Computational Linguistics

December 2013, Vol. 39, No. 4, Pages 917-947
Posted Online November 20, 2013.
(doi:10.1162/COLI_a_00153)

Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection

Alberto Barrón-Cedeño[*†]
Universitat Politècnica de Catalunya

Marta Vila[**†]
Universitat de Barcelona

M. Antònia Martí[‡]
Universitat de Barcelona

Paolo Rosso[§]
Universitat Politècnica de València

*TALP Research Center, Jordi Girona Salgado 1-3, 08034 Barcelona, Spain. E-mail: albarron@lsi.upc.es.

**CLiC, Department of Linguistics, Gran Via 585, 08007 Barcelona, Spain. E-mail: marta.vila@ub.edu.

†Both authors contributed equally to this work.

‡CLiC, Department of Linguistics, Gran Via 585, 08007 Barcelona, Spain. E-mail: amarti@ub.edu.

§NLE Lab-ELiRF, Department of Information Systems and Computation, Camino de Vera s/n, 46022 Valencia, Spain. E-mail: prosso@dsic.upv.es.

Although paraphrasing is the linguistic mechanism underlying many plagiarism cases, little attention has been paid to its analysis in the framework of automatic plagiarism detection. Therefore, state-of-the-art plagiarism detectors find it difficult to detect cases of paraphrase plagiarism. In this article, we analyze the relationship between paraphrasing and plagiarism, paying special attention to which paraphrase phenomena underlie acts of plagiarism and which of them are detected by plagiarism detection systems. With this aim in mind, we created the P4P corpus, a new resource that uses a paraphrase typology to annotate a subset of the PAN-PC-10 corpus for automatic plagiarism detection. The results of the Second International Competition on Plagiarism Detection were analyzed in the light of this annotation.

The presented experiments show that (i) more complex paraphrase phenomena and a high density of paraphrase mechanisms make plagiarism detection more difficult, (ii) lexical substitutions are the paraphrase mechanisms used the most when plagiarizing, and (iii) paraphrase mechanisms tend to shorten the plagiarized text. For the first time, the paraphrase mechanisms behind plagiarism have been analyzed, providing critical insights for the improvement of automatic plagiarism detection systems.

Quarterly (March, June, September, December)
160 pp. per issue
6 3/4 x 10
Founded: 1974
ISSN 0891-2017
E-ISSN 1530-9312
2014 Impact Factor: 1.226

Cited by

Juan D. Velásquez, Yerko Covacevich, Francisco Molina, Edison Marrese-Taylor, Cristián Rodríguez, Felipe Bravo-Marquez. (2016) DOCODE 3.0 (DOcument COpy DEtector): A system for plagiarism detection by applying an information fusion process from multiple documental data sources. *Information Fusion* 2764-75.
Online publication date: 1-Jan-2016.
CrossRef

Horacio Saggion, Stefan Bott, Luz Rello. (2016) Simplifying words in context. Experiments with two lexical resources in Spanish. *Computer Speech & Language* 35200-218.
Online publication date: 1-Jan-2016.
CrossRef

Solange de L. Pertile, Viviane P. Moreira, Paolo Rosso. (2015) Comparing and combining Content- and Citation-based approaches for plagiarism detection. *Journal of the Association for Information Science and Technology*n/a-n/a.
Online publication date: 1-Sep-2015.
CrossRef

Marta Vila, Manuel Bertran, M. Antònia Martí, Horacio Rodríguez. (2015) Corpus annotation with paraphrase types: new annotation scheme and inter-annotator agreement measures. *Language Resources and Evaluation* 4977-105.
Online publication date: 1-Mar-2015.
CrossRef

Marta Vila, M. Antònia Martí, Horacio Rodríguez. (2014) Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology. *Open Journal of Modern Linguistics* 04205-218.
Online publication date: 1-Jan-2014.
CrossRef