Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing*

Marc Franco-Salvador[1 2], Parth Gupta[1], and Paolo Rosso[1]

[1] Natural Language Engineering Lab - ELiRF, DSIC
Universitat Politècnica de València, Valencia, Spain
{mfranco,pgupta,prosso}@dsic.upv.es
[2] Linguistic Computing Laboratory (LCL)
Sapienza Università di Roma, Roma, Italy
francosalvador@it.uniroma1.it

Abstract. Cross-language plagiarism detection attempts to identify and extract automatically plagiarism among documents in different languages. Plagiarized fragments can be translated verbatim copies or may alter their structure to hide the copying, which is known as paraphrasing and is more difficult to detect. In order to improve the paraphrasing detection, we use a knowledge graph-based approach to obtain and compare context models of document fragments in different languages. Experimental results in German-English and Spanish-English cross-language plagiarism detection indicate that our knowledge graph-based approach offers a better performance compared to other state-of-the-art models.

Keywords: Cross-language plagiarism detection, textual similarity, paraphrasing, knowledge graphs, BabelNet.

1 Introduction

One of the biggest problems in literature and science is plagiarism: unauthorized use of the original content. Plagiarism is very difficult to detect, especially when the web is the source of information due to its size. The detection of plagiarism is even more difficult when it concerns documents written in different languages. Recently a survey was done on scholar practices and attitudes [2], also from a cross-language (CL) plagiarism perspective which manifests that CL plagiarism is a real problem: only 36.25% of students think that translating a text fragment and including it into their report is plagiarism.

Plagiarized fragments can be translated verbatim copies, or can be hidden by their authors altering its structure, which is known as paraphrasing. In the recent study on

* The research has been carried out in the framework of the European Commission WIQ-EI IRSES (no. 269180) and DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) projects as well as the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems. We thank Roberto Navigli for offering help to get familiar with the BabelNet API.

paraphrasing in plagiarism [1] it has been shown that paraphrase mechanisms make plagiarism detection more difficult. Moreover, this study also shows that lexical substitutions are the paraphrase mechanisms most used in plagiarism, shortening the plagiarized text. This may be used in future to develop more effective plagiarism detectors.

In recent years there have been a few approaches to CL similarity analysis that can be used for CL plagiarism detection. A simple, yet effective approach is the cross-language character n-gram (CL-CNG) model [9] which is based on the syntax of documents, which uses character n-grams, and offers remarkable performance for languages with syntactic similarities. Cross-language explicit semantic analysis (CL-ESA) [14] is a collection-relative retrieval model, which represents a document by its similarities to a collection of documents. These similarities in turn are computed with a monolingual retrieval model such as the vector space model. The cross-language alignment-based similarity analysis (CL-ASA) model [3, 2] is instead based on a statistical machine translation technology that combines probabilistic translations, using a statistical bilingual dictionary and similarity analysis. Finally, the cross-language conceptual thesaurus based similarity (CL-CTS) model [8] tries to measure the similarity between the documents in terms of shared concepts, using a conceptual thesaurus, and named entities among them. Some of these models have been compared in detecting CL plagiarism in [14]. CL-ASA and CL-CNG obtained the best results. Hence we compare our approach with them. CL setting of plagiarism detection has been also actively addressed in the PAN track[3] at the CLEF[4]. The most popular technique to handle CL plagiarism detection seems to be involving machine translation systems, where all the documents are translated to the language of comparison beforehand [15, 16]. However, this put forward a heavy dependence on availability of Machine Translation (MT) systems and its quality. Hence we propose and compare to CL plagiarism detection systems which do not depend on MT system.

We propose a new approach, named cross-language knowledge graphs analysis (CL-KGA), whose goal is to exploit explicit semantics for a better representation of the documents. CL-KGA provides a context model by generating knowledge graphs that expand and relate the original concepts from suspicious and source paragraphs. Finally, the similarity is measured in a semantic graph space. In this paper we investigate how knowledge graphs as context models can help in detecting CL plagiarism when paraphrasing is employed.

The rest of the paper is structured as follows. In Section 2 we describe the cross-language similarity retrieval models we compare CL-KGA with. In Section 3 we describe the BabelNet multilingual semantic network, i.e. the resource we use to build our knowledge graphs, which are explained in Section 4. The CL-KGA model is described in Section 5. In Section 6, we evaluate our approach using the German-English (DE-EN) and Spanish-English (ES-EN) CL plagiarism cases of the PAN-PC'11 corpus and compare our results with the CL-ASA and CL-CNG models, differentiating plagiarism cases between translated verbatim copies and paraphrase translations.

[3] http://pan.webis.de

[4] http://www.clef-initiative.eu

2 Cross-Language Similarity Estimation Models

In this Section we describe the two state-of-the-art CL similarity retrieval models, CL-CNG and CL-ASA that perform CL plagiarism detection and against we compare.

2.1 Cross-Language Character N-Grams

Cross-language character n-gram (CL-CNG) model have shown to improve the performance of CL information retrieval immensely for syntactically similar languages. This model typically uses character trigrams (CL-C3G) to compare documents in different languages [14].

Given a source document d written in a language L_1 and a suspicious document d' written in language L_2, the similarity $S(d, d')$ between the two documents is measured as follows:

$$S(d, d') = \frac{\vec{d} \cdot \vec{d'}}{|d| \cdot |d'|}, \tag{1}$$

where $\vec{d}$ and $\vec{d'}$ are the vectorial representation of documents d and d' into character n-gram space.

2.2 Cross-Language Alignment based Similarity Analysis

Cross-language alignment based similarity analysis (CL-ASA) model measures the similarity between two documents d and d', from two different languages L_1 and L_2 respectively, by aligning the documents at word level and determining the probability of d' being a translation of d. The similarity $S(d, d')$ between both documents is measured as in equation 2:

$$S(d, d') = l(d, d') * t(d|d'), \tag{2}$$

where $l(d, d')$ is the length factor defined in [17], which is used as normalization since two documents with the same content, in different languages do not have the same length. Moreover, $t(d|d')$ is the translation model defined in equation 3:

$$t(d|d') = \sum_{x \in d} \sum_{y \in d'} p(x, y), \tag{3}$$

where $p(x, y)$ is the probability of a word x from language L_1 being a translation of word y from L_2. These probabilities can be obtained using a bilingual statistical dictionary.

3 Multilingual Semantic Network

A multilingual semantic network (MSN) follows the structure of a traditional lexical knowledge base and accordingly, it consists of a weighted and labeled directed graph

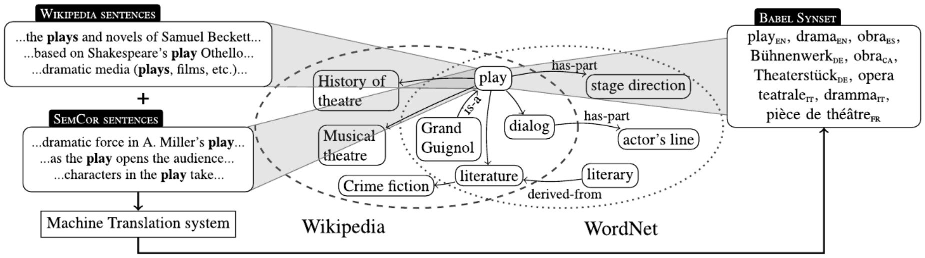
WIKIPEDIA SENTENCES
...the **plays** and novels of Samuel Beckett...
...based on Shakespeare's **play** Othello...
...dramatic media (**plays**, films, etc.)...
+
SEMCOR SENTENCES
...dramatic force in A. Miller's **play**...
...as the **play** opens the audience...
...characters in the **play** take...
Machine Translation system
History of theatre
play
has-part
stage direction
is-a
Musical theatre
Grand Guignol
dialog
has-part
actor's line
literature
literary
Crime fiction
derived-from
Wikipedia
WordNet
BABEL SYNSET
play$_{EN}$, drama$_{EN}$, obra$_{ES}$,
Bühnenwerk$_{DE}$, obra$_{CA}$,
Theaterstück$_{DE}$, opera
teatrale$_{IT}$, dramma$_{IT}$,
pièce de théâtre$_{FR}$

Fig. 1. Structure example of the BabelNet MSN [11].

where nodes represent the concepts and named entities while edges express the semantic relations between them. Each of the nodes contains a set of lexicalizations of the concept in different languages.

Although in this work we employ BabelNet [11], the graph-based approach we propose is generic and could be applied with other available MSNs such as EuroWordNet [21]. BabelNet is a very large MSN available in languages such as: Catalan, English, French, German, Italian and Spanish. Concepts and relations are taken from the largest available semantic lexicon of English, WordNet, and a wide-coverage collaboratively-edited encyclopedia, Wikipedia, which make BabelNet a multilingual "encyclopedic dictionary" that combines lexicographic information with wide-coverage encyclopedic knowledge. BabelNet's concept inventory consists of all WordNet's word senses and Wikipedia's encyclopedic entries, while its set of available relations comprises both semantic pointers between WordNet synsets, and semantically unspecified relations from Wikipedia's hyperlinked text. Multilingual lexicalizations for all concepts are collected from Wikipedia's inter-language links and WordNet's tagged senses in the SemCor corpus [10], using a machine translation system. A BabelNet's structure example is illustrated in Fig. 1.

BabelNet API[5] allows us to use it as a dictionary, statistical dictionary, word-sense disambiguation system and to build knowledge graphs.

4 Knowledge Graphs

A knowledge graph is a weighted and labelled graph that expand and relate the original concepts present in a set of words, providing us a "context model" of its content. Using MSN BabelNet to build the graphs, each one of them has a multilingual dimension of the concepts. Therefore, we can compare directly pairs of graphs built from document fragments in different languages and may be used to detect CL plagiarism.

We can build a knowledge graph using a MSN as follows: having a concept set C, we search the MSN for paths connecting each pair $c, c` \in C$, obtaining the set of paths P, where each $p \in P$ is a set of concepts and relations between concepts from C which include the conceptual expansion. The knowledge graph g is obtained after joining the paths from P including all its concepts and relations. Finally, to weight the

[5] http://babelnet.org/

concepts we use their degree of relateness, i.e. the number of outgoing edges for each node. The relation weighting is performed also in function of the degree of relateness of their source and target concepts.

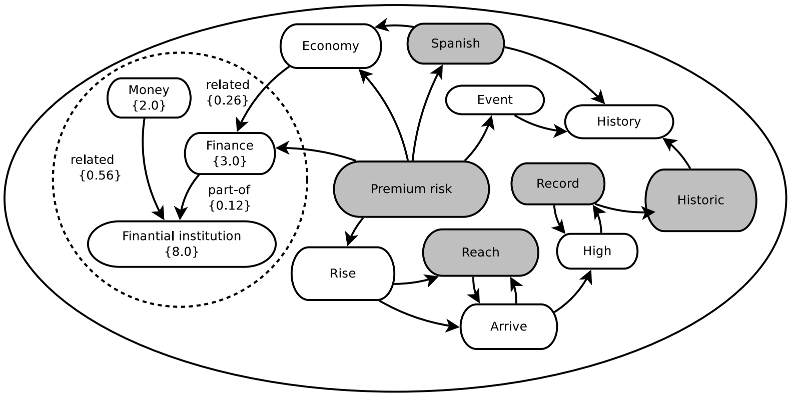

Fig. 2. Knowledge graph built from the sentence "Spanish premium risk reaches historic records", simplified without the multilingual dimension, and with labels and weights only inside the dashed circle.

Example. Having the English sentence "Spanish premium risk reaches historic records", we obtain its concepts $C = \{\text{Spanish, premium risk, reach, record, historic}\}$. Using BabelNet to build a knowledge graph g from C, we obtain a concept set $C_g = C \cup C'$, where $C' = \{\text{economy, finance, history...}\}$ is the expanded concept set. In addition, we obtain a relation set $R \in \{\text{related-to, has-part, belong-to, is-a...}\}$ between concepts of C_g. We can observe the resulting graph g in Fig. 2.

5 Cross-Language Knowledge Graphs Analysis

Our approach, cross-language knowledge graphs analysis (CL-KGA), presented previously in [5, 6], uses knowledge graphs generated from a MSN to obtain a context model of document fragments in different languages. The similarities between document fragments are computed in a semantic graph space.

Given a source document d and a suspicious document d', we compare document fragments in a four-step process:

1. We segment the original document in a set of fragments, using a 5-sentence sliding window with a 2-sentence step on the input document.
2. The paragraphs are lemmatized and tagged according to their grammatical category. For our experiments we use TreeTagger[6] [18], which supports multiple languages.
3. The knowledge graphs from the tagged fragments are built using the MSN.
4. We compare these graphs to measure similarity. The complete CL detection process using CL-KGA is shown in Fig. 3.

[6] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

To compare graphs we use a similarity function S for given graphs g and g' as shown in Eq. 4. It is an adapted version for MSN of flexible comparison of conceptual graphs similarity algorithm presented in [7].

$$S(g, g') = S_c(g, g') * (a + b * S_r(g, g')) \tag{4}$$

$$S_c(g, g') = \frac{\left(2 * \sum_{c \in g_{int}} w(c)\right)}{\left(\sum_{c \in g} w(c) + \sum_{c \in g'} w(c)\right)} \tag{5}$$

$$S_r(g, g') = \frac{\left(2 * \sum_{r \in N(c, g_{int})} w(r)\right)}{\left(\sum_{r \in N(c,g)} w(r) + \sum_{r \in N(c,g')} w(r)\right)} \tag{6}$$

where S_c is the score of the concepts, S_r is the score of the relations, a and b are smoothing variables to give the appropriate relevance to concepts and relations[7], c is a concept, r is a relation, g_{int} is the resulting graph of the intersection between g and g', and $N(c, g)$ is the set of all the relations connected to the concept c in a given graph g.

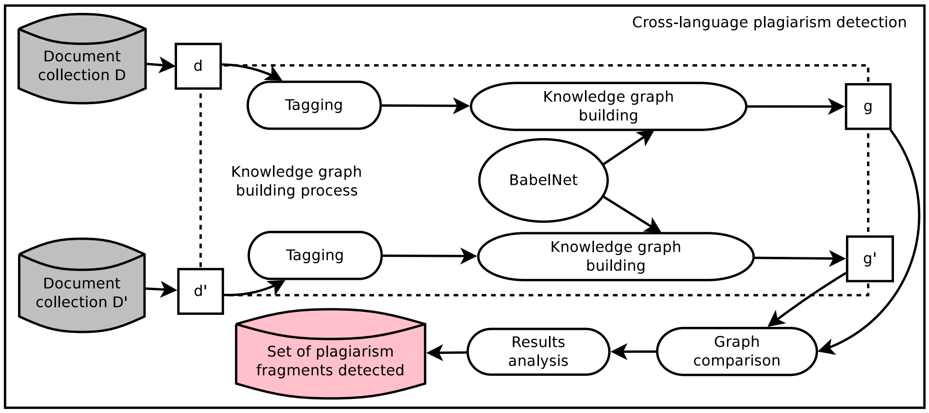

Fig. 3. CL plagiarism detection process between two sets of documents, D and D', in different languages.

6 Experiments and Evaluation

In this section we evaluate the performance of our approach, CL-KGA, for CL plagiarism detection, differentiating plagiarism cases between translated verbatim copies and

[7] In [6] we estimated the values of a and b for DE-EN and ES-EN using the MSN BabelNet.

paraphrase translations. We compare the results obtained by CL-KGA with those provided by CL-ASA and CL-C3G (CL-CNG using 3-grams) for the same task. For CL-ASA model we use two statistical dictionaries: BabelNet's statistical dictionary (CL-ASA$_{BN}$ [4]) and a dictionary trained using the word-aligment model IBM M1 [12] on the JRC-Acquis [20] corpus.

6.1 Corpus and Task Definition

Within automatic plagiarism detection scope, an international competition is celebrated annually since 2009, *Uncovering Plagiarism Authorship and Social Software Misuse*[8] (PAN), in which mono and CL plagiarism detection approaches are presented and tested. In our evaluation we use the CL plagiarism partition of PAN-PC'11[9] [15] corpus from its plagiarism task: given set of suspicious documents D in a language L_1, and their corresponding source documents D', in a language L_2, the task is to compare pairs of documents (d, d'), $d \in D$ and $d' \in D'$, to find all plagiarized fragments in D from D'. For this purpose we use a 5-sentence sliding window on the input documents to extract the fragments, and we analyze the similarities with the models listed above. Once we have the similarities between all the fragments, we use a detailed analysis and a post-processing method [19, 2] to determine the plagiarism cases.

As we can see in the corpus statistics of Table 1, PAN-PC'11 corpus has plagiarism cases generated in two different ways: automatic translations (verbatim copies) and automatic translations+manual correction (paraphrase translations). In our experiments we show the results on the two types of translated plagiarism separately.

ES-EN documents		DE-EN documents	
Suspicious	304	Suspicious	251
Source	202	Source	348
Plagiarism cases {es,de}-en			
Automatic translation			5,142
Automatic translation + Manual correction			433

Table 1. Statistics of PAN-PC'11 external cross-language plagiarism detection partition.

6.2 Measures

For the evaluation, we employ the measures introduced for the PAN competition on plagiarism detection: recall and precision at character level, in addition to granularity, which accounts for the fact that detectors sometimes report overlapping or multiple detections for a single plagiarism case. The three measures were integrated together in order to obtain a overall score for plagiarism detection (plagdet):

$$plagdet(S, R) = \frac{F1}{log_2(1 + granularity(S, R))},$$

[8] http://pan.webis.de/

[9] http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-pc-11.html

where S is the set of plagiarism cases in the corpus, R is the set of plagiarism cases reported by the detector, and $F1$ is the equally weighted harmonic mean of precision and recall. A more detailed description about the corpus and the measures can be found respectively in [13] and [15].

6.3 Results and Discussion

Model	German-English							
	Automatic translations				Paraphrase translations			
	Plagdet	Recall	Precision	Granularity	Plagdet	Recall	Precision	Granularity
CL-KGA	**0.5296**	**0.4671**	**0.6306**	**1.0188**	**0.1006**	**0.2101**	**0.0661**	**1.0**
CL-ASA$_{IBMM1}$	0.4230	0.3690	0.6019	1.1163	0.0462	0.0978	0.0303	1.0
CL-ASA$_{BN}$	0.3019	0.2363	0.5962	1.1753	0.0275	0.0796	0.0166	1.0
CL-C3G	0.0909	0.0564	0.3414	1.0913	0.0185	0.0389	0.0121	1.0

Table 2. DE-EN cross-language plagiarism detection results for automatic and paraphrase translation cases, displayed in the decreasing order of the Plagdet score.

As we can see in Table 2, for the DE-EN CL plagiarism detection, CL-C3G obtains the lowest results, being the baseline for this kind of experiments, due to the simplicity of the approach which uses n-grams. CL-ASA$_{BN}$ uses BabelNet's statistical dictionary. It obtained average results, despite many german words in the dictionary were not found. CL-ASA$_{IBMM1}$, one of the best state-of-the-art approaches for CL plagiarism detection, outperformed the baseline $plagdet$ by 365% in automatic translations and 149% in paraphrase translations. Finally, our novel approach CL-KGA, obtained the best values, surpassing the baseline $plagdet$ by 478% in automatic translations and 443% in paraphrase translations, along with better values for recall, precision and granularity.

Model	Spanish-English							
	Automatic translations				Paraphrase translations			
	Plagdet	Recall	Precision	Granularity	Plagdet	Recall	Precision	Granularity
CL-KGA	**0.6087**	**0.5399**	**0.7036**	**1.0050**	**0.0993**	**0.1979**	**0.0662**	**1.0**
CL-ASA$_{BN}$	0.5793	0.5245	0.6631	1.0154	0.0738	0.1909	0.0457	1.0
CL-ASA$_{IBMM1}$	0.5339	0.4728	0.6911	1.0729	0.0612	0.1501	0.0384	1.0
CL-C3G	0.1756	0.1336	0.6158	1.3796	0.0289	0.0587	0.0192	1.0

Table 3. ES-EN cross-language plagiarism detection results for automatic and paraphrase translation cases, displayed in the decreasing order of the Plagdet score.

As we can see in Table 3, for ES-EN CL plagiarism detection, the models performance was quite similar to DE-EN. CL-C3G is the baseline with the lowest values. CL-ASA$_{BN}$ increased the baseline $plagdet$ by 230% in automatic translations and 155% in paraphrase translations. This time CL-ASA$_{BN}$ obtain better results than CL-ASA$_{IBMM1}$ showing that using BabelNet's statistical dictionary for ES-EN plagiarism detection allowed to obtain a good performance. CL-KGA obtained the best values with all the measures, increasing the baseline $plagdet$ by 246% in automatic translations and 243% in paraphrase translations. The $granularity$ for CL-KGA is the closest to 1.0, the best possible value.

Notice that in both tables, values for paraphrase translation detections remain fairly low in general. All models benefit from the simplicity of the automatic translation cases, obtaining much higher values in all the values of $plagdet$, $recall$ and $precision$. The $precision$ values remain especially low and, looking at the statistics in Table 1, we can see that there are ten times more automatic than paraphrase cases, which may have influenced the false positive detection, with few cases in a large corpus in comparison. This fact explains the $granularity$ value of 1.0 in all the paraphrase detections: due to the small number of paraphrase cases, all the plagiarism cases detected are isolated, making impossible overlappings between detections. Despite the low values, CL-KGA obtained the best performance in detecting paraphrase too, increasing CL-ASA $plagdet$ by 34% in ES-EN and by 118% in DE-EN, which highlights its potential for DE-EN.

All these results exhibit the accuracy of the approach CL-KGA in identifying CL plagiarism. The model benefits from the context model obtained through MSN to measure CL similarity. This provides a tighter bound in estimation and leads to better results. We point out that the knowledge graph construction used in CL-KGA is more time-consuming compared to the other two models and, if time is the priority, the fastest approach is CL-ASA.

7 Conclusions and Future Work

In this study we have shown the good performance and potential of knowledge graphs to detect CL plagiarism even when paraphrasing is employed. CL-ASA using BabelNet's statistical dictionary also has shown his potential for ES-EN plagiarism detection. CL-KGA model obtained better results than CL-ASA and CL-CNG in detecting verbatim copies and paraphrase on the DE-EN and ES-EN CL plagiarism cases of the PAN-PC11 corpus. Nevertheless, experimental results indicate that automatic translations are much easier to detect than translations with paraphrasing. There are many aspects to be improved in order to make plagiarism detectors efficient in the CL task.

In the future we will investigate how the task of CL plagiarism detection can be approached using other MSNs. Moreover, we would like to investigate the knowledge graph suitability for CL information retrieval.

References

1. Barrón-Cedeño, A., Vila, M., Martí, M., Rosso, P.: Plagiarism meets paraphrasing: insights for the next generation in automatic plagiarism detection. In: Computational Linguistics 39(4) (2013)
2. Barrón-Cedeño, A.: On the mono- and cross-language detection of text re-use and plagiarism. Ph.D. thesis, Universitat Politènica de València (2012)
3. Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A.: On cross-lingual plagiarism analysis using a statistical model. In: Proc. of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. PAN'08 (2008)

4. Franco-Salvador, M., Gupta, P., Rosso, P.: Cross-language plagiarism detection using BabelNet's statistical dictionary. Computación y Sistemas, Revista Iberoamericana de Computación 16(4), 383–390 (2012)
5. Franco-Salvador, M., Gupta, P., Rosso, P.: Cross-language plagiarism detection using a multilingual semantic network. In: Proc. of the 35th European Conference on Information Retrieval (ECIR'13). vol. LNCS(7814), pp. 710–713. Springer-Verlag (2013)
6. Franco-Salvador, M., Gupta, P., Rosso, P.: Graph-based similarity analysis: a new approach to cross-language plagiarism detection. Journal of the Spanish Society of Natural Language Processing (Sociedad Espaola de Procesamiento del Languaje Natural) num. 50 (2013)
7. Montes y Gómez, M., Gelbukh, A.F., López-López, A., Baeza-Yates, R.A.: Flexible comparison of conceptual graphs. In: Proc. DEXA. pp. 102–111 (2001)
8. Gupta, P., Barrón-Cedeño, A., Rosso, P.: Cross-language high similarity search using a conceptual thesaurus. In: Proc. 3rd Int. Conf. of CLEF Initiative on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics. CLEF 2012 (2012)
9. Mcnamee, P., Mayfield, J.: Character n-gram tokenization for European language text retrieval. Information Retrieval 7(1), 73–97 (2004)
10. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Proceedings of the workshop on Human Language Technology. pp. 303–308. HLT '93, Association for Computational Linguistics, Stroudsburg, PA, USA (1993)
11. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
12. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
13. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: An evaluation framework for plagiarism detection. In: Proc. of the 23rd Int. Conf. on Computational Linguistics. pp. 997–1005. COLING-2010, Beijing, China (2010)
14. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-language plagiarism detection. Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis 45(1), 45–62 (2011)
15. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd int. competition on plagiarism detection. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
16. Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., et al.: Overview of the 4th international competition on plagiarism detection. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
17. Pouliquen, B., Steinberger, R., Ignat, C.: Automatic linking of similar texts across languages. In: Proc. Recent Advances in Natural Language Processing III. pp. 307–316. RANLP'03 (2003)
18. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proc. Int. Conf. on new methods in language processing (1994)
19. Stein, B., zu Eissen, S.M., Potthast, M.: Strategies for retrieving plagiarized documents. In: Proc. of the 30th annual Int. ACM SIGIR Conf. on Research and development in information retrieval. pp. 825–826. ACM (2007)
20. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The jrc-acquis: A multilingual aligned parallel corpus with +20 languages. In: Proc. 5th Int. Conf. on language resources and evaluation. LREC'2006 (2006)
21. Vossen, P.: Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. In: Proc. Int. Journal of Lexicography. vol. 17 (2004)