Monolingual and Crosslingual Plagiarism Detection

Towards the Competition @ SEPLN09 *

Alberto Barrón-Cedeño and Paolo Rosso

Natural Language Engineering Lab, RFIA,
Departmento de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia
{*lbarron, prosso*}*@dsic.upv.es*

Abstract. Automatic plagiarism detection considering a reference corpus compares a suspicious text to a set of documents in order to relate the plagiarised fragments to their potential source. The suspicious and source documents can be written wether in the same language (monolingual) or in different languages (crosslingual).
In the context of the Ph. D., our work has been focused on both monolingual and crosslingual plagiarism detection. The monolingual approach is based on a search space reduction process followed by an exhaustive word n-grams comparison. Surprisingly it seems that the application of the reduction process has not been explored in this task previously. The crosslingual one is based on the well known IBM-1 alignment model. Having a competition on these topics will make our work available to the Spanish scientific community interested in plagiarism detection.

1 Introduction

The easy access to a wide range of information in multiple languages via electronic resources has favoured the increase of text plagiarism cases of both kinds: monolingual and crosslingual. To plagiarise means to use text written by other people (even adapting it by rewording, insertion or deletion) without credit or citation. From a crosslingual perspective, a text fragment in one language is considered a plagiarism of a text in another language if their contents are considered semantically similar no matter they are written in different languages.

In order to get enough evidence to prove if a text is plagiarised, it is necessary to find its potential source. The objective of plagiarism detection with reference is to give this evidence. This is carried out by searching for the potential source of a suspicious text fragment from a set of reference texts.

Few works have been made from a crosslingual point of view. The first one is based on explicit semantic analysis, where two comparable corpora (one on each implied language) are exploited in order to define how semantically closed two

* We would like to thank the MCyT TEXT-MESS CICYT TIN2006-15265-C06-04 research project as well as the TIMM network CICYT TIN2005-25825-E.

documents are [11]. The second one is based on statistical bilingual models [4, 9] (Section 3). Note that no translation process is carried out in both approaches.

With the aim of bringing together to the researchers interested in these topics, we plan to carry out a competition which will be held in the context of the proposed PAN Satellite Workshop of the SEPLN'09 conference.

2 Monolingual Plagiarism Detection

An important factor in the plagiarism detection with reference is precisely the reference corpus. The best available method would be useless if the source of a plagiarised text is not included into the reference corpus D. Due to this reason, reference corpora are composed of a huge set of potential source documents.

Comparing a suspicious text s to all the reference documents $d \in D$ is practically impossible. Our proposed method carries out a preliminary reduction process, based on the Kulback-Leibler distance, selecting only those documents d with a high probability of being the source of s [3, 1]. Each probability distribution P_d is compared to the probability distribution P_s. The ten most similar reference documents are considered as candidates of being the source of the potentially plagiarised sentences in s. This is the reduced reference set D'.

The following objective is to answer the question "*Is a sentence $s_i \in s$ plagiarised from a document $d \in D'$?*". Due to the fact that plagiarised text fragments use to be rewritten from their source, a rigid search strategy does not give good results. Our flexible search strategy is based on a word n-grams comparison [2]. We consider n-grams due to the fact that independent texts have a small amount of common word n-grams (considering $n \geq 2$).

Our approach is based on the comparison of suspicious sentences and reference documents. We do not split the reference documents into sentences due to the fact that a plagiarised sentence could be made of fragments from multiple parts of a source document. The basic schema is as following: (1) s is split into sentences (s_i); (2) s_i is split into word n-grams, resulting in the set $N(s_i)$; (3) $d \in D'$ is not split into sentences, but simply into word n-grams, resulting in the set $N(d)$; and (4) $N(s_i)$ is compared to $N(d)$. Due to the difference in the size of $N(s_i)$ and $N(d)$, an asymmetric comparison is carried out on the basis of the *containment* measure [7]:

$$C(s_i \mid d) = \frac{|N(s_i) \cap N(d)|}{|N(s_i)|} \tag{1}$$

If the maximum $C(s_i \mid d)$, after considering every $d \in D'$, is greater than a given threshold, s_i becomes a candidate of being plagiarised from d.

For our experiments we have used the *METER corpus* [6]. This corpus is not a real plagiarism corpus. It is composed of a set of journalistic notes and was originally created in order to analyse the reuse of information in the British newspapers. The interesting fact about this corpus is that the text of a set of newspaper (suspicious) notes is identified as *verbatim*, *rewrite* or *new*, for exact copy, rewritten or nothing to do with the Press Association (reference) notes.

Our experiments show that the best results for the exhaustive comparison are obtained by considering bigrams and trigrams. In both cases, the word n-grams are short enough to handle modifications in the plagiarised sentences and long enough to compose strings with a low probability of appearing in any (but the plagiarism source) text. Trigram based search is more rigid, resulting in a better Precision. Bigram based search is more flexible, allowing better Recall. The search space reduction process improves the obtained F-measure (from 0.68 to 0.75 for bigrams) and the time it takes to analyse a suspicious document is reduced (from 2.32 to only 0.19 seconds in average).

3 Crosslingual Plagiarism Detection

Given the suspicious and reference texts x and y (written in different languages), the objective is to answer the question "*Is x plagiarised (and translated) from y?*". In some way, crosslingual plagiarism analysis is related to crosslingual information retrieval [10]. In fact, the aim is to retrieve those fragments that have been plagiarised in a language with respect to the one originally employed.

In our current research [4, 9] we have composed a minicorpus of original-plagiarised text pairs. The original fragments (y), in English, were extracted from a set of documents on Information Retrieval written by one only author. Around ten plagiarised versions of each fragment y have been obtained in Spanish and Italian (x). Each fragment x has been created by a different "human plagiariser" or automatic machine translator.

The set of y-x pairs was divided into training and test subsets. The training subset was used in order to compose a statistical bilingual dictionary. This dictionary was created on the basis of the IBM-1 alignment model [5], commonly used in statistical machine translation. The test set was only composed of the suspicious fragments from the test pairs. In order to obtain a realistic experiment, text fragments originally written in Spanish (and Italian) were added.

The objective of our experiment was to know if a suspicious fragment x was a plagiarism case from one of our reference fragments y. In order to determine if x is plagiarised from any y fragment, we compute the probability $p(y \mid x)$ of each fragment y given x. This probability is calculated as in Eq. 2.

$$p(y \mid x) = \frac{1}{(|x| + 1)^{|y|}} \prod_{i=1}^{|y|} \sum_{j=0}^{|x|} p(y_i \mid x_j) \tag{2}$$

where $p(y_i \mid x_j)$ is simply calculated on the basis of the statistical bilingual dictionary previously obtained and $| \cdot |$ is the length of $\cdot$ in words.

Our proposal calculates the probabilistic association between two terms in two different languages. After considering this probability, we are able to determine how likely is that a fragment x is a translation (plagiarism) from y. If the maximum $p(y \mid x)$ after considering every reference fragment y is higher than a given threshold, we consider that x is plagiarised from y.

The results obtained up to now with this method are promising. The application of a statistical machine translation technique, has demonstrated to be a valuable resource for the crosslingual plagiarism analysis. Due to the fact that we determine the similarity between suspicious and original text fragments on the basis of a dictionary, the word order is not relevant and we are able to find good candidates even when the plagiarised text has been modified.

4 Current and Future Work

Currently, we are creating corpora containing both kinds of plagiarism cases. These corpora will be used during the proposed competition as well as for our own research work. We plan also to tackle the problem of plagiarism of ideas "...in which an original thought from another is used but without any dependence on the words or form of the source..." [8]. This is a more general (and hard to detect) case of plagiarism.

References

1. Barrón-Cedeño, A. 2008. Detección automática de plagio en texto. Master's thesis, Universidad Politécnica de Valencia.
2. Barrón-Cedeño, A., Rosso, P.: On Automatic Plagiarism Detection based on n-grams Comparison. In: ECIR. LNCS, *in press* (2009)
3. Barrón-Cedeño, A., Rosso, P., Benedí, J.M.: Reducing the Plagiarism Detection Search Space on the basis of the Kullback-Leibler Distance. IN: CICLing 2008. LNCS, *in press* (2009)
4. Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A.: On Crosslingual Plagiarism Analysis Using a Statistical Model. In: ECAI'08 PAN Workshop Uncovering Plagiarism, Authorship and Social Software Misuse, pp. 9–13, Patras, Greece (2008)
5. Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., Roossin, P.: A Statistical Approach to Machine Translation. Computational Linguistics, 16(2), 79-85 (1990)
6. Clough, P., Gaizauskas, R., Piao, S.: Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In: 3rd International Conference on Language Resources and Evaluation (LREC-02), vol. V, pp. 1678–1691. Las Palmas, Spain (2002)
7. Lyon, C., Malcolm, J., Dickerson, B.: Detecting Short Passages of Similar Text in Large Document Collections. In: Conference on Empirical Methods in Natural Language Processing, pp. 118–125. Pennsylvania (2001)
8. Martin, B.: Plagiarism: a Misplaced Emphasis. Information Ethics. 3(2) 36-47 (1994)
9. Pinto, D., Civera, J., Juan, A., Rosso, P., Barrón-Cedeño, A.: A Statistical Approach to Crosslingual Natural Language Tasks. In: Fourth Latin American Workshop on Non-Monotonic Reasoning, Puebla, Mexico (2008)
10. Pinto, D., Juan, A., Rosso, P.: Using Query-Relevant Documents Pairs for Cross-Lingual Information Retrieval. In: TSD 2007. LNAI 4629, pp. 630–637 (2007)
11. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-Based Multilingual Retrieval Model. In: ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer-Verlag (2008)