Using Word Embedding for Cross-Language Plagiarism Detection

Jérémy Ferrero
Compilatio
276 rue du Mont Blanc
74540 Saint-Félix, France
LIG-GETALP
Univ. Grenoble Alpes, France
jeremy.ferrero@imag.fr

Frédéric Agnès
Compilatio
276 rue du Mont Blanc
74540 Saint-Félix, France
frederic@compilatio.net

Laurent Besacier
LIG-GETALP
Univ. Grenoble Alpes, France
laurent.besacier@imag.fr

Didier Schwab
LIG-GETALP
Univ. Grenoble Alpes, France
didier.schwab@imag.fr

Abstract

This paper proposes to use distributed representation of words (word embeddings) in cross-language textual similarity detection. The main contributions of this paper are the following: (a) we introduce new cross-language similarity detection methods based on distributed representation of words; (b) we combine the different methods proposed to verify their complementarity and finally obtain an overall F_1 score of 89.15% for English-French similarity detection at chunk level (88.5% at sentence level) on a very challenging corpus.

1 Introduction

Plagiarism is a very significant problem nowadays, specifically in higher education institutions. In monolingual context, this problem is rather well treated by several recent researches (Potthast et al., 2014). Nevertheless, the expansion of the Internet, which facilitates access to documents throughout the world and to increasingly efficient (freely available) machine translation tools, helps to spread *cross-language plagiarism*. Cross-language plagiarism means plagiarism by translation, *i.e.* a text has been plagiarized while being translated (manually or automatically). The challenge in detecting this kind of plagiarism is that the suspicious document is no longer in the same language of its source. We investigate how distributed representations of words can help to propose new cross-lingual similarity measures, helpful for plagiarism detection. We use word embeddings (Mikolov et al., 2013) that have shown promising performances for all kinds of NLP tasks, as shown in Upadhyay et al. (2016), Ammar et al. (2016) and Ghannay et al. (2016), for instance.

Contributions. The main contributions of this paper are the following:

- we augment some state-of-the-art methods with the use of word embeddings instead of lexical resources;
- we introduce a syntax weighting in distributed representations of sentences, and prove its usefulness for textual similarity detection;
- we combine our methods to verify their complementarity and finally obtain an overall F_1 score of 89.15% for English-French similarity detection at chunk level (88.5% at sentence level) on a very challenging corpus (mix of Wikipedia, conference papers, product reviews, Europarl and JRC) while the best method alone hardly reaches F_1 score higher than 50%.

2 Evaluation Conditions

2.1 Dataset

The reference dataset used during our study is the new dataset recently introduced by Ferrero et al.

(2016)[1]. The dataset was specially designed for a rigorous evaluation of cross-language textual similarity detection.

More precisely, the characteristics of the dataset are the following:

- it is multilingual: it contains French, English and Spanish texts;
- it proposes cross-language alignment information at different granularities: document level, sentence level and chunk level;
- it is based on both parallel and comparable corpora (mix of Wikipedia, conference papers, product reviews, Europarl and JRC);
- it contains both human and machine translated texts;
- it contains different percentages of named entities;
- part of it has been obfuscated (to make the cross-language similarity detection more complicated) while the rest remains without noise;
- the documents were written and translated by multiple types of authors (from average to professionals) and cover various fields.

In this paper, we only use the French and English sub-corpora.

2.2 Overview of State-of-the-Art Methods

Plagiarism is a statement that someone copied text deliberately without attribution, while these methods only detect textual similarities. However, textual similarity detection can be used to detect plagiarism.

The aim of cross-language textual similarity detection is to estimate if two textual units in different languages express the same or not. We quickly review below the state-of-the-art methods used in this paper, for more details, see Ferrero et al. (2016).

Cross-Language Character N-Gram (CL-CnG) is based on Mcnamee and Mayfield (2004) model. We use the Potthast et al. (2011) implementation which compares two textual units under their 3-grams vectors representation.

Cross-Language Conceptual Thesaurus-based Similarity (CL-CTS) (Pataki, 2012) aims to measure the semantic similarity using abstract concepts from words in textual units. In our implementation, these concepts are given by a linked lexical resource called *DBNary* (Sérasset, 2015).

Cross-Language Alignment-based Similarity Analysis (CL-ASA) aims to determinate how a textual unit is potentially the translation of another textual unit using bilingual unigram dictionary which contains translations pairs (and their probabilities) extracted from a parallel corpus (Barrón-Cedeño et al. (2008), Pinto et al. (2009)).

Cross-Language Explicit Semantic Analysis (CL-ESA) is based on the explicit semantic analysis model (Gabrilovich and Markovitch, 2007), which represents the meaning of a document by a vector based on concepts derived from Wikipedia. It was reused by Potthast et al. (2008) in the context of cross-language document retrieval.

Translation + Monolingual Analysis (T+MA) consists in translating the two units into the same language, in order to operate a monolingual comparison between them (Barrón-Cedeño, 2012). We use the Muhr et al. (2010) approach using *DBNary* (Sérasset, 2015), followed by monolingual matching based on bags of words.

2.3 Evaluation Protocol

We apply the same evaluation protocol as in Ferrero et al. (2016)'s paper. We build a distance matrix of size N x M, with M = 1,000 and $N = |S|$ where S is the evaluated sub-corpus. Each textual unit of S is compared to itself (to its corresponding unit in the target language, since this is cross-lingual similarity detection) and to M-1 other units randomly selected from S. The same unit may be selected several times. Then, a matching score for each comparison performed is obtained, leading to the distance matrix. Thresholding on the matrix is applied to find the threshold giving the best F_1 score. The F_1 score is the harmonic mean of precision and recall. Precision is defined as the proportion of relevant matches (similar cross-language units) retrieved among all the matches retrieved. Recall is the proportion of relevant matches retrieved among all the relevant matches to retrieve. Each method is applied on each EN-FR sub-corpus for chunk and sentence granularities. For each configuration (*i.e.* a particular method applied on a particular sub-corpus considering a particular granularity), 10 folds are carried out by changing the M selected units.

[1] `https://github.com/FerreroJeremy/Cross-Language-Dataset`

3 Proposed Methods

The main idea of word embeddings is that their representation is obtained according to the context (the words around it). The words are projected on a continuous space and those with similar context should be close in this multi-dimensional space. A similarity between two word vectors can be measured by cosine similarity. So using word-embeddings for plagiarism detection is appealing since they can be used to calculate similarity between sentences in the same or in two different languages (they capture intrinsically synonymy and morphological closeness). We use the *MultiVec* (Berard et al., 2016) toolkit for computing and managing the continuous representations of the texts. It includes word2vec (Mikolov et al., 2013), paragraph vector (Le and Mikolov, 2014) and bilingual distributed representations (Luong et al., 2015) features. The corpus used to build the vectors is the News Commentary[2] parallel corpus. For training our embeddings, we use CBOW model with a vector size of 100, a window size of 5, a negative sampling parameter of 5, and an alpha of 0.02.

3.1 Improving Textual Similarity Using Word Embeddings (*CL-CTS-WE* and *CL-WES*)

We introduce two new methods. First, we propose to replace the lexical resource used in *CL-CTS* (*i.e. DBNary*) by distributed representation of words. We call this new implementation *CL-CTS-WE*. More precisely, *CL-CTS-WE* uses the top 10 closest words in the embeddings model to build the BOW of a word. Secondly, we implement a more straightforward method (*CL-WES*), which performs a direct comparison between two sentences in different languages, through the use of word embeddings. It consists in a cosine similarity on distributed representations of the sentences, which are the summation of the embeddings vectors of each word of the sentences.

Let U a textual unit, the n words of the unit are represented by u_i as:

$$U = \{u_1, u_2, u_3, ..., u_n\} \tag{1}$$

If U_x and U_y are two textual units in two different languages, *CL-WES* builds their (bilingual) common representation vectors V_x and V_y and applies a cosine similarity between them.

A distributed representation V of a textual unit U is calculated as follows:

$$V = \sum_{i=1}^{n}(vector(u_i)) \tag{2}$$

where u_i is the i^{th} word of the textual unit and $vector$ is the function which gives the word embedding vector of a word. This feature is available in *MultiVec*[3] (Berard et al., 2016).

3.2 Cross-Language Word Embedding-based Syntax Similarity (CL-WESS)

Our next innovation is the improvement of *CL-WES* by introducing a *syntax flavour* in it. Let U a textual unit, the n words of the unit are represented by u_i as expressed in the formula (1). First, we syntactically tag U with a part-of-speech tagger (*TreeTagger* (Schmid, 1994)) and we normalize the tags with Universal Tagset of Petrov et al. (2012). Then, we assign a weight to each type of tag: this weight will be used to compute the final vector representation of the unit. Finally, we optimize the weights with the help of *Condor* (Berghen and Bersini, 2005). *Condor* applies a Newton's method with a trust region algorithm to determinate the weights that optimize the F_1 score. We use the first two folds of each sub-corpus to determinate the optimal weights.

The formula of the syntactic aggregation is:

$$V = \sum_{i=1}^{n}(weight(pos(u_i)).vector(u_i)) \tag{3}$$

where u_i is the i^{th} word of the textual unit, pos is the function which gives the universal part-of-speech tag of a word, $weight$ is the function which gives the weight of a part-of-speech, $vector$ is the function which gives the word embedding vector of a word and . is the scalar product.

If U_x and U_y are two textual units in two different languages, we build their representation vectors V_x and V_y following the formula (3) instead of (2), and apply a cosine similarity between them. We call this method *CL-WESS* and we have implemented it in *MultiVec* (Berard et al., 2016).

It is important to note that, contrarily to what is done in other tasks such as neural parsing (Chen

[2] http://www.statmt.org/wmt14/translation-task.html

[3] https://github.com/eske/multivec

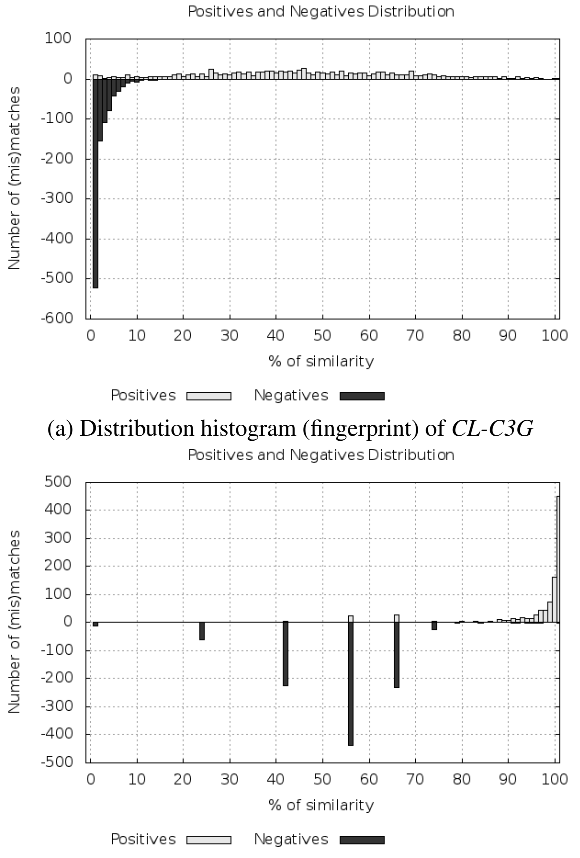
and Manning, 2014), we did not use POS information as an additional vector input because we considered it would be more useful to use it to weight the contribution of each word to the sentence representation, according to its morpho-syntactic category.

4 Combining multiple methods

4.1 Weighted Fusion

We try to combine our methods to improve cross-language similarity detection performance. During weighted fusion, we assign one weight to the similarity score of each method and we calculate a (weighted) composite score. We optimize the distribution of the weights with *Condor* (Berghen and Bersini, 2005). We use the first two folds of each sub-corpus to determinate the optimal weights, while the other eight folds evaluate the fusion. We also try an average fusion, *i.e.* a weighted fusion where all the weights are equal.

4.2 Decision Tree Fusion



(a) Distribution histogram (fingerprint) of *CL-C3G*

(b) Distribution histogram (fingerprint) of *CL-ASA*

Figure 1: Distribution histograms of two state-of-the-art methods for 1000 positives and 1000 negatives (mis)matches.

Regardless of their capacity to predict a (mis)match, an interesting feature of the methods is their clustering capacity, *i.e.* their ability to correctly separate the positives (similar units) and the negatives (different units) in order to minimize the doubts on the classification. Distribution histograms on Figure 1 highlight the fact that each method has its own fingerprint. Even if two methods look equivalent in term of final performance, their distribution can be different. One explanation is that the methods do not process on the same way. Some methods are lexical-syntax-based, others process by aligning concepts (more semantic) and still others capture context with word vectors. For instance, *CL-C3G* has a narrow distribution of negatives and a broad distribution for positives (Figure 1 (a)), whereas the opposite is true for *CL-ASA* (Figure 1 (b)). We try to exploit this complementarity using decision tree based fusion. We use the C4.5 algorithm (Quinlan, 1993) implemented in *Weka* 3.8.0 (Hall et al., 2009). The first two folds of each sub-corpus are used to determinate the optimal decision tree and the other eight folds to evaluate the fusion (same protocol as weighted fusion). While analyzing the trained decision tree, we see that *CL-C3G*, *CL-WESS* and *CL-CTS-WE* are the closest to the root. This confirms their relevance for similarity detection, as well as their complementarity.

5 Results and Discussion

Use of word embeddings. We can see in Table 1 that the use of distributed representation of words instead of lexical resources improves *CL-CTS* (*CL-CTS-WE* obtains overall performance gain of +3.83% on chunks and +3.19% on sentences). Despite this improvement, CL-CTS-WE remains less efficient than *CL-C3G*. While the use of bilingual sentence vector (*CL-WES*) is simple and elegant, its performance is lower than three state-of-the-art methods. However, its syntactically weighted version (*CL-WESS*) looks very promising and boosts the *CL-WES* overall performance by +11.78% on chunks and +14.92% on sentences. Thanks to this improvement, *CL-WESS* is significantly better than *CL-C3G* (+2.97% on chunks and +7.01% on sentences) and is the best single method evaluated so far on our corpus.

Fusion. Results of the decision tree fusion are reported at both chunk and sentence level in Table 1. Weighted and average fusion are only re-

Chunk level						
Methods	**Wikipedia (%)**	**TALN (%)**	**JRC (%)**	**APR (%)**	**Europarl (%)**	**Overall (%)**
CL-C3G	63.04 ± 0.867	40.80 ± 0.542	36.80 ± 0.842	80.69 ± 0.525	53.26 ± 0.639	50.76 ± 0.684
CL-CTS	58.05 ± 0.563	33.66 ± 0.411	30.15 ± 0.799	67.88 ± 0.959	45.31 ± 0.612	42.84 ± 0.682
CL-ASA	23.70 ± 0.617	23.24 ± 0.433	33.06 ± 1.007	26.34 ± 1.329	55.45 ± 0.748	47.32 ± 0.852
CL-ESA	64.86 ± 0.741	23.73 ± 0.675	13.91 ± 0.890	23.01 ± 0.834	13.98 ± 0.583	14.81 ± 0.681
T+MA	58.26 ± 0.832	38.90 ± 0.525	28.81 ± 0.565	73.25 ± 0.660	36.60 ± 1.277	37.12 ± 1.043
CL-CTS-WE	58.00 ± 1.679	38.04 ± 2.072	31.73 ± 0.875	73.13 ± 2.185	49.91 ± 2.194	46.67 ± 1.847
CL-WES	37.53 ± 1.317	21.70 ± 1.042	32.96 ± 2.351	39.14 ± 1.959	46.01 ± 1.640	41.95 ± 1.842
CL-WESS	52.68 ± 1.346	34.49 ± 0.906	45.00 ± 2.158	56.83 ± 2.124	57.06 ± 1.014	53.73 ± 1.387
Average fusion	81.34 ± 1.329	65.78 ± 1.470	61.87 ± 0.749	91.87 ± 0.452	79.77 ± 1.106	75.82 ± 0.972
Weighed fusion	84.61 ± 2.873	69.69 ± 1.660	67.02 ± 0.935	94.38 ± 0.502	83.74 ± 0.490	80.01 ± 0.623
Decision Tree	95.25 ± 1.761	74.10 ± 1.288	72.19 ± 1.437	97.05 ± 1.193	95.16 ± 1.149	89.15 ± 1.230

Sentence level						
Methods	**Wikipedia (%)**	**TALN (%)**	**JRC (%)**	**APR (%)**	**Europarl (%)**	**Overall (%)**
CL-C3G	48.24 ± 0.272	48.19 ± 0.520	36.85 ± 0.727	61.30 ± 0.567	52.70 ± 0.928	49.34 ± 0.864
CL-CTS	46.71 ± 0.388	38.93 ± 0.284	28.38 ± 0.464	51.43 ± 0.687	53.35 ± 0.643	47.50 ± 0.601
CL-ASA	27.68 ± 0.336	27.33 ± 0.306	34.78 ± 0.455	25.95 ± 0.604	36.73 ± 1.249	35.81 ± 1.036
CL-ESA	50.89 ± 0.902	14.41 ± 0.233	14.45 ± 0.380	14.18 ± 0.645	14.09 ± 0.583	14.44 ± 0.540
T+MA	50.39 ± 0.898	37.66 ± 0.365	32.31 ± 0.370	61.95 ± 0.706	37.70 ± 0.514	37.42 ± 0.490
CL-CTS-WE	47.26 ± 1.647	43.93 ± 1.881	31.63 ± 0.904	57.85 ± 1.921	56.39 ± 2.032	50.69 ± 1.767
CL-WES	28.48 ± 0.865	24.37 ± 0.720	33.99 ± 0.903	39.10 ± 0.863	44.06 ± 1.399	41.43 ± 1.262
CL-WESS	45.65 ± 2.100	40.45 ± 1.837	48.64 ± 1.328	58.08 ± 2.459	58.84 ± 1.769	56.35 ± 1.695
Decision Tree	80.45 ± 1.658	80.89 ± 0.944	72.70 ± 1.446	78.91 ± 1.005	94.04 ± 1.138	88.50 ± 1.207

Table 1: Average F_1 scores and confidence intervals of cross-language similarity detection methods applied on EN→FR sub-corpora – 8 folds validation.

ported at chunk level. In each case, we combine the 8 previously presented methods (the 5 state-of-the-art and the 3 new methods). Weighted fusion outperforms the state-of-the-art and the embedding-based methods in any case. Nevertheless, fusion based on a decision tree looks much more efficient. At chunk level, decision tree fusion leads to an overall F_1 score of 89.15% while the precedent best weighted fusion obtains 80.01% and the best single method only obtains 53.73%. The trend is the same at the sentence level where decision tree fusion largely overpasses any other method (88.50% against 56.35% for the best single method). In our evaluation, the best decision tree, for an overall higher than 85% of correct classification on both levels, involves at a minimum *CL-C3G*, *CL-WESS* and *CL-CTS-WE*. These results confirm that different methods proposed complement each other, and that embeddings are useful for cross-language textual similarity detection.

6 Conclusion and Perspectives

We have augmented several baseline approaches using word embeddings. The most promising approach is a cosine similarity on syntactically weighted distributed representation of sentence (*CL-WESS*), which beats in overall the precedent best state-of-the-art method. Finally, we have also demonstrated that all methods are complementary and their fusion significantly helps cross-language textual similarity detection performance. At chunk level, decision tree fusion leads to an overall F_1 score of 89.15% while the precedent best weighted fusion obtains 80.01% and the best single method only obtains 53.73%. The trend is the same at the sentence level where decision tree fusion largely overpasses any other method.

Our future short term goal is to work on the improvement of *CL-WESS* by analyzing the syntactic weights or even adapt them according to the plagiarist's stylometry. We have also made a submission at the SemEval-2017 Task 1, *i.e.* the task on Semantic Textual Similarity detection.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively Multilingual Word Embeddings. arXiv.org: http://arxiv.org/pdf/1602.01925v2.pdf. Computing Research Repository.

Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. 2008. On Cross-lingual Plagiarism Analysis using a Statistical Model. In Benno Stein and Efstathios Stamatatos and Moshe Koppel, editor, *Proceedings of the ECAI'08 PAN Workshop:*

Uncovering Plagiarism, Authorship and Social Software Misuse, pages 9–13, Patras, Greece.

Alberto Barrón-Cedeño. 2012. On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism. In *PhD thesis*, València, Spain.

Alexandre Berard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portoroz, Slovenia, May. European Language Resources Association (ELRA).

Frank Vanden Berghen and Hugues Bersini. 2005. CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, 181:157–175, September.

Danqi Chen and Christopher D. Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 740–750, Doha, Qatar.

Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, and Didier Schwab. 2016. A Multilingual, Multi-style and Multi-granularity Dataset for Cross-language Textual Similarity Detection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portoroz, Slovenia, May. European Language Resources Association (ELRA).

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, pages 1606–1611, Hyderabad, India, January. Morgan Kaufmann Publishers Inc.

Sahar Ghannay, Benoit Favre, Yannick Estève, and Nathalie Camelin. 2016. Word Embedding Evaluation and Combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portoroz, Slovenia, May. European Language Resources Association (ELRA).

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*, volume 11, pages 10–18, July.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning (ICML'14)*, volume 32, pages 1188–1196, Beijing, China, June. JMLR Proceedings.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st NAACL Workshop on Vector Space Modeling for Natural Language Processing*, Denver, Colorado, USA, May.

Paul Mcnamee and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. In *Information Retrieval Proceedings*, volume 7, pages 73–97. Kluwer Academic Publishers.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS'13)*, pages 3111–3119, Lake Tahoe, USA, December. .

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. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *CLEF Notebook*, Padua, Italy, September.

Màté Pataki. 2012. A New Approach for Searching Translated Plagiarism. In *Proceedings of the 5th International Plagiarism Conference*, Newcastle, UK.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

David Pinto, Jorge Civera, Alfons Juan, Paolo Rosso, and Alberto Barrón-Cedeño. 2009. A Statistical Approach to Crosslingual Natural Language Tasks. In *CEUR Workshop Proceedings*, volume 64 of *Journal of Algorithms*, pages 51–60, January.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A Wikipedia-Based Multilingual Retrieval Model. In *30th European Conference on IR Research (ECIR'08)*, volume 4956 of *LNCS of Lecture Notes in Computer Science*, pages 522–530, Glasgow, Scotland, March. Springer.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Cross-Language Plagiarism Detection. In *Language Resources and Evaluation*, volume 45, pages 45–62.

Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein. 2014. Overview of the 6th International Competition on Plagiarism Detection. In *PAN at CLEF 2014*, Sheffield, UK, September.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. In *Semantic Web Journal (special issue on Multilingual Linked Open Data)*, volume 6, pages 355–361.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, Berlin, Germany, August.