Lexical Generalisation for Word-level Matching in Plagiarism Detection

Miranda Chong
University of Wolverhampton
Miranda.Chong@wlv.ac.uk

Lucia Specia
University of Wolverhampton
L.Specia@wlv.ac.uk

Abstract

Plagiarism has always been a concern in many sectors, particularly in education. With the sharp rise in the number of electronic resources available online, an increasing number of plagiarism cases has been observed in recent years. As the amount of source materials is vast, the use of plagiarism detection tools has become the norm to aid the investigation of possible plagiarism cases. This paper describes an approach to improve plagiarism detection by incorporating a lexical generalisation technique. The goal is to identify plagiarised texts even if they are paraphrased using different words. Experiments performed on a subset of the PAN'10 corpus show that the matching approach involving lexical generalisation yields promising results, as compared to standard n-gram matching strategies.

1 Introduction

Plagiarism is a growing challenge in modern society. In an attempt to maintain academic integrity, the use of plagiarism detection tools has become the norm in many higher education institutions. However, the methods used in these tools are mostly limited to comparisons of suspicious plagiarised texts and potential source texts at the string level. If the texts have not been copied verbatim, these tools are not able to identify the obfuscated texts effectively. Therefore, the accuracy of these methods is yet to reach a satisfactory level.

This paper investigates the use of preprocessing, morphological and lexical semantics techniques from Natural Language Processing (NLP) in automatic plagiarism detection. The hypothesis is that by enhancing standard string matching approaches with linguistic information it is possible to improve the accuracy of plagiarism identification at the document level. More specifically, the goal is to generalise the text comparison to include morphological and lexical variations (synonyms). Different from previous work, instead of restricting the expansion of words in the documents to synonyms with the same *sense*, we use a simpler approach that considers all possible expansions. This approach does not require word sense disambiguation and is therefore less prone to common errors due to incorrect disambiguation.

This paper is organised as follows: in Section 2 we describe related work in the plagiarism detection field using NLP; in Section 3 we outline the experimental settings; in Section 4 we present the results of our experiments; in Section 5 we discuss the findings; and in Section 6 we conclude and suggest future work.

2 Related Work

Our focus is on external monolingual plagiarism detection of English documents. External detection refers to cases where potential source texts are available for comparison against suspicious plagiarised texts. Following the standard terminology in the field, we name *suspicious document* a potentially plagiarised text, and *source document* the possible origin of the plagiarised material.

Current studies in this area have suggested the use of approaches such as n-gram matching between suspicious and potential source documents. NLP has only recently started to be exploited for this problem. However, most approaches focus on shallow techniques or the processing of very small corpora.

The PAN workshop series "Uncovering Plagiarism, Authorship, and Social Software Misuse" has been organised in recent years to provide a common ground for developing and testing plagiarism detection systems. Each year, the workshop provides a corpus for large-scale detection experiments (Barrón-Cedeno and Rosso, 2010). Reports from the 1st and 2nd competition (PAN'09 and PAN'10) have shown that most competitors used n-gram-based hashed-indexing approach, but little or no effort was made to use NLP techniques. Although some levels of shallow NLP techniques

such as stemming were used to generalise string matching (Costa-jussà et al., 2010; Pereira et al., 2010; Torrejón and Ramos, 2010), the reports did not specify whether the application of these techniques contributed to the detection accuracy. Due to the very short time given to participants to process the corpus for the official competitions, little effort has been made in these competitions to further explore NLP techniques.

Outside of these competitions, lexical resources with synonymy information have been used in a few approaches. Similar to our work, the idea is to generalise the words in the texts by considering synonyms when searching for lexical matching between suspicious and source texts, in addition to exact matching of words.

The use of a lexical thesaurus such as WordNet (Fellbaum, 1998) was investigated by Nahnsen et al. (2005). The paper described the use of lexical resources in text similarity detection, which involved the use of cosine similarity on n-grams of lexical chains, with word sense disambiguation applied to nouns, verbs and adjectives. They computed *tf-idf* of the disambiguated words as a similarity measure but if the WSD process is not accurate, it would affect the similarity scores.

Another research by Chen et al. (2010) has concluded that using WordNet to perform synonym recognition can help determine whether a sentence pair contains similar words. They measure the similarity by comparing the synonyms within each synsets, ie. they compare the synonyms in synset 1 for suspicious document word A and synonyms in synset 1 for source document word B, however, this method would not return any similarity scores if the synonyms are in different synsets even if they belonged to the same word. In comparison, the use of WordNet in Ceska (2009)'s experiment did not show significant improvement over the other shallow text-processing methods. Ceska performed synonymy recognition with word sense disambiguation and it was said using the *ad hoc* rule to choose the "first synset" or word sense disambiguation techniques to choose the "most suitable synset" were not effective.

In previous work we performed experiments on a small-scale manually created corpus to incorporate shallow text pre-processing, morphological, lexical and syntactic information (Chong et al, 2010). The results suggested NLP techniques can help to improve the identification of plagiarised documents. However, besides being small, the corpus contained easily detectable short cases of induced plagiarism. In this paper we concentrate on the subset of linguistic processing techniques identified as the most promising in our previous work and apply it to the much larger PAN'10 corpus. More specifically, we evaluate the use of lexical generalisation in this large-scale scenario, without the need of word sense disambiguation. Since word sense disambiguation is a complex task on its own, we can avoid mistakes resulting from incorrect disambiguation. Syntactic processing was not used here due to the nature of the corpus: a large proportion of the plagiarised cases are artificially created by random text operations including automatically replacing, adding and removing words and changing sentence structure, resulting in text that is not always grammatical.

3 Experimental Settings

3.1 Corpus

The corpus used in the experiment is the PAN'10 [1] corpus. It consists of a total of 11,147 source and 15,925 suspicious documents. Plagiarism cases refer to segments in suspicious documents, annotated in terms of character offsets. Of all the plagiarism cases, 40% are verbatim copies from multiple sources (no obfuscation). Other 40% of cases contain artificially inserted passages with two levels, low or high, of automatic obfuscations such as modifying sentence structures and replacing words with their synonyms. A small proportion of cases (6%) are simulated plagiarism cases where texts were manually rewritten with different wordings using the *Amazon Mechanical Turks*. The remaining cases consisted of translated plagiarism texts, that is, suspicious texts produced from automatically translating source documents using a machine translation system. The length of plagiarism segments in a suspicious document range from a minimum of 50 words to a maximum of 5,000 words, and the segments can come from 1 to more than 50 sources. 50% of the suspicious and source documents contain 1 to 10 pages, 35% contain 10 to 100 pages, and 15% contain 100-1000 pages. The corpus contains both external and in-

[1] 2nd International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse PAN-10 http://pan.webis.de/

trinsic plagiarism cases, that is, cases where plagiarism is to be identified within the actual suspicious document, without referring to a source document.

For practical reasons, in this paper we selected a subset of the PAN corpus: the first 1,000 suspicious documents, along with all 11,147 source documents. Since our goal is to investigate external plagiarism of English texts, all intrinsic and translated plagiarism cases were excluded from the dataset. We therefore removed 186 cases from the subset of 1,000 suspicious documents and 731 non-English cases from the source documents. The experiments presented here are thus based on 814 suspicious documents and 10,416 source documents, which gives a total of 8,478,624 possible pairwise comparisons.

The method used in this paper is a binary classification of documents, that is, we classify each suspicious-source document pair as *plagiarised* or *not plagiarised*. Although in the PAN competition plagiarised cases are expected to be reported at the segment level, in this paper cases are treated at document level, where a pair of documents is considered as *plagiarised* whenever at least one segment within the suspicious document is plagiarised from the source document. Given that NLP techniques are much more computationally expensive than simple string matching techniques, document level processing is a more realistic scenario for this feasibility study. Moreover, flagging plagiarised documents can be a helpful aid for humans checking potential plagiarism cases by filtering out a very large amout of documents from the process.

3.2 Processing Techniques

We follow the standard 2-phase methodology in plagiarism detection. The first phase is *candidate document selection*, that is, filtering documents in order to narrow down the search space to document pairs that can contain plagiarised segments. The second phase is a detailed analysis of the remaining candidate document pairs.

In order to generalise the texts for subsequent similarity comparisons, both source and suspicious documents were processed using the following preprocessing and morphological processing techniques as available in NLTK[2] (Bird et al., 2010).

Tokenisation: determine token (words, punctuation symbols) boundaries in sentences.

Lowercasing: substitute every uppercase letter with their lowercase form.

Punctuation removal: remove all punctuation symbols.

Stemming: morphological analysis to transform words into their stems by removal of derivational affixes, for example: 'computational', 'computing' and 'compute' will be returned to the base form 'comput'. Stemming is used as a common pre-processing method in plagiarism detection task and we have followed this approach.

For the experiment with lexical generalisation, functional words (**stop words**) were removed and all remaining (content) words were generalised using their **WordNet synsets**, that is, groups of synonym words. In other words, we expanded the source and suspicious documents by replacing each of its content word by the words in all of its synsets from WordNet. It is important to notice that WordNet performs morphological generalisation by lemmatising words, that is, converting them into their basic form, for example: 'operative', 'operational' and 'operation' into 'operate'.

3.3 Similarity Metrics

Based on the corpus processed with the techniques described above, the next step is to measure the similarity between source-suspicious document pairs. As shown in Table 1, we differentiate between the proposed approach (Dataset (II)) and a baseline using the same pre-processing steps, but having stemming as a morphological generalization technique, as opposed to the use of WordNet for morphological and lexical generalization (Dataset (I)). We propose a synset overlap metric and compare it against a standard 5-gram overlap metric for our baseline dataset.

Data set	Techniques	Similarity Metric
(I)	Tokenisation Lowercasing Punctuation Removal Stemming	5-gram overlap
(II)	Tokenisation Lowercasing Punctuation Removal Stopwords Removal WordNet All Synsets	Synset overlap

Table 1: Similarity metrics applied to the baseline and proposed approaches

[2] http://www.nltk.org/

The use of overlapping n-grams is a common practise in the PAN competitions; the use of hashed 5-grams was one of the techniques contributing to the top-ranked approaches (Kasprzak and Brandejs, 2010; Zou et al., 2010). Therefore, in this experiment the overlap of chunks of 5-grams was used as our baseline. More specifically, we used the overlap coefficient, a common n-gram similarity metric (Clough and Stevenson, 2009).

$$Sim_{Overlap}(A,B) = \frac{|S(A,n) \cap S(B,n)|}{min(|S(A,n)|,|S(B,n)|)} \quad (1)$$

where $S(A,n)$ and $S(B,n)$ are the unique 5-grams contained in the suspicious and source documents, respectively. The number of common 5-grams in both sets is normalized by the smaller of $S(A,n)$ or $S(B,n)$ to account for differences in the sizes of suspicious and source documents.

For the version of the corpus expanded with WordNet synsets, the matching is performed based on unigrams of synsets. In other words, the number of common synsets of the source $S(A)$ and suspicious $S(B)$ documents is computed and then normalised by the total number of synsets in both suspicious and source documents, using the *Jaccard coefficient*:

$$Sim_{WordNet}(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|} \quad (2)$$

3.4 Filtering

In plagiarism detection tasks, it is essential to perform initial filtering with superficial techniques to reduce the number of potential source documents, and therefore the number of document pairs to be processed in the next stage. The use of progressive filtering makes the application of deeper NLP techniques more feasible in the remaining document pairs. The filtering stage is referred to as the *candidate document selection* and the suspicious-source documents selected for further processing are referred to as *candidate documents*.

In this paper, the filtering strategy is based on empirical observation and consists in applying the following steps to all document pairs in the dataset processed with superficial techniques and 5-gram overlap coefficient (Dataset (I) in Table 1):

1. Rank the documents pairs in descending order according to their similarity scores.
2. For each suspicious document, select the top 10 potential source doc. This resulted in 8,140 document pairs.
3. Remove document pairs that do not have at least 10 common 5-grams or with an overlap coefficient score (Equation 1) of less than 0.01. This resulted in 1,534 candidate document pairs in Dataset I.

The 1,534 candidate document pairs are then processed for lexical generalisation using WordNet (resulting in Dataset II). We then compare and evaluate both datasets using the 1,534 document pairs.

4 Results

We treat the detection problem as a binary classification task where the documents are said to be *plagiarised* when their similarity score is above a certain threshold, or *not plagiarised* if the similarity score is below that threshold. Therefore, standard evaluation metrics of precision, recall and F-score can be employed to measure detection performance. The number of correctly classified plagiarised documents - True Positives (TP), correctly classified non-plagiarised documents - True Negatives (TN), non-plagiarised documents incorrectly classified as plagiarised - False Positives (FP), and the plagiarised documents incorrectly classified as non-plagiarised - False Negatives (FN) are used for the standard calculation of precision, recall, and F-score.

The similarity scores are tested with various thresholds to investigate the trade-off between precision and recall. Ideally, a detection approach should make sure that all potential plagiarised documents are flagged (high recall), but also make sure that non-plagiarised documents are not flagged (high precision), to save humans' time when manually analysing the flagged documents. However, as in most classification tasks, a high recall may come at the price of a low precision, and vice-versa. Therefore, depending on the detection task, it may be more important to favour one metric or another. For this reason, instead of fixing a threshold, we show, in Figures 1, the precision and recall at different thresholds.

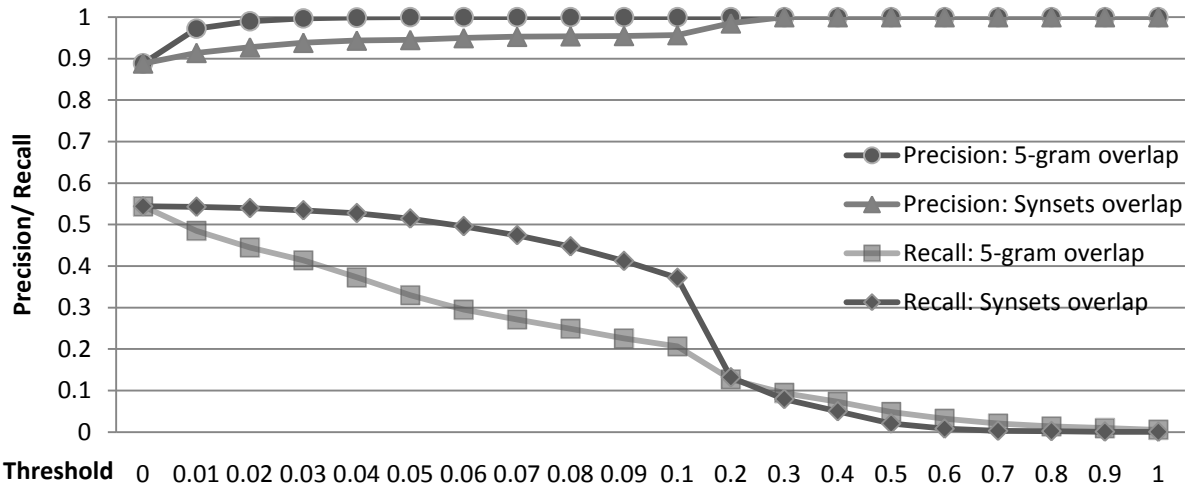

Figure 1: Precision and Recall for several thresholds in the similarity metrics. Statistically significant differences were observed according to pair-wise *t*-test (*p*-value < 0.05) between the baseline 5-gram overlap (Dataset I) and proposed approach Synsets overlap (Dataset II).

5 Discussion

As we can see in Figure 1, the WordNet-based similarity metric shows improvement over the baseline, achieving similar precision and a significantly higher recall for lower thresholds. The high recall figure indicates that using all synsets in the similarity metric can help reduce the number of false negative cases. However, the slightly lower precision indicates that using all synsets may be too lenient. This suggests that the use of WordNet may be more appropriate to investigate a subset of highly suspicious plagiarism cases after filtering by using other methods.

Upon further analysis based on individual levels of obfuscation, that is, the four levels of plagiarism annotation in the PAN'10 corpus (manual paraphrase, low artificial obfuscation, high artificial obfuscation, and no obfuscation), we noticed that the use of WordNet synsets matching is more effective than the 5-gram overlap baseline in all obfuscation levels. Although the baseline is effective in detecting direct verbatim copies, the WordNet synsets matching is capable of achieving better results regardless of how the plagiarised texts have been produced. In particular, this strategy has identified significantly more simulated and obfuscated plagiarism cases than the baseline.

For example, Table 2 shows the the recall of both approaches on different levels of obfuscation, based on a threshold of 0.03.

Obfuscation level	Dataset (I)	Dataset (II)
None	0.48	0.62
Artificial - low	0.42	0.54
Artificial - high	0.35	0.46
Simulated	0.27	0.37

Table 2: Recall obtained by the of 5-gram overlap baseline and the synset-based similarity matching for different obfuscation levels

Although this initial experiment is based on a subset of the corpus, we believe that by using a combination of 5-gram overlap and WordNet-based similarity metrics, a more accurate detection performance could be achieved. Further experiments need to be performed on this direction.

6 Further Work and Conclusions

In this paper we proposed using lexical generalisation to improve the performance of string-based matching plagiarism detection approaches. The experiments were performed with a subset of the PAN'10 corpus, but a similar performance is expected with larger datasets. The results have shown the influence of lexical generalisation on plagiarism detection performance in terms of precision and recall. Different levels of threshold have different effects on precision, recall and F-score. Therefore, the threshold needs to be set in accordance to the detection task requirement. A future direction is to use machine learning algorithms to

set this threshold. Machine learning algorithms will also allow a principled way of classifying documents based on a combination of similarity scores generated from different metrics, such as scores from 5-gram overlap and WordNet synsets.

Further investigation is needed to seek for better filtering strategies to optimise the detection performance, as well as better similarity metrics to account for other linguistic variations. Areas such as Recognising Textual Entailment (RTE) and stylistic approaches used in authorship attribution may provide additional improvements. Semantic parsing by using tools such as semantic role labellers can provide deeper analysis in terms of the semantic structure of texts. It is expected that such rich features will be more effective in identifying simulated plagiarism cases.

Last but not least, future experiments using the PAN corpus will be performed on passage level instead of document level in order to allow comparative evaluation to be performed using the standard PAN evaluation measures.

References

Barrón-Cedeno, A. and Rosso, P. (2010). Towards the 2nd International Competition on Plagiarism Detection and Beyond. Proceedings for the 4th International Plagiarism Conference. Newcastle, UK.

Bird, S., Klein, E. and Loper, E. (2010). Natural Language Processing with Python--- Analyzing Text with the Natural Language Toolkit.

Ceska, Z. (2009). Automatic Plagiarism Detection Based on Latent Semantic Analysis. Doctoral Thesis. University of West Bohemia, CR.

Chen, C.-Y., Yeh, J.-Y. and Ke, H.-R. (2010). Plagiarism Detection using ROUGE and WordNet.

Chong, M., Specia, L. and Mitkov, R. (2010). Using Natural Language Processing for Automatic Detection of Plagiarism. Proceedings of the 4th International Plagiarism Conference. Newcastle, UK.

Clough, P. and Stevenson, M. (2009). Developing a Corpus of Plagiarised Short Answers. Language Resources and Evaluation, 1–20. Springer.

Costa-jussà, M. R., Banchs, R. E., Grivolla, J. and Codina, J. (2010). Plagiarism Detection Using Information Retrieval and Similarity Measures Based on Image Processing Techniques. Proceedings of the Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) 2010 Workshop. Padua, Italy.

Fellbaum, C. (1998, ed.). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Kasprzak, J. and Brandejs, M. (2010). Improving the Reliability of the Plagiarism Detection System Lab Report for PAN at CLEF 2010. Proceedings of the Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) 2010 Workshop. Padua, Italy.

Nahnsen, T., Uzuner, O. and Katz, B. (2005). Lexical chains and sliding locality windows in content-based text similarity detection. CSAIL Memo.

Pereira, R. C., Moreira, V. P. and Galante, R. (2010). UFRGS @ PAN2010 : Detecting External Plagiarism Lab Report for Pan at CLEF 2010. Proceedings of the Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) 2010 Workshop. Padua, Italy.

Torrejón, D. A. R. and Ramos, J. M. M. (2010). CoReMo System (Contextual Reference Monotony) Lab Report for PAN at CLEF 2010. Proceedings of the Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) 2010 Workshop. Padua, Italy.

Zou, D., Long, W.J., and Ling, Z. (2010). A Cluster-Based Plagiarism Detection Method. Proceedings of the Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) 2010 Workshop. Padua, Italy.