Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 103-105 recibido 21-11-12 revisado 15-01-13 aceptado 19-02-13

On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism *

Detección de texto reutilizado y plagio monolingüe y translingüe

Alberto Barrón Cedeño
DSIC Universitat Politècnica de València

TALP Research Center
Universitat Politècnica de Catalunya
albarron@[lsi.upc.edu | gmail.com]

Resumen: Tesis de doctorado en ciencias de la computación (con mención europea del doctorado) escrita por Alberto Barrón Cedeño bajo la supervisión del Dr. Paolo Rosso en la Universitat Politècnica de València. El autor fue examinado en Valencia en julio de 2012 por un jurado compuesto por los siguientes doctores: Paul Clough (*University of Sheffield*), Benno Stein (*Bauhaus-Universität Weimar*), Ricardo Baeza-Yates (*Yahoo! Research*), Fabio Crestani (*Università della Svizzera italiana*) y José Miguel Benedí (Universitat Politècnica de Valencia). La mención europea fue obtenida tras una estancia de 4 meses en la *Information School* de la *University of Sheffield* (Reino Unido) bajo la supervisión del Dr. Paul Clough.
Palabras clave: recuperación de información translingüe, plagio traducido, texto reutilizado, plagio parafrástico, plurilingüismo en Wikipedia

Abstract: Ph.D. thesis (European doctorate mention) in Computer Science written by Alberto Barrón Cedeño under the advice of Dr. Paolo Rosso at the Universitat Politècnica de València. The author was examined in Valencia in July 2012 by a jury composed of the following doctors: Paul Clough (University of Sheffield), Benno Stein (Bauhaus-Universität Weimar), Ricardo Baeza-Yates (Yahoo! Research), Fabio Crestani (Università della Svizzera italiana), and José Miguel Benedí (Universitat Politècnica de Valencia). The European mention was received after a 4 months internship at the Information School of the University of Sheffield (UK) under the advice of Dr. Paul Clough.
Keywords: cross-language information retrieval, re-used text, cross-language plagiarism, paraphrase plagiarism, Wikipedia multilingualism

1. Introduction

Automatic text re-use detection is the task of determining whether a text has been produced by considering another as its source. Plagiarism, the unacknowledged re-use of text, has gained the greatest notoriety. Favoured by the easy access to information through electronic media, plagiarism has raised in recent years, requesting for the attention of experts in text analysis.

Automatic text re-use detection takes advantage of NLP and IR technology to compare thousands of documents —looking for the potential source of a presumably case of re-use. Machine translation technology can be used in order to uncover cases of cross-language re-use. By exploiting such technology, thousands of exhaustive comparisons are possible, also across languages, something impossible to manually achieve.

In this dissertation we pay special attention to three aspects of text re-use:

1. Cross-language text re-use: we propose a cross-language similarity assessment model that represents one of the best

* This PhD research was supported by the National Council of Science and Technology of Mexico (CONACyT) through the 192021/302009 scholarship. The Ministry of Education of Spain supported my internship in the University of Sheffield through the TME2009-00456 grant. The investigation was carried out in the framework of the MICINN project Text-Enterprise 2.0 (TIN2009-13391-C04-03).

ISSN 1135-5948

options when looking for exact translations.

2. Paraphrase text re-use: we investigate what types of paraphrasing are more frequently applied when plagiarising and how they difficult plagiarism detection; something never done before.

3. Mono- and cross-language re-use within and from Wikipedia: the encyclopedia is explored as a multi-authoring framework, where texts are re-used within versions of an article and across languages.

2. Thesis Overview

The dissertation consists of 9 chapters, describing our efforts to approach the main difficulties of automatic text re-use detection. The contents are described following.

Chapters 2 and 3 are an overall introduction of the covered topics. Chapter 2 offers an overview of text re-use, with special emphasis on plagiarism. Our contribution comes in the form of the survey we held in different Mexican universities; aiming to assess how often students plagiarise across languages and their attitudes respect to paraphrase plagiarism (factors never analysed before). Chapter 3 introduces the IR and NLP concepts used through the rest of the thesis.

Chapter 4 describes corpora for (automatic) analysis of text re-use and plagiarism available up to date. Our participation in the construction of three corpora —coderivatives, CL!TR, and to a smaller extent PAN-PC— are cutting edge contributions discussed in this chapter. Evaluation metrics are also discussed: some are well known in IR and related areas, whereas others were recently proposed —and specially designed— for evaluating text re-use detection.

Chapter 5 defines the two main approaches to re-use detection: intrinsic and external. Our contributions to external (monolingual) detection are discussed. Our main contribution is a model for retrieving those related documents to the suspicious one, hence reducing the load when performing the actual plagiarism detection process. Such a problem is often neglected in the plagiarism detection literature, that *assumes* that either the step is not necessary or it is already solved; an absolutely false idea.

Chapter 6 describes our model for cross-language detection (this is one of the least approached problems of re-use detection!): CL-ASA. CL-ASA is compared to state-of-the-art models over different sub-tasks of the detection process. A variety of languages is considered to analyse the strengths and weaknesses of the different models.

Chapter 7 discusses the international competitions we ran during three years: the PAN International Competition on Plagiarism Detection. We also experiment with our detection models on the generated test-beds and discuss the obtained results.

Chapter 8 analyses plagiarism from the point of view of paraphrasing, providing a bridge between the two disciplines: plagiarism detection and paraphrase analysis. Our findings on the use of paraphrasing when plagiarising represent useful insights to take into account when developing the next generation of plagiarism detection systems.

In Chapter 9 we analyse monolingual coderivation among revisions of Wikipedia articles and cross-language text re-use from Wikipedia. Related to the latter issue, we offer a preliminary discussion on the PAN competition we organised at FIRE on cross-language text re-use: PAN Cross-Language !ndian Text Re-Use; where the potentially re-used documents were written in Hindi and the potential source documents were written in English.

3. Thesis Contributions

The main contributions of this research are described below.

Detection of text re-use across languages. We explored a range of cross-language information retrieval techniques. We observed that (*i*) a simple model based on characterising texts by short character n-grams (CL-CNG) was worth considering when dealing with common-alphabet languages (and different alphabets, after transliteration), and particularly if they have some influence (Barrón-Cedeño et al., 2010; Potthast et al., 2011); (*ii*) the model cross-language explicit semantic analysis (CL-ESA), based on large comparable corpora such as Wikipedia, performs well when looking for related documents across languages (Potthast et al., 2011). We proposed a model —cross-language alignment-based similarity analysis, CL-ASA—, based on translation probabilities and length distributions between texts (Barrón-Cedeño et al., 2008; Pinto et al., 2009). Our empirical results showed that

CL-ASA is competitive when looking for reused texts, regardless if they were manually or automatically translated. CL-ASA performs better than CL-ESA and CL-CNG, identified as two of the most appealing models for cross-language similarity assessment, when dealing with translations at document and fragment level (Potthast et al., 2011).

Creation of standard collections of documents for the study and development of plagiarism detection. We helped in the creation of two "sister" corpora with simulated cases of re-use and plagiarism. The PAN-PC series look at composing a realistic IR challenge: it includes thousands of documents, with thousands of plagiarism cases (both manually and automatically generated) (Potthast et al., 2010). The CL!TR corpus looks at composing a realistic cross-language challenge: it contains a few thousand documents, with hundreds of re-use cases (manually generated across distant languages) (Barrón-Cedeño et al., 2011). These corpora (particularly the PAN-PC series) have become a reference in the development of models for plagiarism detection, filling an important gap.[1]

Analysis of paraphrase plagiarism and its detection. The vast majority of models for text re-use detection are designed to uncover "cut and paste" cases, as they consider surface information only. These models are unsuccessful when facing paraphrase plagiarism. For the first time, we analysed the paraphrase phenomena applied when text is plagiarised (Barrón-Cedeño et al., 2013 (to appear)). Our seminal study showed that lexical substitutions are the paraphrase mechanisms used the most. Moreover, the paraphrasing tends to be used to generate a simplified version of the re-used text. A model intended to succeed in detecting paraphrase re-use requires robust text pre-processing and characterisations: the expansion (or contraction) of related vocabulary, the normalisation of formatting and word forms, and the inclusion of mechanisms that model the expected length of a re-used fragment given its source.

[1]The PAN-PC corpora, created in the framework of the PAN International Competition on Plagiarism Detection, are available at `http://www.uni-weimar.de/cms/medien/webis/research/corpora.html`. The CL!TR corpus, created in the framework of the PAN Cross-Language !ndian Text Reuse challenge, is available at `http://memex2.dsic.upv.es/workshops/2011/clitr/`

References

Barrón-Cedeño, Alberto, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In Huang and Jurafsky (Huang and Jurafsky, 2010).

Barrón-Cedeño, Alberto, Paolo Rosso, Sobha Lalitha Devi, Paul Clough, and Mark Stevenson. 2011. PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition. In FIRE, editor, *FIRE 2011 Working Notes. Third Workshop of the Forum for Information Retrieval Evaluation.*

Barrón-Cedeño, Alberto, Paolo Rosso, David Pinto, and Alfons Juan. 2008. On Cross-Lingual Plagiarism Analysis Using a Statistical Model. In Benno Stein, Efstathios Stamatatos, and Moshe Koppel, editors, *ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008)*, volume 377, pages 9–13, Patras, Greece. CEUR-WS.org. http://ceur-ws.org/Vol-377.

Barrón-Cedeño, Alberto, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013 (to appear). Plagiarism meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics.*

Huang, Chu-Ren and Dan Jurafsky, editors. 2010. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, August. COLING 2010 Organizing Committee.

Pinto, David, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, and Paolo Rosso. 2009. A Statistical Approach to Crosslingual Natural Language Tasks. *Journal of Algorithms*, 64(1):51–60.

Potthast, Martin, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Cross-language plagiarism detection. *Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis*, 45(1):1–18.

Potthast, Martin, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In Huang and Jurafsky (Huang and Jurafsky, 2010), pages 997–1005.