do the last two scenarios that related to Word2vec representation. In the scenario 3, we compare between using CBOW and Skip-Gram algorithm without doing resampling. Its result will be compared with the result of scenario 1, because both of them use all the extracted features. The most result from this scenario is not good. When using CBOW algorithm, the F-measure average from using IBk classifier achieved only 27.58% and J48 achieved 33.77%. If compared with the result of scenario 1, this result has decreased from 2.5% to 5.5%. It is different when using Skip-Gram. Although in J48 tree classifier, the F-measure has decreased becoming 31.36% more than when using CBOW, in IBk classifier it shows the contrary that using Skip-Gram is better than CBOW. Moreover the only one of all result in scenario 3 that increases from scenario 1 is acquired when using Skip-Gram and IBk classifier. In [13], Skip-Gram is indeed better than CBOW. It could be because Skip-Gram predicts surrounding words given the current word. In this research we check the similarity between words existing in sentences and words existing in vocabulary in Word2Vec representation. So that every word in sentence has weight of the semantic similarity between vocabulary clearly. It is different when using CBOW. It will predict the current word based on the context, in which sentence can be that context.

Therefore in the scenario 4, we only involve Skip-Gram algorithm then it is combined with resampling. This scenario is compared with scenario 2. Its result shows that the F-measure average from using IBk classifier achieved only 84.97% and J48 achieved 80.04%. After using resampling, the performance has always increased. It means that combination of Skip-Gram of Word2Vec representation and resampling give the higher performance for rhetorical sentence categorization. Resampling helps to solve the imbalanced data set and Skip-Gram Word2Vec to capture semantic similarity words in the data set.

6. Conclusion and Further Work

For the first, the implementation of previous label as sequential feature is used to catch class pattern which always appears that is correlated to the target class. So that it could be used to predict the next rhetorical category for sentences. For the second, the implementation of resample can get the performance up. It is because this method make the imbalanced data set in every class become balanced. For this research, we get the highest performance of using resampling when employing IBk classifier. It means that resampling is more suitable with IBk classifier than with J48 tree. Furthermore our work focused also in employing Word2Vec semantic representation. There are two architecture of Word2Vec, namely Skip-Gram and CBOW. Actually, after doing this research, we got that the performance of using Skip-Gram is better than CBOW. The last is implementation of combining Word2Vec and resampling can give the higher performance.

Actually, in this work we still do Word2Vec representation in the annotated data set itself and have not yet conducted testing and training. These two things can affect the performance whether it becomes better or not. In the further work, it will be considered to achieve good performance of rhetorical sentence categorization. Moreover the data set should be added more also to get higher performance.

Acknowledgement

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

References

[1] R A Schwegler and L K Shamoon 1982 *The Aims and Process of the Research Paper* (College English vol 44) no 8 pp 817-824

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

[3] M Taboada and W C Mann 2006 *Rhetorical structure theory: Looking back and moving ahead* (Discourse studies vol 8) no 3 pp 423-459

[4] M L Khodra, D H Widyantoro, E A Aziz, and B R Trilaksono 2012 *Automatic Tailored Multi-Paper Summarization based on Rhetorical Document Profile and Summary Specification* (Journal of ICT Research and Applications vol 6) no 3 pp 220-239

[5] Y Yang and J Pedersen 1997 *A Comparative Study on Feature Selection in Text Categorization, Int. Conf. on Machine Learning* (Nashville: ICML) pp 412-420

[6] X Rong 2014 word2vec Parameter Learning Explained (Cornell University Library) *Preprint* cs/1411.2738

[7] S Teufel 1999 *Argumentative Zoning: Information Extraction from Scientific Text* (Edinburgh: University of Edinburgh)

[8] S Teufel and M Moens 2002 *Summarizing scientific articles: experiments with relevance and rhetorical status* (Computational Linguistics - Summarization vol 28) issue 4 pp 409-445

[9] S Merity, T Murphy, and J Curran 2009 *Accurate Argumentative Zoning with Maximum Entropy models, Proc. of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries* (Suntec: World Scientific Publishing Co Pte Ltd) pp 19-26

[10] D H Widyantoro, M L Khodra, B R Trilaksono and E A Aziz 2013 *A Multiclass-based Classification Strategy Sentence Categorization from Scientific Papers* (Journal of ICT Research and Applications vol 7) no 3 pp 235-249

[11] T Mikolov, K Chen, G Corrado, and J Dean 2013 Efficient estimation of word representations in vector space (Cornell University Library) *Preprint* cs/1301.3781

[12] K Heffernan and S Teufel 2016 *Identifying Problem Statements in Scientific Text, Foundations of the Language of Argumentation, Proc. of COMMA 2016 Workshop* (University of Potsdam) p 18

[13] D Rahmawati and M L Khodra 2016 *Word2vec Semantic Representation in Multilabel Classification for Indonesian News Article, Proc. of 2016 Int. Conf. on Advanced Informatics: Concepts, Theory and Application (ICAICTA)* (Penang)

[14] Y A Putra and M L Khodra 2016 *Deep Learning and Distributional Semantic Model for Indonesian Tweet Categorization, Proc. of 2016 International Conference on Data and Software Engineering (ICoDSE)* (Bali)

[15] N V Chawla, N Japkowicz and A Kotcz 2004 *Editorial: special issue on learning from imbalanced data sets* (ACM Sigkdd Explorations Newsletter vol 6) no 1 pp 1-6

[16] G H Rachman and M L Khodra 2016 *Automatic Rhetorical Sentence Categorization on Indonesian Meeting Minutes, Proc. of 2016 International Conference on Data and Software Engineering (ICoDSE)* (Bali)

[17] N Lesh, M J Zaki and M Oglhara 2000 *Scalable feature mining for sequential data* (IEEE Intelligent Systems and their Applications vol 15) no 2 pp 48-56