Automatic sentence extraction for the detection of scientific paper relations

Y Sibaroni1, S S Prasetiyowati2, M Miftachudin3
1,2,3Computational Science, School of Computing, Universitas Telkom

1yuliant@telkomuniversity.ac.id, 2srisuryani@telkomuniversity.ac.id,
3Muhammad.Miftachudin55@gmail.com

Abstract. The relations between scientific papers are very useful for researchers to see the interconnection between scientific papers quickly. By observing the inter-article relationships, researchers can identify, among others, the weaknesses of existing research, performance improvements achieved to date, and tools or data typically used in research in specific fields. So far, methods that have been developed to detect paper relations include machine learning and rule-based methods. However, a problem still arises in the process of sentence extraction from scientific paper documents, which is still done manually. This manual process causes the detection of scientific paper relations longer and inefficient. To overcome this problem, this study performs an automatic sentences extraction while the paper relations are identified based on the citation sentence. The performance of the built system is then compared with that of the manual extraction system. The analysis results suggested that the automatic sentence extraction indicates a very high level of performance in the detection of paper relations, which is close to that of manual sentence extraction.

1. Introduction
Relations between scientific papers are forms of an interrelationship between one paper and another. Such relations are very useful and important in the fields of research and education. By observing the relationships between the existing scientific articles, one can obtain a description of recent research in particular fields as well as those which can be developed further. In addition, scientific paper relations can be used to investigate the development trend in future research. Manually, the forms of such paper relations can be obtained through an in-depth analysis of a collection of scientific papers.

Large number of worldwide produced scientific papers which reach millions of papers annually will certainly complicate the process of manually identified relations [1]. Therefore, the need for the process of identifying relations between papers automatically becomes more urgent.

In the case of citation-based relations, the actual applications actually have been created. Some websites which serve as a paper recommender system have been available, such as scholar.google.com, www.acm.org, www.sciencedirect.com, and citeseerx.ist.psu.edu. Websites may provide recommendations of desired scientific papers on the basis of similarity relations or citation relations. Similarity relations in question mean the level of content similarity between scientific papers.

Developing an automatic paper recommendation system takes several stages. In a supervised learning-based stage, the process includes file reading and sentence splitting based on a collection of
pdf scientific papers, creation of a scientific paper dataset and class annotation, feature identification and extraction, model learning, and, finally, model evaluation [2]. All of these steps (except class annotation) should be done automatically to be effective for large data applications.

One of the problems that arise in those studies is the process of automatic file reading and sentence splitting in which some sentences cannot be separated completely. An improper sentence separation can certainly affect the identification of the whole paper relations. So far, no studies have been done on research paper relations which measure the effect of automatic extraction error on the performance of scientific paper relation identification.

This study was aimed to develop an automatic sentence extraction process for detecting scientific paper relations and, at the same time, to measure the effectiveness of the automatic sentence extraction compared to the manual sentence extraction. Therefore, this study seeks to obtain a system which is able to automatically separate sentences in scientific papers whose results are expected to approach the results of manual sentence extraction. The result of this study is very important to know when we want to develop business applications of papers’ relation detection based on input pdf-type papers that come from the internet.

2. Related work
The relations between scientific papers are formed when the papers have similar contents or when a paper cites other papers. The first is known as a content-based approach while the latter is known as a citation analysis-based approach [3]. This citation-based approach is further developed into a citation context approach. A citation context is defined as a collection of words or phrases in a sentence which quotes or discusses other papers [2]. This citation context-based approach provides a more detailed relation form of the citation analysis approach which only contains similarity relations. Some examples of paper relations which can be developed based on this approach include extend, criticize, compare, and improve [4].

The research focus in each approach is different. In the content-based approach, the research focus of the majority of researchers is on developing better similarity measurements. Several similarity measurements developed include identity measures [5], Jaccard coefficient [6] and cosine similarities [7]. Identity measures are used to determine the co-derivative (plagiarism) of an article based on a collection of article database. Jaccard coefficient is used to measure the similarity of text in an article, while cosine similarities are used to measure the power level of co-citation between two sentences in an article.

In the citation analysis approach, the research focus is actually also on developing similarity measurements like the previous approach. However, in this approach, the similarity measurements are calculated according to the occurring citations. Some basic similarity measurements include bibliographic coupling, co-citation analysis, reference list, and cited by [8], [9]. In this approach, the similarity of a paper is not only defined by its content, but also by its citation pattern. The citation pattern in question includes citation sequence, citation location in a paragraph or chapter, and so on. Some of the more complex citation-based similarity measurements developed by researchers include in-text citation frequency analysis (ICFA) and in-text citation distance analysis (ICDA) by Gipp et al. [8], Citation Proximity Analysis (CPA) and Citation Order Analysis (COA) by Gipp and Bell [9] and Global Relation Strength (GRS) by Liang et.al [10].

In the citation context approach, the research focus is on detecting paper relations based on a citation context contained in citation sentences. Teufel et al. classify citation context into 12 types [11] in which these 12 categories are sentences which contain inter-paper relationships. Teufel et al. use a supervised machine learning as the classification method. The development of an explicit paper relation scheme based on this citation context is then performed by Wang. Based on the results of citation context analysis, Wang divides paper relations into extend, criticize, compare, and improve [4]. The method used by Wang to classify paper relations is the rule-based method. In both studies, the
main feature used for classifying or determining paper relations is the cue phrase feature based on the citation sentence identified previously.

The existence of paper relations, therefore, depends on the results of citation sentence identification. In relation to the process of citation sentence identification, some studies are found to be very relevant. Two of them are those conducted by Sugiyama et al. [12] and Widyantoro et al. [13]. Both studies use the supervised machine learning method for the classification process. The accuracy of this method is largely defined by the reliability of the used features. Sugiyama et al. [12] use 6 main features for the classification process, including unigram, bigram, position, proper Noun, previous and next-sentence, and orthographic, while Widyantoro et al. [13] use citation style-related features. Widyantoro et al. use four forms of citation style to identify the citation sentence.

Unfortunately, in the implementation, Sugiyama et al. only use the "accuracy" measure in determining system performance based on the proposed feature even though their research focus is on investigating the extent to which the proposed feature is able to identify the citation sentence appropriately. The use of "accuracy" measure becomes less precise because the obtained value is greatly influenced by the system performance in identifying non-citation sentence as well. The use of citation style-related features is even more promising to be used for identifying citation sentence.

3. Identification of citation-based paper relations
The existence of a paper relation can be first identified using citation and subsequently be detected using a particular cue phrase [4]. The accuracy of paper relation identification significantly depends on the accuracy of citation detection in every sentence in scientific papers. There are four types of scientific paper relations that can be identified according to the citation context existing in citation sentence, including extend, criticize, compare, and improve. These four types of paper relation also intersect with the citation function grouping of a citation sentence contained in the 12 categories as performed by [14].

4. The proposed system designs
The system design of learning classifier model for detecting paper relations conducted in this study can be seen in Figure 1. The process begins with Reading Data from a collection of pdf papers that will be processed. The data used in this study is 30 pdf English language articles that are obtained from the internet for free. The data reading process utilizes pdfBox library so that the pdf files can be read as text files. The process is also combined with the Sentence Splitting process, which is an automatic extraction of sentences from a parent article using the breakIterator function. The whole sentences obtained from the reading and separation process are later stored in a single database. The stored database (citation dataset) contains information on id_file, id_sentence, and sentence content. The number of sentences generated from this automatic extraction process is as many as 5878.

For the learning process, an annotation must be done for each sentence as generated from the separation process. This annotation means class labelling process for each sentence as a "citation" or "non-citation" class. The annotation process generates 874 citation sentences and 5004 non-citation sentences. The number of the automatically generated sentences is slightly different from the number of manually generated sentences considering that the accuracy of breakIterator function has not reached 100%. After the annotation process is finished, the next is feature extraction, i.e. a process to generate citation style-based features. The dataset and the resulting features are then used for the learning process. The classifier model used in this study includes Support Vector Machine (SVM) and Naïve Bayes (NB).

The evaluation method used in this study is 5-fold Cross Validation based on id-paper. The determination of this evaluation method aims to make the formation of a model more objective than cross validation based on random sentences. This is because in a paper, there will likely be several sentences that have similar writing patterns. The performance of relation identification through automatic sentence extraction process is then compared with the performance of relation identification through manual sentence extraction process. In addition to comparing the number of citation sentences
identified, the performance measurement is also done by considering the F-Measure value obtained from each process. The comparison of these two processes will identify the accuracy of automatic extraction process in the detection of paper relations.

**Figure 1.** Model learning of relation identification

5. **Features used**

The accuracy of citation identification is defined by the use of appropriate features. In this study, the citation feature used to detect citation includes four citation style features as proposed by [13]. We use 4 features in [13] because these four features provide good results. The four citation style features along with the regular expression and its feature explanation can be seen in Table 1.

| Regular Expressions | Name of Feature                     | Description                                                                 |
|---------------------|-------------------------------------|----------------------------------------------------------------------------|
| m/(\d{4})/g         | Textual Syntactic                   | Declare patterns in the form of: "digits with 4 numbers" flanked by brackets. Example: .... Sibaroni (2015) |
| m/(\[A-Z][^()]*\d{4}\])/g | Textual Parenthetical            | Patterns of letters, punctuation, and numbers consisting of 4 digits appear in sequence. Example: Ahuja, Magnanti, and Orlin, 1995 |
| (m/(\{1,3\})/g)     | Numbered 1                         | Pattern: There are 1 to 3 digit numbers in a square bracket. Example: .... and PolyBench [5,7] |
| (m/(\{1,3\}.*\})/g | Numbered 2                         | Pattern: There are 1 to 3 digit numbers (and can be repeated more 1 time) in an square bracket |

---

Table 1. Features for citation detection [13]
The performance of citation sentence identification through automatic sentence extraction is then compared with that of citation sentence identification through manual sentence extraction. The number of citation sentences in the data set, which is generated through manual sentence splitting, serves as the reference to the performance measurement of automatic sentence splitting. The evaluation method used is a 5-cross validation method based on id-paper, while the classifiers used are Naïve Bayes and Support Vector Machine (SVM).

6. Experimental results and evaluation results
The results of citation sentence identification through manual sentence extraction are shown in Table 2, while the ones of citation sentence identification through automatic sentence extraction are presented in Table 3. The system performance is evaluated by taking into account the ratio of the number of successfully detected citations between automatic and manual sentence extraction. Furthermore, the F-Measure values of the two processes are also compared.

Table 2. Citation identification performance for manual sentence extraction

| Actual/prediction | Naïve Bayes / SVM | 
|-------------------|-------------------|
|                   | citation | non-citation |
| citation          | 632      | 264         |
| non-citation      | 34       | 5738        |

Based on the experimental results as shown in Table 2, the SVM and NB classifiers generate the same results in detecting citations contained in all sentences obtained using manual extraction. As many as 632 citation sentences are identified from a total of 896 citation sentences. Meanwhile, the F-Measure calculation results in 81%, which means the performance is quite good.

Table 3. Citation identification performance for automatic sentence extraction

| Actual/prediction | Naïve Bayes / SVM | 
|-------------------|-------------------|
|                   | citation | non-citation |
| citation          | 616      | 258         |
| non-citation      | 74       | 4930        |

According to the experimental results as illustrated in Table 3, the SVM and NB classifiers give the same result in detecting citations contained in all sentences obtained using automatic extraction. As many as 616 citation sentences are identified from a total of 874 citation sentences.

The number of citation sentences generated from automatic extraction is slightly different from that generated from manual extraction. The calculation of F-Measure results in 79%, which means the performance is quite good, even approaching the automatic extraction performance (81%). The comparison between the number of citation sentences obtained through automatic extraction and that obtained through manual extraction generates a value of 97.5%.

Based on the analysis results, it can be concluded that the process of automatic sentence extraction using the breakIterator function gives very satisfactory results from which the results of citation sentence detection are close to those of citation sentence detection using manual sentence extraction.

7. Conclusion and future work
In general, the obtained results are very satisfactory. The proposed of automatic sentence extraction system by utilizing the library function available in java programming language for the process of relation detection provides excellent performance. This means the development of business applications for the detection of paper relations based on this extraction system is very feasible to do.

Further research that can be done is to conduct experiments on a larger scale on scientific article data with larger size and more varied fields. The text quality of scientific papers to be processed
should also be considered because it will affect the obtained file reading results. The identification process of paper relations, including extend, criticize, compare, and improve by directly inputting pdf paper data can also be done for the detection of a more specific paper relation. The encouraged to develop appropriate features to identify each type of paper relation is the next research challenge that can be tried to do. Furthermore, an experiment also needs to be conducted in other classifiers in addition to Naïve Bayes and SVM to obtain the best system performance.

References
[1] M. Ware and M. Mabe, “The stm report,” *STM Rep.*, no. September, p. 68, 2009.
[2] Y. Sibaroni, D. H. Widyanto, and M. L. Khodra, “Extend Relation Identification in Scientific Papers Based On Supervised Machine Learning,” in *The 2016 International Conference on Advanced Computer Science and Information Systems (ICACSI 2016)*, 2016.
[3] Y. Sibaroni, D. H. Widyanto, and M. L. Khodra, “Survey on Research Paper’s Relations,” *Int. Conf. Inf. Technol. Syst. Innov.*, 2015.
[4] W. Wang, P. Villavicencio, and T. Watanabe, “Analysis of reference relationships among research papers, based on citation context,” *Int. J. Artif. Intell. Tool*, vol. 21, no. 2, pp. 1–24, 2012.
[5] T. C. Hoad and J. Zobel, “Methods for Identifying Versioned and Plagiarised Documents,” *JASIST*, vol. 54, pp. 203–215, 2003.
[6] H. Xu, E. Martin, and A. Mahidadia, “Contents and time sensitive document ranking of scientific literature,” *J. Informetr.*, vol. 8, no. 3, pp. 546–561, 2014.
[7] Y. Kyung Jeong, M. Song, and Y. Ding, “Content-based author co-citation analysis,” *J. Informetr.*, vol. 8, no. 1, pp. 197–211, 2014.
[8] B. Gipp, J. Beel, and C. Hentschel, “Scienstein : A Research Paper Recommender System,” *Proc. Int. Conf. Emerg. Trends Comput. (ICETIC’09), pages 309–315, Virudhunagar (India), January 2009. Kamaraj Coll. Eng. Technol. India, IEEE*, no. January, pp. 309–315, 2009.
[9] B. Gipp and J. Beel, “Identifying related Documents for Research Paper Recommender by CPA and COA,” *Proc. World Congr. Eng. Comput. Sci. 2009 Vol I WCECS 2009, Oct. 20-22, 2009, San Fr. USA*, vol. I, pp. 20–23, 2009.
[10] Y. Liang, Q. Li, and T. Qian, “Finding Relevant Papers Based on Citation Relations,” *Lect. Notes Comput. Sci.*, vol. 6897, no. Department of Computer Science, City University of Hong Kong, pp. 403–414, 2011.
[11] S. Teufel, a Siddharthan, and D. Tidhar, “An annotation scheme for citation function,” *Proc. 7th SIGdial Work. Discourse Dialogue*, pp. 80–87, 2006.
[12] K. Sugiymama, T. Kumar, M.-Y. Kan, and Ramesh C. Tripathi, “Identifying Citing Sentences in Research Papers Using Supervised Learning,” 978-1-4244-5651-2/10/$26.00 ©2010 IEEE, pp. 67–72, 2010.
[13] D. H. Widyanto and I. Amin, “Citation Sentence Identification and Classification for Related Work Summarization,” *Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSI), Jakarta, Indones.*, 2014.
[14] S. Teufel and D. Tidhar, “Automatic classification of citation function,” *Proc. EMNLP-06, Sydney, Aust.*, 2006.