Cross-Language Plagiarism Detection using a Multilingual Semantic Network*

Marc Franco-Salvador, Parth Gupta, and Paolo Rosso

Natural Language Engineering Lab - ELiRF, DSIC
Universitat Politècnica de València, Valencia, Spain
{mfranco,pgupta,prosso}@dsic.upv.es

Abstract. Cross-language plagiarism refers to the type of plagiarism where the source and suspicious documents are in different languages. Plagiarism detection across languages is still in its infancy state. In this article, we propose a new graph-based approach that uses a multilingual semantic network to compare document paragraphs in different languages. In order to investigate the proposed approach, we used the German-English and Spanish-English cross-language plagiarism cases of the PAN-PC'11 corpus. We compare the obtained results with two state-of-the-art models. Experimental results indicate that our graph-based approach is a good alternative for cross-language plagiarism detection.

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 is among documents written in different languages. Recently a survey was done on scholar practices and attitudes [1], 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.

In recent years there have been a few approaches to CL plagiarism detection that go beyond translating the source document to the language of the suspicious document, and performing monolingual plagiarism analysis. Cross-language character n-gram (CL-CNG) model [4] is a model 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) [7] is a collection-relative retrieval model, which means that a document is represented 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. Cross-language alignment-based similarity analysis (CL-ASA) model [2, 1] is based on a statistical machine translation technology that

* We thank the *Consellería D′educació, Formació i Ocupació* of the *Generalitat Valenciana* for funding the work of the first author with the *Gerónimo Forteza* program. The research has been carried out in the framework of the European Commission WIQ-EI IRSES project (no. 269180) and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

combines probabilistic translation, using a statistical bilingual dictionary and similarity analysis. The three models have been compared in [7]. CL-ASA and CL-CNG produced the best results. Hence we compare our approach with them.

Our new approach, named cross-language knowledge graphs analysis (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.

2 Multilingual Semantic Network

A multilingual semantic network (MSN) follows the structure of a traditional lexical knowledge base and accordingly, it consists of a labeled directed graph where nodes represent the concepts and named entities while edges express the semantic relations between them. Each of its nodes contain a set of lexicalizations of the concept in different languages. In this work we employ MSN to build knowledge graphs to obtain a multilingual context model from document fragments and compare them to detect CL plagiarism.

Although in this work we employ BabelNet [6], the graph-based approach we propose is generic and could be applied with other available multilingual semantic networks such as ConceptNet [3] or EuroWordNet[1]. BabelNet is a very large multilingual semantic network 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 inventory concepts consist 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 SemCor corpora, using a machine translation system. BabelNet API[2] allows us to use it as a dictionary, statistical dictionary, word-sense disambiguation and to building knowledge graphs.

3 Graph based Similarity Analysis

Given a source document d and a suspicious document d', to compare document fragments we have a four-step process.

1. We segment the original document in a set of paragraphs, using a 5-sentence sliding window on the input document.
2. The paragraphs are tagged according to their grammatical category and use the terms in their infinitive form. For our experiments we use TreeTagger[3], which supports multiple languages.

[1] http://www.illc.uva.nl/EuroWordNet/

[2] http://lcl.uniroma1.it/babelnet/

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

3. The knowledge graphs from the tagged paragraphs are prepated. A knowledge graph is a weighted and labelled graph with the concepts present in the document. In addition we add the neighbour concepts from the MSN with labelled links according to their relations. Using knowledge graphs, we expand the original vocabulary obtaining a context model from the input document.
4. We compare these graphs to measure similarity.

To compare graphs we use similarity function S for given graphs G_1 and G_2 as shown in Eq. 1. It is an adapted version of conceptual graphs similarity algorithm presented in [5] for MSN to allow flexible comparison.

$$S(G_1, G_2) = S_c * (a + b * S_r) \tag{1}$$

$$S_c = \left(2 * \sum_{c \in O} weight(c)\right) / \left(\sum_{c \in G_1} weight(c) + \sum_{c \in G_2} weight(c)\right) \tag{2}$$

$$S_r = \left(2 * \sum_{r \in N(c,O)} weight(r)\right) / \left(\sum_{r \in N(c,G_1)} weight(r) + \sum_{r \in N(c,G_2)} weight(r)\right) \tag{3}$$

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, c is a concept, r is a relation, O is the resulting graph of the intersection between $G1$ and $G2$, and $N(c, G)$ is the set of all the relations connected to the concept c in a given graph G.

After a graph intersection, the relation weights in O must be updated according to the value of the concepts that form it. The value of a concept is measured as the number of its outgoing relations. We also have to re-estimate the relation weights from graph O taking into account the original value of the concepts in the graphs $G1$ and $G2$. For this purpose we propose the following algorithm:

```
for each concept $c$ from $O$ do
  dif($c, G_1$) = number_of_outgoing_edges($c, G_1$) / number_of_outgoing_edges($c, O$)
  dif($c, G_2$) = number_of_outgoing_edges($c, G_2$) / number_of_outgoing_edges($c, O$)
  for each outgoing edge $X$ of concept $c$ do
    weight($c, X, O$) = (weight($c, X, G_1$) $*$ dif($c, G_1$) $+$ weight($c, X, G_2$) $*$ dif($c, G_2$)) / 2
```

Fig. 1. Graph relation scores re-estimation algorithm

4 Evaluation

We use the cross-language plagiarism partition of PAN-PC'11 where for given set of suspicious documents D and their corresponding source documents D', 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'. 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[4].

As we can see in Table 1, for the DE-EN CL plagiarism detection, our novel approach increased the *plagdet* by 26.73% with respect to CL-ASA along with better

[4] Space constraints do not allow for describing corpus and measures. A more detailed description about the corpus and the measures can be found in the PAN-PC'11 overview [8]

	DE-EN				ES-EN			
Model	Plagdet	Recall	Precision	Granularity	Plagdet	Recall	Precision	Granularity
CL-KGA	**0.5144**	**0.4433**	**0.6319**	**1.0179**	**0.5943**	**0.5183**	**0.7059**	**1.0080**
CL-ASA	0.4059	0.3438	0.6039	1.1132	0.5170	0.4480	0.6891	1.0709
CL-C3G	0.0778	0.0473	0.3302	1.0896	0.1700	0.1278	0.6168	1.3721

Table 1. DE-EN and ES-EN cross-language plagiarism detection results

values for recall, precision and granularity. Similar behaviour is noticed for ES-EN pair too where the $plagdet$ score is increased by 14.95% compared to CL-ASA. These results exhibit the accuracy of the proposed algorithm in identifying CL plagiarism. The proposed model benefit from the context model obtained througn MSN to measure CL similarity which provides tighter bound in estimation and leads to better result. The graph construction is much computation costly compared to other two models.

5 Conclusion and Future Work

We described the necessary steps to use effectively a MSN such as BabelNet in order to detect cross-language plagiarism in documents. The proposed CL-KGA model obtained better results than CL-ASA and CL-CNG on the DE-EN and ES-EN cross-language plagiarism cases of the PAN-PC'11 corpus. It is important to point out that our approach is generic, and can be applied to other available MSNs such as ConceptNet or EuroWordNet in order to support more languages. In future we would like to investigate its suitability for cross-language information retrieval.

References

1. Barrón-Cedeño, A.: On the mono- and cross-language detection of text re-use and plagiarism. Ph.D. thesis, Universitat Politènica de València (2012)
2. Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A.: On cross-lingual plagiarism analysis using a statistical model. In: Proceedings of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. PAN'08 (2008)
3. Havasi, C.: Conceptnet 3: A flexible, multilingual semantic network for common sense knowledge. In: the 22nd Conference on Artificial Intelligence (2007)
4. Mcnamee, P., Mayfield, J.: Character n-gram tokenization for European language text retrieval. Inf. Retr. 7(1-2), 73–97 (2004)
5. Montes-Gómez, M., Gelbukh, A.F., López-López, A., Baeza-Yates, R.A.: Flexible comparison of conceptual graphs. In: DEXA. pp. 102–111 (2001)
6. Navigli, R., Ponzetto, S.P.: Babelnet: building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 216–225. ACL '10, Stroudsburg, PA, USA (2010)
7. 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) (2011)
8. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd international competition on plagiarism detection. In: CLEF (Notebook Papers/Labs/Workshop) (2011)