Using Clustering to Identify Outlier Chunks of Text

Notebook for PAN at CLEF 2011

Navot Akiva

Department of Computer Science
Bar Ilan University, Ramat Gan, Israel
Navot@cs.biu.ac.il

Abstract. Intrinsic plagiarism detection is a sub-task of authorship identification in which outlier chunks must be detected solely on the basis of stylistic differences from the main body of the text. We present a first attempt at utilizing words that appear infrequently in a text as stylistic markers for distinguishing outlier chunks in the text. In the first phase of our method we cluster chunks of text represented by usage of infrequent words. In the second phase, we use a training corpus to identify cluster properties of outlier chunks.

Keywords: Intrinsic plagiarism, clustering, outlier detection.

1 Introduction

One of the main difficulties in any plagiarism task is identifying the boundaries of plagiarized text[1]. In the case of intrinsic plagiarism detection, we face the additional difficulty that no source text is available for comparison. We thus need to automatically identify sudden shifts in writing style.

2 Outlier Chunks Identification

Our approach consists of two phases: chunks clustering and cluster properties detection.

2.1 Chunks Clustering Phase

Chunking: For a given text, we first divide the text into chunks consisting of 1000 characters. We then identify the 100 rarest words that appear in at least 5% of the chunks. (Thus we have a set of words that are infrequent in the text but not so infrequent as to be useless. These parameters were not optimized and no doubt can be significantly improved.) Each chunk is now represented by a numerical vector of

length 100 corresponding to the presence or absence of each of the rare words in the chunk. We measure the similarity of pairs of chunks using cosine.

Clustering: We then use a spectral clustering method called n-cut [2] to cluster the chunks. We cluster the texts into only two clusters, which we hope will correspond to the true text and the plagiarized text, respectively. This hope is often unrealistic because there is no guarantee that the plagiarized material is taken from a single source. It might be that different plagiarized sections are not similar to each other; it might also be that there is little or no plagiarized material and the clustering will be along lines that are unrelated to plagiarism.

2.2 Cluster Properties Detection

We thus use a second phase to identify clusters that really do consist of plagiarized text. To do this, we run our clustering method on the training corpus and we measure a variety of properties of each cluster and each chunk in each cluster. These properties include the relative and absolute size of each cluster, the similarity of each chunk to its own cluster, to the other cluster and to the whole document and so forth. The intuition is that plagiarized chunks are those that are close to the centroid of the small cluster and very far from the centroid of the whole document.

We represent each chunk in the training set as a numerical vector recording each of the above values and we label each chunk as including plagiarized material or not. We then use supervised learning methods to learn decision trees using WEKA [3] for distinguishing plagiarized chunks from non-plagiarized chunks.

3 Evaluation Results

We found it was best to learn separate classifiers for short (up to 20K), medium (20K to 250K) and long (above 250K) documents.

We used ten-fold cross-validation to optimize parameter settings and to estimate accuracy results. For reasons of efficiency, we did not use the full training set. In particular, we ignored all documents with more than 40% plagiarism. We also randomly selected chunks from among the remaining documents.

Our cross-validation results are shown in Table 1. Unfortunately, these numbers turned out to be optimistic. On the PAN-2011 evaluation set, we achieved precision 12.7% and recall of 6.6%.

Table 1. Cross validation result.

Training Group	Best Algorithm Applied	Precision	Recall	F-Measure	# Plag. Chunks	# Non-Plag. Chunks
Up to 20K chars.	JRip	48%	14%	21.6%	2000	3700
20K-250K chars.	J48	62%	32%	42.2%	9000	17,000
Above 250K chars.	J48	72%	76%	74%	3000	6,000

Analysis of the results indicates that the method achieved especially poor precision on short documents.

4 Conclusions and Future Work

Despite the poor evaluation results, we believe that our overall method is promising. We should significantly increase the number of training examples on our future experiments. There are a number of parameters that first need to be optimized, including the choice of rare words and the size of the chunks (which need not be constant). There are a number of other considerations that might improve results, including using the full training set and clustering into k>2 clusters.

References

1. Stein, B., Lipka, N., Prettenhofer, P., "Intrinsic plagiarism analysis," Lang. Resources & Evaluation, Volume 45, Number 1, 63-82 (2010).

2. Dhillon,I. S., Guan,Y., Kulis, B., Kernel kmeans: spectral clustering and normalized cuts. Proc. ACM International Conference on Knowledge Discovery and Data Mining (KDD), pp. 551-556. (2004)

3. Hall, M., Frank, E.,, Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H, The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1. (2009)