SCIENCE, TECHNOLOGY & INNOVATION INDICATORS

Open indicators: innovation, participation and actor-based STI indicators
PARIS 2017

First Stage of an Automated Content-Based Citation Analysis Study: Detection of Citation Sentences[1]

Zehra Taşkın[*], Umut Al[*] and Umut Sezen[**]

[*]{ztaskin; umutal}@hacettepe.edu.tr
Department of Information Management, Hacettepe University, Beytepe, Çankaya, Ankara, 06800 (Turkey)

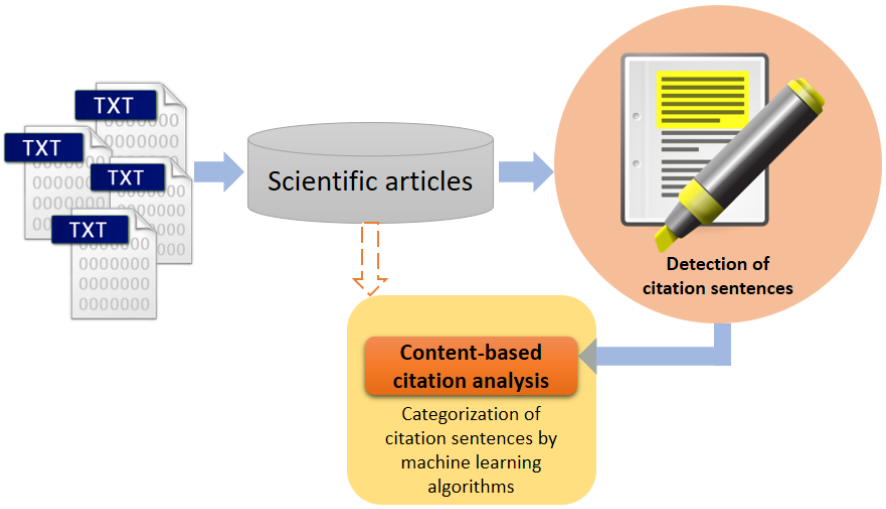
[**]u.sezen@ee.hacettepe.edu.tr
Department of Electrical and Electronics Engineering, Hacettepe University, Beytepe, Çankaya, Ankara, 06800 (Turkey)

INTRODUCTION

Content-based citation analyses, which are mainly focused on giving meaning to the citations, are important studies in the literature recently in terms of ethical problems and manipulations on citations. As the number of published articles increases year-by-year enormously in the world, an uncontrollable mass of publications has begun to arise, and with this mass of publications, authors have lost their motivations to cite by accessing the best of them. This phenomenon is known as "publish or perish" in the academia (Angell, 1986). Academic promotions or funds are given by using the number of publications and citations. However, as mentioned, authors do not act much selective when they cite to someone. These types of citations are denominated as "perfunctory citations" in the literature (Moravcsik & Murugesan, 1975). Undoubtedly, all citations are not perfunctory, besides, studies in the literature confirm that the vast majority of references (app. up to 70%) are perfunctory citations (e. g. Athar, 2011, p. 82; Cano, 1989, p. 286; Xu, Zhang, Wu, Wang, Dong & Xu, 2015, p. 1338). In this case, it becomes important to evaluate the meanings of citations, so the concept called "content-based citation analysis" is emerged. The main capability of content-based analysis is the analysis of a citation's context within the full text of the scientific paper rather than its simple frequency (Ding, Zhang, Chambers, Song, Wang & Zhai, 2014, p. 1821). These analyses make it possible to distinguish between perfunctory citations and meaningful ones. Especially with the development of computational linguistics techniques, the analyses have started to be done more easily. The main phases of an automated content-based citation study is shown on Figure 1.

[1] This study is supported by a research grant (no: 115K440) of the Turkish Scientific and Technological Research Center (TÜBİTAK).

Figure 1: The main phases of a content-based citation analysis study.



At the beginning of a content-based study, it is important to create a database, which contains scientific articles. After creating the database, there are two options to conduct the process; gathering citations manually or detecting citation sentences by using machine learning algorithms or other automated techniques. Automated techniques provide a useful method that requires least human effort.

This study is a research in progress based on a project entitled "Designing a Content-Based Citation Analysis Model for Turkish Citations" which is supported by a research grant of the Turkish Scientific and Technological Research Center (TÜBİTAK). The main aim of the project is to categorize citations by semantic and syntactic classes automatically to pre-determined citation classes. In the process of designing this model, the references and citation sentences of 423 articles, which were published in the two main library and information science journals, *Turkish Librarianship* and *Information World*, in Turkey, are classified by experts in this field. Firstly, experts determined the citation sentences manually. Then, at least two experts tagged each citation sentence to provide inter-annotator agreement which provides accuracy and cross-validation of tagging quality (Bhowmick, Mitra & Basu, 2008, p. 58). The categorization performance of the positive, negative and neutral citations is measured as 96% (*f*=0,965/Naïve Bayes Multinomial) (Taşkın, 2017, p. 56). One of the important findings of this study is revealing patterns of the citation sentences. In this case, it would be possible to develop an automated citation extraction system using the tagging patterns made by the experts. This paper proposes an automated citation extraction model by using finite state grammars to make content-based citation analysis easy and practical.

Citation Extraction and Finite State Grammars

The method of information extraction depends on the identification, tagging and extraction of key elements (such as person, institution, location, country information) from high volume texts (Liddy, 2010, p. 3871). This method may form the basis of the natural language processing tasks. Because, in some studies, while the information in the text leads to the conclusion; the data obtained by the extraction of this information is processed by the other studies. For this reason, information extraction is used with other methods such as summarization, text categorization and etc. (Blake, 2013, p. 129). Finite state grammars are one way of information extraction tasks. These grammars are used -for determining whether a

word or a whole sentence has an acceptable regular language. The algorithm reads each word from left to right and labels the transitions. If a transition is labelled with the same symbol as the initial state, the algorithm goes to the next state when the current state is ended. This process continues until the last state (Galvez & Moya-Anegón, 2007b, p. 9).

Finite state grammars are used in a wide range of domains, including pattern matching and recognition, speech processing, handwriting recognition, optical character recognition, encryption algorithms or data compression and indexing (Roche and Schabes, 1997, p. 227). Finite state grammars are used in bibliometrics for recent studies generally for accessing accurate data (e.g. Galvez & Moya-Anegón, 2006; Galvez & Moya-Anegón, 2007a; Galvez & Moya-Anegón, 2012; Taşkın & Al, 2014). However, studies which extract citation sentences from full texts have not taken place in the literature sufficiently. There is only one study that focused on the extraction of citation sentences is published in 2014 (Kim, Le & Thoma, 2014). This study uses the text categorization method and SVM algorithm to obtain citation sentences. The success rate of the algorithm is 96% (f=96.99). Text categorization seems to be an effective technique for extracting citation sentences.

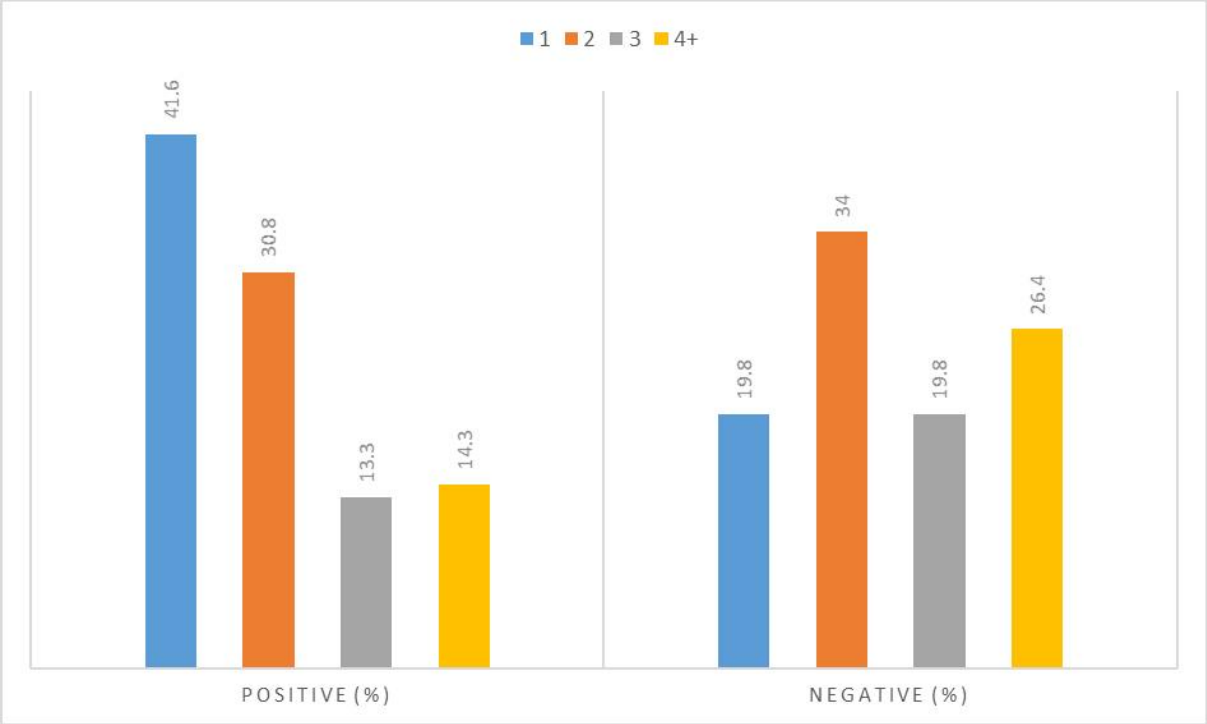
METHODOLOGY AND PRELIMINARY FINDINGS

In order to create the automated citation sentence extraction model, a software called *NooJ: A Linguistic Development Environment*, which is developed to construct large-coverage formalized descriptions of natural languages and to apply them to large corpora in real time, is used (Silberztein, 2003, p. 7). Nooj graphical grammar editor is used to create the structure of the model. First of all, a graphical model is created to access citation sentences which are determined and categorized by the experts. The two journals use APA style; therefore, the model is designed to extract citations given only in the APA style. The rules defined in this pattern creation phase are as follows:

- Each sentences are ended with "." (dot) or ":" (colon) marks. However, there is some cases where the dot sign does not represent the end of the sentence such as *"vb. (e.g.)" and "vs. (e.g.)"*, numbers in Latin alphabet or titles *(e.g. Dr.)*. All of these usages are determined and are excluded from the system structure.
- A citation sentence may contain one or more sentences. The connection between sentences are provided by the conjunctions such as *"bu bağlamda (in this context)"*, *"bu (this)"*, *"ancak (however)"*, *"burada (along this)"* and *"yine de (nevertheless)"*.
- There are different arrays of citations possible in citation sentences. All the mentioned options below are considered during the creation of the model.
    - Citation to single author (e.g. *(xxx, 1997), (yyy, 1998, p. 28) or (zzz, 2000: 13)*).
    - Citation to multiple authors in a single citation sentence (e.g. *(xxx, 1997; yyy, 1998, p. 28; zzz, 2000)*).
    - Citation to the authors by mentioning the author name (e.g. *According to Taşkın's study (2017) …*)

Preliminary findings show that 63.8% of the citations contain single sentence constructs. The 23.4% of them are two-sentence citations. Rest of the citations (12.8%) consist of three or more sentences. There are even citations, which consist of 35 sentences. It is important to extract citations accurately to enhance the quality of content-based citation analysis. Because if an important sentence is skipped and is not included to the analysis, the correct meaning of citation may not be detected. It is revealed that the authors, who make positive or negative citations, prefer more than one sentence. Figure 2 shows the distribution of number of sentences into the positive and negative citation classes.

Figure 2: Distribution of the number of citations into two citation classes.



According to the Figure 2, it is obvious that authors prefer more than one sentence to make a citation.

The importance of this study is to create an automated extraction tool for scientific articles in any language or field, which use APA style. This tool may be used easily with the purpose of providing meaningful citation data to all content-based citation analysis studies carried out by the decision makers and managers. In this way, it is possible to reduce the need for human power in such studies and to increase focus on the meanings of the citations.

References

Angell, M. (1986). Publish or perish: A proposal. *Annals of Internal Medicine*, 104(2), 261-262.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *HLT-SS '11 Proceedings of the ACL 2011 Student Session* (p. 81-87). Stroudsburg: Association for Computational Linguistics.

Bhowmick, P.K., Mitra, P. & Basu, A. (2008). An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics* (p. 58-65). Manchester: ACLWeb.

Blake, C. (2013). Text mining. *Annual Review of Information Science and Technology*, 45(1), 121-125.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the Association for Information and Technology*, 40(4), 284-290.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X. & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information and Technology*, 65(9), 1820-1833.

Galvez, C. & Moya-Anegón, F. (2006). An evaluation of conflation accuracy using finite-state transducers. *Journal of Documentation*, 62(3), 328-349.

Galvez, C. & Moya-Anegón, F. (2007a). Approximate personal name-matching through finite-state graphs. *Journal of the American Society for Information Science and Technology*, 58(13), 1-17.

Galvez, C. & Moya-Anegón, F. (2007b). Standardizing formats of corporate source data. *Scientometrics*, 70(1), 3-26.

Galvez, C. & Moya-Anegón, F. (2012). A dictionary-based approach to normalizing gene names in one domain of knowledge from the biomedical literature. *Journal of Documentation*, 68(1), 5-30.

Kim, C., Le, D.X. & Thoma, G.R. (2014). Automated method for extracting "citation sentences" from online biomedical articles using SVM based text summarization technique. In *IEEE International Conference on Systems, Man and Cybernetics, October 5-8, 2014*. San Diego, CA: IEEE.

Liddy, E.D. (2010). Natural language processing. In *Encyclopedia of Library and Information Sciences, Third Edition* (p. 3864-3873). New York: Taylor and Francis.

Moravcsik, M.J. & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5, 86-92.

Silberztein, M. (2003). *NooJ manual*. Retrieved 27 April, 2017 from https://www.researchgate.net/publication/276186794_NooJ_Manual

Roche, E. & Schabes, Y. (1997). *Finite-state language processing (language, speech and communication)*. Cambridge: The MIT Press.

Taşkın, Z. & Al, U. (2014). Standardization problem of author affiliations in citation indexes. *Scientometrics*, 98(1), 347-368.

Taşkın, Z. (2017). Designing a model for content-based citation analysis: An application for Turkish citations based on text categorization. Unpublished PhD dissertation, Hacettepe University, Turkey.

Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., ve Xu, H. (2015). Citation sentiment analysis in clinical trial papers. *AMIA Annual Symposium Proceedings*, 2015, 1334-1341.