Cross Lingual Text Reuse using machine translation & similarity measures

Nitish Aggarwal[1], Kartik Asooja[2] & Paul Buitelaar[1]

[1]Unit for Natural Language Processing, DERI, National Univeristy of Ireland Galway

[2]Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

Abstract

In this article, we briefly describe our approach in detecting cross-lingual text reuse. We are given a set of suspicious (possibly reused) files and the source files. The objective is to detect the corresponding source files for the reused ones. We handled the cross-lingual problem with the help of machine translation. Then we applied similarity measures with the help of Apache Lucene, an open-source information retrieval library. Along with the Lucene scoring, we tried to match the order of the sentences or events, but this did not result in any observed improvement.

Introduction

The advent of the Web and its subsequent expansion with services such as online machine translation systems has provided the general public with access to enormous volumes of information across language barriers. Unfortunately this has also increased the occurrence of cross-lingual text plagiarism as it is very easy to find related texts on a topic. Cross-Lingual Text Reuse (CLITR) is the scenario where text is reused in a language that is different from the language of the source text [1]. The CLITR scenario can be said to be composed of two steps: i) identify relevant text in the source language; ii) customize and translate it into the target language (the language in which the plagiarism is done). The amount of customization gives some idea of the intensity of the text reuse. Therefore, it can be concluded that the important part of text reuse lies in customizing the text that is being reused.

Due to the large size of resources available for text reuse on the Web, it is not feasible for a human plagiarism detection expert to validate all of it. Therefore, it is necessary to develop an automatic process that can efficiently detect text reuse and reduce the burden of validating it. Moreover, in the case of cross-lingual text reuse, the problem becomes more difficult as there can be quite different ways of verbalizing (translating) the same text into the target language. Sometimes, it may also happen that some text fragment is wrongly classified as plagiarized text due to the complexity of the problem. Even then, plagiarism detection systems may prove to be of great help to experts by finding the evidence for plagiarism and providing them to human experts for validation purposes [2].

Methodology

To address the CLITR problem, we divide its analysis into two subtasks. Firstly, our methodology dealt with converting the cross-lingual text reuse detection problem into a monolingual one. For this purpose we translated all the suspicious text files into the language of the source files. Secondly, we analyzed the files along two features, i.e. vocabulary matching and sentence/event ordering. For the first subtask, we relied on the

Google and Bing machine translation APIs for automatically translating all the suspicious text files, giving us both the reused and the source files in one language. To check the text reuse, we used similarity measures provided by Lucene, which inherently includes vocabulary matching as it incorporates primarily an inverse document frequency scoring algorithm. We also experimented with matching the ordering of the sentences against the corresponding source files but could not come up with a nice formula that embeds both the Lucene scoring and sentence ordering. So, we submitted only the results that use the Lucene similarity scores without sentence ordering.

Results

We submitted two runs of our system, incorporating two different settings over the threshold of the similarity scores given by Lucene. In the first run we did not set any threshold over the score and associated each suspicious file with a source file. In the second run we set a threshold over the similarity score by taking an average of all the similarity scores leaving the outliers which we got by performing the same methodology over the training data. Table 1 shows the results for both of our runs.

F-measure	**Recall**	**Precision**	***Run***
0.609	0.821	0.484	***1***
0.589	0.795	0.468	***2***

Table 1: Results

Conclusions & Future Work

We were able to achieve respectable result by using the Lucene similarity function. However, we think that the results can be further improved by taking sentence ordering and other lingistic information into account. We obtained a high recall in the first run where we did not put any threshold as we think that the test data is biased towards more reused files. Furthermore, the training data was not grammatically correct, so it was hard to apply some natural language processing techniques that make use of the sentence structure. The use of machine translation constrains us by giving only one way of verbalizing a particular sentence into the target language. Information is being lost in the first step because of this. So, it restricts the results in terms of vocabulary matching.

References:

[1] CLITR workshop, FIRE, India. http://users.dsic.upv.es/grupos/nle/fire-workshop-clitr.html

[2] Barrón-Cedeño, A. and Rosso, P. (2009) Monolingual and Crosslingual Plagiarism Detection. Towards the Competition @ SEPLN09. In: III Jornadas PLN-TIMM, February 5-6, Madrid, Spain