Cross-Language Plagiarism Detection Methods

Vera Danilova
Dept. of Romance Languages, Autonomous University of Barcelona, Spain
maolve@gmail.com

Abstract

The present paper provides a summary on the existing approaches to plagiarism detection in multilingual context. Our aim is to organize the available data for the further research. Considering distant language pairs is of a particular interest for us. Cross-language plagiarism detection issue has acquired pronounced importance lately, since semantic contents of a document can be easily and discreetly plagiarized through the use of translation (human or machine-based). We attempt to show the development of detection approaches from the first experiments based on machine translation pre-processing to the up-to-date knowledge-based systems that proved to obtain reliable results on various corpora.

1 Introduction

According to Barrón-Cedeño *et al.* (2008), cross-language plagiarism detection (CLPD) consists in discriminating semantically similar texts independent of the languages they are written in, when no reference to the original source is given. However, here *similar* means that the objects (texts) share only certain characteristics and are comparable, whereas plagiarism has to do with the cases when author's original words and ideas are copied (with or without formal modifications). As follows from an updated version of the definition in Barrón-Cedeño (2012) a cross-language plagiarism case takes place when we deal with unacknowledged reuse of a text involving its translation from one language to another.

As indicated by Barrón Cedeño (2012) no technologies were developed for CLPD purposes before 2008. Since the establishment of the International Competition on Plagiarism Detection as a part of the workshop PAN (*Uncovering Plagiarism, Authorship and Social Software Misuse*) in 2009, cross-lingual issues started to draw attention of the participants. In 2010 there were attempts of using machine translation (MT) at the document pre-processing step in order to deal with non-English documents as possible sources of plagiarism. The detailed comparison of sections was implemented using traditional monolingual methods. The main problems that manifested themselves immediately were computational cost and quality of MT that is so far unable to permit reliable comparison of suspicious texts and sources. Moreover, authors tend to modify translated texts using paraphrases, which makes the discrimination process even more complicated. Also, one of the main challenges is the presence of salient distinctions in syntactic structures of languages belonging to different families.

It was already in 2008 that the researchers started to come up with new strategies for avoiding the MT step. Barrón Cedeño (2008) proposed a statistical approach based on parallel corpora for the CLPD. In Lee *et al.* (2008), a text categorization approach was posited. Domain-specific classification was performed using support vector machine model and parallel corpora containing Chinese-English text pairs. Similarity measurement was carried out by means of language-neutral clustering based on Self-Organizing Maps (SOM). Ceska *et al.* (2008) proposed a tool named MLPlag based on the word location analysis. EuroWordNet thesaurus was used for language-independent text representation (synonym normalization). Detailed comparison was performed by computing both symmetric (VSM-based) and asymmetric similarity measures, which required a preliminary calculation of occurrence frequency of plagiarized words. Multilingual pre-processing involving lemmatization and inter-lingual indexing anticipated the comparison.

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 51–57,
Hissar, Bulgaria, 9-11 September 2013.

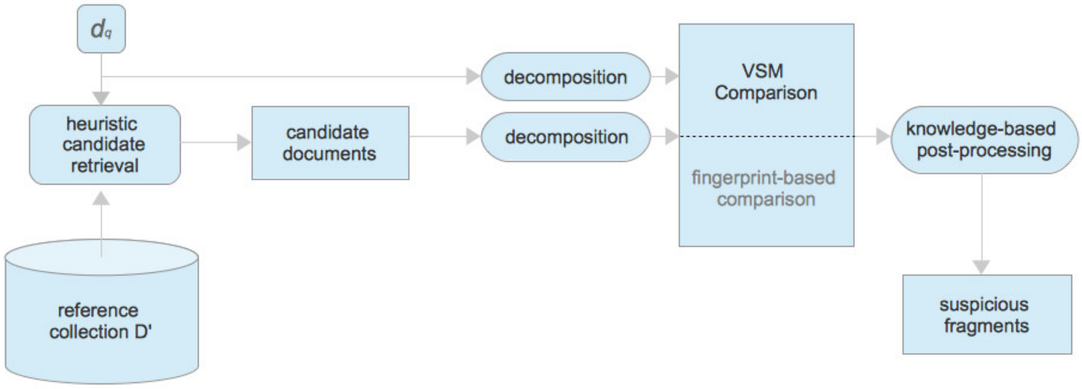

Figure 1: Plagiarism detection process (adapted from Potthast *et al.* (2011).

Despite of the disadvantages of MT-based approach, it was not discarded by the researchers. As Meuschke and Gipp (2013) point out, it is suitable for small document collections. In the subsequent sections we describe the application of MT and other approaches more in detail.

2 Related Work

The surveys by Potthast *et al.* (2011) and Barrón Cedeño *et al.* (2013) were dedicated exclusively to the classification and evaluation of CLPD methods. Also, a large description of CLPD technology is provided in the doctoral thesis by Barrón Cedeño (2012). Potthast *et al.* (2011) outline the steps of CLPD process, provide some strategies of heuristic retrieval and evaluate the performance of three models for the detailed analysis. Barrón Cedeño *et al.* (2013) enrich this survey by describing the whole architecture of plagiarism analysis. Also, a modification to the classification of detailed analysis methods is introduced and an evaluation of three other models is provided.

The rest of the article is organized as follows: Section 3 introduces the main approaches to CLPD, explains the prototypical structure of analysis and outlines the performance evaluation, presented in the previous surveys; Section 4 concludes the paper.

3 Approaches to CLPD

3.1 Intrinsic VS External CLPD

Barrón Cedeño (2012) divides CLPD methods into intrinsic and external, because, as shown in the literature, intrinsic plagiarism detection techniques allow to discriminate the so called *effects of translation process* inside the text. Some of the relevant indicators found by researchers are as follows: function words, morphosyntactic categories, personal pronouns, adverbs (in 2006 by Baroni and Bernardini); animate pronouns, such as *I, we, he*, cohesive markers, such as *therefore, thus* (in 2011 by Koppel and Ordan); a high number of *hapax legomena* (in 2006 by Somers).

Some researchers, cited in Pataki (2012), argue that no regularities indicating MT within texts were revealed as a result of a series of experiments with German-English translation, which is one of the best qualities. Thus, they regard this solution as infeasible due to the randomness and variable nature of features.

3.2 CLPD Process Structure

The majority of authors attribute CLPD to the external PD approach, as in Meuschke and Gipp (2013), therefore, the same conventional detection steps, namely, candidate retrieval, detailed comparison and knowledge-based post-processing are distinguished and remain unchanged, as shown in the surveys by Potthast *et al.* (2011) and Barrón Cedeño *et al.* (2013). The standard plagiarism detection workflow is presented in Fig. 1.

3.3 Retrieval and Comparison

The candidate retrieval stage applies heuristics in order to reduce the search space (included topic/genre filtering of the potential source documents). Potthast *et al.* (2011) outlined three approaches: the first one implies query formulation on the basis of keywords extracted from the suspicious document and translated into the corresponding language (a CLIR solution); the next two approaches rely on the results of machine translation and make use of either standard keyword retrieval (an IR solution) or hash coding. Detailed comparison step includes measuring the similarity between suspicious text and the potential source documents resulting from the candidate retrieval step. The corresponding methods outlined in Potthast *et al.* (2011) are as follows: syntax-based (CL-CNG), dictionary-based (Eurovoc thesaurus-based, CL-VSM), parallel corpora based (CL-ASA, CL-LSI, CL-KCCA) and comparable corpora-based (CL-ESA). Some of them rely on the use of tools, containing language- and topic-specific information, e.g. dictionary based, parallel corpora-based, comparable corpora-based and some of them do not, such as syntax-based. In what follows a detailed explanation is provided for each one of the comparison models.

Syntax-Based Models

CL-CNG or Cross-Language Character N-Gram model uses overlapping character 4-gram tokenization on the basis of the Hopkins Automated Information Retriever for Combing Unstructured Text (HAIRCUT) system and was created by McNamee and Mayfield (2004). The key distinction of this approach lies in the possibility of comparing multilingual documents without translation. The best results were achieved for the languages sharing similar syntactic structure and international lexicon (e.g., related European language pairs).

The rest of the methods depends on the use of lexico-conceptual knowledge bases, corpora and dictionaries.

Dictionary-Based Models

CL-VSM (Cross-Language Vector Space Model) approach consists in constructing vector space models of the documents using indexed thesauri, dictionaries and other concept spaces. Eurovoc and corpora developed in the JRC(Joint Research Centre), e.g. JRC-Acquis Multilingual Parallel Corpus, presented in Steinberger (2012) link texts through the so called "language-independent anchors", multilingual pairs of words that denote entity names, locations, dates, measurement units etc.. In Gupta (2012) CL-CTS, Cross-Language Conceptual Thesaurus-Based Similarity method, is proposed, which is an algorithm that measures the similarity between texts written in different languages (English, German and Spanish in that particular case) on the basis of the domain-specific mapping presented in Eurovoc. An *ad-hoc* function defines whether a document belongs to some thesaurus concept *id*, represented by vector dimension in multidimensional vector space. The main advantage of this method lies in robustness to topic variance. In Pataki (2012) a dictionary-based language-independent approach is presented that consists of three main stages, namely, search space reduction, similarity estimation and filtering of results. Retrieval space is reduced by means of document pre-processing (fragmentation, stemming, elimination of stop-words), key words extraction and translation of their lemmas. It was estimated that the optimum number of translations equals to five. The main distinction of the present method lies in the use of an *ad-hoc* metric based on the minimum function, which allows to discard word number variance. Its purpose is to verify whether the compared documents are likely to be translations of one another. Post-processing step is rule-based and considers two thresholds for the obtained similarities. In order to reduce the computational cost of candidate retrieval and similarity analysis it was proposed in Pataki and Marosi (2012) to use SZTAKI desktop grid. It dynamically uploads and preprocesses information from the Wikipedia database and stores it to the KOPI system. Torrejón and Ramos (2011) presented a combination of *n*-gram and dictionary-based approach as an extension to "CoReMo" System developed earlier for external plagiarism detection purposes. ***Direct2stem*** and ***stem2stem*** dictionaries are integrated into the system and are based on Wiktionary and Wikipedia interlanguage links dictionaries. ***Direct2stem*** takes full words as entries and provides translations of the most frequent stems as output. ***Stem2stem*** gets activated in case the previous dictionary could not find any translation variant: original roots are taken as input in this case. If both dictionaries fail, the word gets stemmed by the English rules. CoReMo System's core rests on CTNG or *Contextual n-grams*, and RM, *Referential Monotony*. Contextual *n*-gram modelling is used to obtain the inverted index and uncover plagiarized fragments, which is performed by alphabetic ordering of overlapping 1-grams. Pre-processing includes case folding, elimination of stopwords, Porter stemming and internal sorting. Referential monotony is an algorithm that selects the longest sequences of text splits that indicate possible plagiarism and compares them to the whole source text. CoReMo system algorithm's advantages, as observed by the authors, are good runtime performance (obtaining of global results in 30 minutes), integrated dictionary and low computer requirements.

Comparable Corpora-Based Models

CL-ESA or Cross-Language Explicit Similarity Analysis, as reported in Potthast *et al.* (2011) represents approaches based on comparable corpora. According to Talvensaari (2008), as opposed to parallel corpora (CL-LSI, CL-KCCA and CL-ASA models), comparable corpora concept does not involve sentence-aligned translations. It is represented by topic-related texts with common vocabulary. Wikipedia encyclopedia and similar resources can serve as an example. These corpora are noisier, but at the same time more flexible. CL-ESA approach implies automatic creation of word associations for bilingual document representation in order to perform comparison of vocabulary correlation. As explained in Cimiano *et al.* (2009), concept space *C* is associated precisely to the article space in Wikipedia, therefore the approach is called ”explicit”. The association strength between the suspicious document and the concept space is evaluated by calculating the sum of the *tf-idf* values of the article for all words of the analysed text. Later, for cross-language retrieval purposes, the method was extended by the employment of Wikipedia language links to index the document with respect to the corresponding articles in any language.

Parallel Corpora-Based Models

CL-ASA or Cross-Language Alignment Similarity Analysis introduced by Barrón Cedeño *et al.* (2008) implies creation of bilingual statistical dictionary (core of CLiPA (Cross-Lingual Plagiarism Analysis) system) on the basis of parallel corpus being aligned using the well-known IBM Model 1. As observed in Ceska *et al.* (2008) word positions are taken into account. At the second step expectation maximization algorithm is applied in order to calculate statistical dictionary probabilities. The model was modified, as presented in Potthast *et al.* (2011): translation model probability $p(d_q/d')$ was changed to weight measure $w(d_q/d')$ and language model probability *p(d')* was substituted by a length model in order to apply it similarity analysis of full-scale documents of variable length.

CL-LSI or Cross-Language Latent Semantic Indexing also uses parallel corpora. It is a common strategy applied in IR systems for term-document association. It is ”latent” in the way that it extracts topic-related lexemes from the data itself and not from the external sources as opposed to CL-ESA. In Potthast *et al.* (2011) it is observed that CL-LSI is characterized by poor runtime performance due to the use of linear algebra technique, singular value decomposition of the original term-document matrix, as the core of the algorithm. According to Cimiano *et al.* (2009), concepts are latently contained in the columns of one of the orthogonal matrices (term-concept correlation weights) resulting from the main matrix decomposition.

CL-KCCA or Cross-Language Kernel Canonical Correlation Analysis performs much better than LSI on the same datasets, although it is based on SVD as well, according to Vinokourov *et al.* (2002). However, Potthast *et al.* (2011) observe that for the same reasons of runtime performance this approach cannot compete with CL-CNG and CL-ASA. As explained in Vinokourov *et al.* (2002), CL-KCCA analyses the correspondence of points in two embedding spaces that represent bilingual document pair and measures the correlation of the respective projection values. It provides detection of certain semantic similarities, represented by word sets with the same patterns of occurrence values for given bilingual document pairs.

One of more recent approaches named CL-KGA was not included into this classification. It can be considered both dictionary- and comparable corpora-based. It is described as follows. CL-KGA or Cross-Language Knowledge Graph Analysis, presented in Franco-Salvador *et al.* (2013), is substantially new in that it involves the use of the recently created multilingual semantic network BabelNet and graph-based text representation and comparison. In BabelNet, WordNet synsets and Wikipedia pages form concepts (nodes), meanwhile semantic pointers and hyperlinks constitute relations (edges) respectively, as explained in Navigli (2012). This structure enhances word-sense disambiguation and concept mapping of the analysed documents. However, any other knowledge base can be integrated into this system, as pointed out by the authors. Text fragmentation at the pre-processing step is performed using 5-sentence sliding window, grammatical categories are tagged with the TreeTagger tool. Similarity is measured basing on relation and concept weight values. CL-KGA, as observed by Franco-Salvador *et al.* (2013), refines the results of the other state-of-the-art approaches, according to plagdet evaluation results.

Barrón Cedeño *et al.* (2013) update this classification by adding the fifth model (MT-based) and attributing the whole set to the retrieval step, not the detailed comparison. Thus, as a result we have five families of retrieval models: lexicon-based, thesaurus-based, comparable corpus-based, parallel corpus-based and MT-based. Authors define them as *systems*. Lexicon-based systems (an amplified version syntax-based model class, presented in Potthast *et al.* (2011)) comprise the following techniques: *cognateness*, based on prefixes and other tokens; *dot-plot* model, based on character *n*-grams; CL-CNG (Cross-Language Character N-Grams). The rest of the models, except the MT-based one, are identical to those described in Potthast *et al.* (2011). MT-based model (or *T+MA*) involves determination of the suspicious document language with a language detector, translation and monolingual analysis. In Barrón Cedeño (2012) T+MA includes *web-based CL models* and *multiple translations*. The approach by Kent and Salim (2009 and 2010) belongs to the first type. They use Google Translate API to obtain English versions of texts that were originally written in Malay, with that the further pre-processing and comparison using three least-frequent four-grams fingerprint matching are performed. The approach by Muhr *et al.* (2010) is attributed to the second type. Instead of a full-scale automatic translation, they make use only of the main component of the corresponding systems: word alignment algorithm. German and Spanish texts form the corpus for the subsequent experiments. The words are aligned using BerkeleyAligner and 5 translation candidates are assigned on the basis of the Europarl corpus. As observed in Barrón Cedeño *et al.* (2013), T+MA proved its efficiency in PAN 2011, however, the same translation system (Google Translator) was used for generation and analysis. Therefore, an evaluation of T+MA performance using other translation systems was implemented.

3.4 Results of Performance Evaluation

In Potthast *et al.* (2011) the performance of CL-C3G (based on 3-grams), CL-ESA and CL-ASA was compared. Three experiments (cross-language ranking, bilingual rank correlation and cross-language similarity distribution) were carried out on the basis of two aligned corpora: comparable Wikipedia and parallel JRC-Acquis corpus (legal documents of the European Union aligned in 22 languages). Language pairs included English as the first language and Spanish, German, French, Dutch, or Polish as the second one. CL-C3G and CL-ESA show better results when suspicious and original documents share topic-specific information, whereas CL-ASA performs better with professional and automatic translations (due to the nature of the corpora used). CL-ASA and CL-ESA, as opposed to CL-CNG, can be applied for distant language pairs with alphabet and syntax unrelated, as pointed out in Barrón Cedeño (2012). CL-ESA, as compared to CL-ASA and CL-C3G, proved to be more a general purpose retrieval model, however, it depends much on the languages involved. CL-C3G outperformed the other approaches within the framework of these experiments.

In Barrón Cedeño (2012) the performance of CL-CNG, CL-ASA and CL-T+MA was compared. The author was interested in studying the behaviour of the models with respect to distant language pairs (Basque-English and Basque-Spanish). T+MA outperformed the other models, because it doesn't depend neither on corpora nor on syntactic/lexical similarities between languages. However, it is a computationally expensive method and there is still a lack of good automatic translators for most language pairs.

In Barrón Cedeño *et al.* (2013) another evaluation of CL-CNG, CL-ASA and CL-T+MA is presented, which is base on PAN-PC-11 corpus (Spanish-English). This is a standard corpus for plagiarism detection that allows for the analysis of plagiarism cases from exact copy to paraphrase and translation. Three experiments are carried out in order to assess the models performance with respect to precision and recall values. The respective scenarios are as follows. In Experiment A the suspicious document is an exact copy of a reference collection document. This experiment is designed to adjust the parameters of CL-ASA. In Experiment B the candidate and source are known and the aim is to detect plagiarized fragments. In Experiment C plagiarized fragments shall be retrieved from the noisy set of reference collection documents. According to the results of Experiment A, performance of the models depends on the document length: when considering an exact copy case, CL-CNG and T+MA work better with longer documents as opposed to CL-ASA (due to the use of length model). CL-CNG appears to outperform the other models in paraphrase uncovering. As to the results of Experiment B, T+MA shows the best recall in fragment detection, whereas CL-ASA provides the highest precision values, particularly in case of long texts (chunks have a fixed length of 5 sentences). Short plagiarism cases appear to be the hardest to detect. Within the framework of the Experiment C, CL-ASA provided better values of F-measure on short texts than T+MA model. Those obtained using CL-CNG, despite of not being influenced by the length and nature of plagiarism, turned out to be the worst ones. On the basis of the experiments performed authors conclude that T+MA and CL-CNG can be considered as recall-oriented systems and CL-ASA as a precision-oriented one.

4 Conclusions

The paper in hand outlines the existing approaches to translated plagiarism detection for the purposes of further research in the context of distant language pairs. The problem-oriented surveys by Potthast *et al.* (2011) and Barrón Cedeño *et al.* (2013) are summarized. It can be seen that the prototypical detection process remains unchanged: it includes heuristic retrieval, detailed comparison and knowledge-based filtering. Retrieval and comparison algorithms are being modified and knowledge bases are being expanded. CL-CNG was developed in 2004 and it is still one of the best-performing approaches that does not require the availability of any concept bases, such as dictionaries, thesauri, semantic networks or corpora, however it performs well only for languages sharing syntactic and lexical similarities (Indoeuropean families). All of the other analysis approaches depend on the availability of knowledge bases. In Torrejón and Ramos (2011) and Pataki (2012) *ad-hoc* dictionaries are used; Steinberger (2012) and Gupta (2012) describe the application of Eurovoc thesaurus; CL-ESA makes use of comparable corpora and such models as CL-ASA, CL-KCCA, CL-LSI require the availability of parallel corpora to properly perform the analysis; CL-KGA approach relies on the use of large semantic network BabelNet that combines WordNet synsets with Wikipedia articles, thus ensuring a more precise concept mapping. MT+A, according to the comparison by Barrón Cedeño *et al.* (2013), provides the best results, however, the translation of the whole reference collection is too costly and the corresponding translation services are far from being perfect, particularly for the cases of distant language pairs. Within the framework of the considered approaches, linguistic features are taken into account at the pre-processing step (lemmatization, case-folding, grammatical categories tagging etc.). Due to the variation in languages structures, their analysis is being avoided at the comparison step for the purposes of preserving runtime characteristics. The core analysis unit for the present methods is either character (CL-CNG) or word with the underlying concepts and connections.

References

Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. 2008. *On cross-lingual plagiarism analysis using a statistical model.* Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse, pp. 9-13. Patras, Greece.

Alberto Barrón-Cedeño. 2012. *On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism (Thesis).* Departmento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.

Alberto Barrón-Cedeño, Parth Gupta, and Paolo Rosso. 2013 *Methods for cross-language plagiarism detection.* Knowledge-Based Systems 50, 211-217.

Zdenek Ceska, Michal Toman, and Karel Jezek. 2008. *Multilingual Plagiarism Detection.* AIMSA 2008, LNAI 5253, pp. 83-92, 2008.

Philipp Cimiano, Antje Schultz, Sergey Sizov, Philipp Sorg, and Steffen Staab 2009. *Explicit Versus Latent Concept Models for Cross-Language Information Retrieval.* Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09).

Marc Franco-Salvador, Parth Gupta, and Paolo Rosso. 2013. *Cross-Language Plagiarism Detection Using a Multilingual Semantic Network.* IECIR 2013, LNCS 7814, pp. 710-713.

Parth Gupta, Alberto Barrón-Cedeño, and Paolo Rosso. 2012. *Cross-Language High Similarity Search Us-*

ing a Conceptual Thesaurus. ACLEF 2012, LNCS 7488, pp. 67-75, 2012.

Chow Kok Kent, and Naomie Salim. 2009. *Web Based Cross Language Plagiarism Detection*. Journal of Computing, Volume 1, Issue 1.

Chow Kok Kent, and Naomie Salim. 2010. *Web Based Cross Language Plagiarism Detection*. Second International Conference on Computational Intelligence, Modelling and Simulation, pages 199-204, IEEE.

Chung-Hong Lee, Chih-Hong Wu, and Hsin-Chang Yang. 2008. *A Platform Framework for Cross-lingual Text Relatedness Evaluation and Plagiarism Detection*. The 3rd International Conference on Innovative Computing Information and Control (ICICIC'08).

Paul McNamee, and James Mayfield. 2004. *Character N-Gram Tokenization for European Language Text Retrieval*. Information Retrieval,7,773-97.

Norman Meuschke, and Bela Gipp 2013. *State-of-the-art in detecting academic plagiarism*. International Journal for Educational Integrity Vol. 9 No.1, pp. 50-71 .

Markus Muhr, Roman Kern, Mario Zechner, and Michael Granitzer. 2010. *External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System*. Lab report for PAN at CLEF 2010.

Roberto Navigli 2012. *Babelplagiarism: What can BabelNet do for cross-language plagiarism detection?*. Keynotes for PAN 2012: Uncovering, Authorship, ad Social Software Misuse.

Máté Pataki. 2012. *A new approach for searching translated plagiarism*. Proceedings of the 5th International Plagiarism Conference. Newcastle, UK.

Máté Pataki, and Attila Csaba Marosi 2012. *Searching for Translated Plagiarism with the Help of Desktop Grids*. Journal of Grid Computing, 1-18.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. *Cross-language plagiarism detection*. Language Resources and Evaluation 45:45-62.

Ralf Steinberger 2012. *Cross-lingual similarity calculation for plagiarism detection and more - Tools and resources*. Keynotes for PAN 2012: Uncovering, Authorship, ad Social Software Misuse.

Tuomas Talvensaari. 2008. *Comparable Corpora in Cross-Language Information Retrieval (Academic Dissertation)*. Acta Electronica Universitatis Tamperensis 779.

Diego Antonio Rodríguez Torrejón, and José Manuel Martí Ramos. 2011. *Crosslingual CoReMo System*. Notebook for PAN at CLEF 2011.

Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. 2002. *Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis*. Advances of Neural Information Processing Systems 15.