Linguistic and Statistical Traits Characterising Plagiarism

Miranda Chong[1] *Lucia Specia*[2]
(1) University of Wolverhampton, Stafford Street, Wolverhampton, WV1 1SB UK
(2) University of Sheffield, 211 Portobello, Sheffield, S1 4DP UK
`miranda.chong@wlv.ac.uk, l.specia@sheffield.ac.uk`

Abstract

This paper investigates the problem of distinguishing between original and rewritten text materials, with focus on the application of plagiarism detection. The hypothesis is that original texts and rewritten texts exhibit significant and measurable differences, and that these can be captured through statistical and linguistic indicators. We propose and analyse a number of these indicators (including language models, syntactic trees, etc.) using machine learning algorithms in two main settings: (i) the classification of individual text segments as original or rewritten, and (ii) the ranking of two or more versions of a text segment according to their "originality", thus rendering the rewriting direction. Different from standard plagiarism detection approaches, our settings do not involve comparisons between supposedly rewritten text and (a large number of) original texts. Instead, our work focuses on the sub-problem of finding segments that exhibit rewriting traits. Identifying such segments has a number of potential applications, from a first-stage filtering for standard plagiarism detection approaches, to intrinsic plagiarism detection and authorship identification. The corpus used in the experiments was extracted from the PAN-PC-10 plagiarism detection task, with two subsets containing manually and artificially generated plagiarism cases. The accuracies achieved are well above a by chance baseline across datasets and settings, with the statistical indicators being particularly effective.

Keywords: Text Reuse, Plagiarism Detection, Plagiarism Direction.

1 Introduction

Current studies in plagiarism detection are mostly focused on the detection of plagiarised segments in a collection of documents or within a document. The direction of plagiarism is thus predetermined. Original documents and plagiarised documents are provided separately and the task is to determine which segments of plagiarised texts (if any) are copied or rewritten from which segments of original texts. This is generally done through a large number of pairwise comparisons: the "suspicious" text is compared against original texts using similarity metrics, which are mostly based on word overlap.

To date, very superficial metrics such as n-gram matching achieve the state of the art performance on verbatim plagiarism cases. While this is a perfectly reasonable approach for plagiarism detection, it has some limitations. Firstly, pairwise comparisons in large collections are computationally very expensive and in practice very simple filtering strategies are used to rule out most of the original texts. Secondly, for real-world, open data collections such as the web, pairwise comparisons may be less reliable. It is not uncommon to find multiple versions of a plagiarised material on the web, and thus the concept of an "original" text becomes less clear.

This study looks at the plagiarism practice from a novel perspective: instead of measuring the similarity between pairs of texts, the goal is to investigate traits that distinguish original from rewritten texts based on examples of both types of texts. We make use of machine learning algorithms and exploit a number of linguistically and statistically-motivated features – e.g. statistical language models, syntactic trees and features from *translationese* studies – to (i) determine whether an individual text segment is original or plagiarised, and (ii) determine the direction of plagiarism, that is, rank a pair of texts according to their originality. This approach requires observing patterns of features in individual texts, without any direct comparison between texts.

2 Related Work

Research on distinguishing original from plagiarised texts is very limited. The only existing work analyses character 16-grams on artificially generated plagiarised documents from the PAN-PC-10 corpus (Grozea and Popescu, 2010). These plagiarism cases are generated via automatic means with various obfuscation levels through the insertion, deletion, replacement of words, and other simple operations. At document level, overall accuracies reached about 75%. Tests on highly obfuscated artificial documents reached an accuracy of 69.77%. This analysis has indicated that there seem to be significant differences between original and plagiarised texts in the PAN corpus. However, given the way the artificial plagiarism cases were produced, this finding is somehow trivial. To the best of the authors' knowledge, no research has been done on the more challenging cases of manually plagiarised documents, nor at the level of segments, as opposed to documents.

Considering translation as a process of text rewriting (in a different language), studies on *translationese* (i.e. on distinguishing original from translated texts in a given language) and on detecting translation direction in a bilingual pair of texts are also relevant for this work. Most work in this area follows the *Translation Universals* theory (Gellerstam, 1986), which hypothesises that translated texts tend to exhibit characteristics that are different from non-translated texts. The theory was further explored by (Baker, 1993, 1996) and based on such a theory, research has been done for identifying specific properties that reflect these universals and using them to automatically test these universals. For example, on a corpus of original (non-translated) and translated texts in Italian, (Baroni and Bernardini, 2006) finds that features such as the distribution of function words, personal pronouns and adverbs are very relevant. (Pastor et al., 2008) explored the existence of the *simplification* universal – which states that translated texts are simpler than their source counterpart –, suggesting that this universal does affect Spanish translated texts. Also focusing on the simplification universal, the studies by (Ilisei et al., 2010; Ilisei and Inkpen, 2011) on Romanian and Spanish translationese use morphological and simplification-based features.

A six-lingual study by (Halteren, 2008) using frequencies of word n-grams shows that it is possible to distinguish between translated and non-translated texts down to their respective original languages. This is followed by the work of (Lembersky et al., 2011, 2012) which uses statistical language models for each language. Furthermore, a study by (Volansky et al., 2012) explores the differences between original, manually translated and machine translated texts.

The experiments on translation direction identification suggest that translated texts have lower lexical richness and higher number of frequent words. They point out that simplification-based features are very helpful, but alone they are not sufficient to distinguish original from translated texts. Although by nature plagiarised texts are very different from translated texts, we exploit insights gained from these and other related studies in the features we use, including many of the simplification-based features.

3 Methodology and Experimental Settings

A supervised machine learning approach is proposed to test the hypothesis that original and plagiarised texts exhibit significant and measurable differences. We build models based on various linguistically and statistically-motivated features. The models are tested on manually simulated and artificially generated plagiarism cases. Each case consists of a segment of text. Well-known machine learning algorithms are used for two tasks: binary classification and ranking. These two variations of the approach are evaluated in the same way: computing the accuracy of each algorithm in categorising segments as original or plagiarised.

3.1 Corpus

This study uses the PAN-PC-10 plagiarism detection task corpus (Potthast et al., 2010), which comprises books from the Project Gutenberg.[1] Two datasets were extracted from this corpus, as shown in Table 1. The segments are extracted according to the annotation provided in the corpus: pre-defined labels for manually simulated and artificial plagiarism sequences of words.

The *Artificial Dataset* is composed of a randomly selected subset of plagiarised texts that were generated automatically by three types of edits: (i) a set of text operations, which include replacing, shuffling, removing or inserting words at random, (ii) semantic word variations by replacing words by similar or related words (such as synonyms), and (iii) POS-preserving word shuffling, which shuffles words at random but keeps the same ordering of part-of-speech tags. The *Simulated Dataset* is composed of all the manually simulated plagiarised segments available in the corpus. The plagiarism cases were manually written using mechanical turks to simulate plagiarism by paraphrasing the original texts.

Given the way the artificial dataset was created, it is expected that our approach will perform significantly better on this dataset, while the simulated dataset represents a much more challenging, but more realistic, problem.

[1] http://www.gutenberg.org

	Statistics	Simulated Dataset	Artificial Dataset
Original texts	Number of segments	4067	4000
	Minimum length	74 words	46 words
	Maximum length	745 words	4506 words
	Average length	409.5 words	2276 words
Plagiarised texts	Number of segments	4067	4000
	Minimum length	21 words	38 words
	Maximum length	1190 words	3917 words
	Average length	605.5 words	1977.5 words

Table 1: Corpus statistics

3.2 Machine learning algorithms

In the **binary classification task** the goal is to assign each instance in the collection to one of the two classes: *original* or *plagiarised*. In the **ranking task**, the goal is to sort two (or potentially more) versions of a segment according to the order in which they were created, in other words, to identify the plagiarism direction.

The algorithms applied here are as follow: the rule-based learner Repeated Incremental Pruning to Produce Error Reduction (RIPPER) for binary classification and Support Vector Machines (SVM) for ranking. RIPPER[2] was selected as a good representative of symbolic classifiers: the rules produced can help identify relevant features for specific cases. SVM is one of the most robust and best performing algorithms in many language processing tasks. For ranking, the SVMrank algorithm[3] (Joachims, 2006) is used with a linear kernel. Both classification and ranking models are trained and tested using 4-fold cross-validation. In addition, a structured prediction version of SVM was applied as an alternative binary classifier: SVM-light-TK[4] (Moschitti, 2006), wich uses tree kernels with (partial) syntactic trees as features.

3.3 Feature extraction and selection

The datasets are pre-processed with sentence segmentation, tokenisation and lowercasing. The part-of-speech (POS) tags and lemmas of words and the syntactic trees of sentences are generated using the Stanford CoreNLP toolkit[5] (Klein and Manning, 2003). Pre-defined lists of function words (Koppel and Ordan, 2011) and stopwords[6] are used.

N-gram language models (with n = 3 & 5) are built using the KenLM toolkit[7] (Heafield, 2011). The corpus used to build such models consisted in a random selection of 1.7M segments extracted from the entire “original” collection of the PAN-PC-10 corpus, excluding all the documents containing one or more segments present in our two datasets. We then use these language models to calculate the scores for both plagiarised and original segments.

Features that capture simplification, morphological, statistical and syntactic aspects of texts are investigated. Based on the simplification universal, we extract the following **simplification-based features**:

[2] We used the Jrip Weka implementation of this algorithm: `http://www.cs.waikato.ac.nz/ml/weka/`

[3] `http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

[4] `http://disi.unitn.it/moschitti/Tree-Kernel.htm`

[5] `http://nlp.stanford.edu/software/corenlp.shtml`

[6] `http://nltk.org/`

[7] `http://kheafield.com/code/kenlm/`

1. Average token length: number of characters normalised by the number of tokens.
2. Average sentence length: average number of tokens in all sentences of the segment.
3. Information load: proportion of lexical words to tokens. Lexical words refer to nouns, verbs, adjectives, adverbs and numerals.
4. Lexical variety: type/token ratio obtained by normalising the word types over all words.
5. Lexical richness: proportion of type lemmas per tokens. Different from lexical variety, lexical richness considers the lemmatised word types normalised by all words.
6. Proportion of sentences without finite verbs.
7. Proportion of simple sentences: sentences that contain only one finite verb.
8. Proportion of complex sentences: sentences that contain more than one finite verb.

To capture plagiarism traits that may occur at the **morphological** level, the following features are extracted:

9. Proportion of nouns over tokens.
10. Proportion of prepositions over tokens.
11. Proportion of pronouns over tokens.
12. Proportion of stopwords over tokens.
13. Proportion of finite verbs over tokens.
14. Grammatical cohesion rate: proportion of grammatical words over lexical words. Grammatical words are determiners, articles, prepositions, auxiliary verbs, pronouns, conjunctions and interjections.
15. Individual function words: each function word in the pre-defined list is extracted as an individual feature, such as "the", "of", "and", "to", "be", "someone", "self" etc.
16. Proportion of function words in texts: number of function words over word tokens.

The following shallow **statistical features** are proposed:

17. Number of sentences in the segment.
18. Number of tokens in the segment.
19. Number of characters in the segment.
20. Language model 3-gram log probability.
21. Language model 3-gram perplexity (all tokens).
22. Language model 3-gram perplexity (without end of sentence tags).
23. Language model 5-gram log probability.
24. Language model 5-gram perplexity (all tokens).
25. Language model 5-gram perplexity (without end of sentence tags).

Finally, from a more linguistically motivated perspective, (partial) syntactic trees are used with the tree-kernel algorithm (the other algorithms do not allow structured features):

26. Syntactic trees: dependency-based parse trees for all sentences in the segments.

4 Results and Discussion

The baseline results are defined according to the distribution of the two classes in the datasets, which is 50:50. Therefore, the baseline accuracy is 50%. The machine learning algorithms described in Section 3.2 are used with different feature sets as shown in Table 2, along with the results for each combination of algorithm and feature set.

With respect to the algorithms, the comparison shows that the rule-based classification (RIPPER) and the ranking (SVM-rank) algorithms using pre-selected features perform very similarly, and well above the by chance classifier, with the rule-based algorithm doing slightly better. The precision, recall and f-score of the feature sets with RIPPER are given in Table 3.

The *pre-selected* feature set contains the top 12 features ranked according to their Information Gain computed on the training set: F2, F3, F6, F13, F14, F19, F20, F21, F22, F23, F24, F25. These features include some morphological, statistical and simplification indicators, showing that these feature families are complementary. The improvement using these features over the set of all features is not consistent across datasets.

Algorithm	Feature set	Simulated	Artificial
Baseline: by chance	-	50%	50%
RIPPER	All	74.67%	98.15%
RIPPER	Pre-selected	75.66%	97.94%
RIPPER	Simplification-based	59.81%	70.24%
RIPPER	Morphological	59.53%	68.08%
RIPPER	Statistical	74.17%	97.78%
SVM-rank	Pre-selected	74%	95%
SVM-tree kernels	Syntactic	56.17%	79.9%

Table 2: Accuracy of algorithms and feature sets in classifying cases as “original” or “plagiarised”

Dataset	Class	Feature set	Precision	Recall	F-score
Simulated	Original	Pre-selected	75.8%	75.4%	75.6%
		Statistical	73.6%	75.5%	74.5%
		Simplification-based	59.9%	59.4%	59.7%
		Morphological	59.8%	58.2%	59%
	Plagiarised	Pre-selected	75.5%	75.9%	75.7%
		Statistical	74.8%	72.9%	73.8%
		Simplification-based	59.7%	60.2%	60%
		Morphological	59.3%	60.8%	60%
Artificial	Original	Pre-selected	98.4%	97.5%	97.9%
		Statistical	97.8%	97.7%	97.8%
		Simplification-bases	67.8%	72.2%	72.2%
		Morphological	66.1%	74.1%	69.9%
	Plagiarised	Pre-selected	97.5%	98.4%	97.9%
		Statistical	97.7%	97.8%	97.8%
		Simplification-based	73.5%	63.3%	68%
		Morphological	70.5%	62.1%	66%

Table 3: Precision, recall and f-score of various feature sets using RIPPER

On the comparison between the features sets, it was observed that using statistical features alone yields nearly the same performance as using all features. Features involving language models are amongst the best performing. Statistical features performed significantly better in the Artificial Dataset. The relative improvement of these features in the Simulated Dataset over simplification or morphological features is 14%. In the Artificial Dataset, the relative improvement over the other features is 27%. Morphological, simplification and syntactic features are not as discriminative on their own, but their performance is well above the baseline. Interestingly, tree kernels on the Artificial Dataset performs significantly better than tree kernels on the Simulated Dataset with respect to the baseline. This may be a consequence of the fact that artificial cases consistently exhibit malformed syntax, which makes it easier for syntactic features to capture relevant distinctions.

4.1 Discussion and examples

Across all experiments with different algorithms and feature sets, the problem of identifying artificially generated plagiarism cases proved significantly easier than that of identifying manually plagiarised cases. Given the nature of the operations applied to generate artificial cases, this result is not surprising. Nevertheless, the near-100% performance for these cases is a very positive result. It shows that this approach can be used for the filtering of candidates in a real plagiarism detection system, one of the applications suggested in this paper.

It is arguable that the experiments above show only a marginal gain from using a combination of simplification, morphological and statistical methods with respect using simple statistical features. Although previous studies have also pointed out that statistical features are generally relevant for related problems, confirming this finding for the specific problem we address is an interesting contribution of this study.

With respect to the novel, linguistically motivated features suggested here, they perform well on artificially generated texts, which exhibit a considerable proportion of ungrammatical constructions. Along with statistical features, these may help future work in identifying not only the existence and direction of plagiarism, but also several types or levels of plagiarism.

We found no strong evidence that the simplification universal applies to plagiarism. Although some simplification-based features do seem relevant, they could be interpreted from different perspectives, which are not necessarily related to simplification.

A closer inspection on some examples of pairs of segments is given below.

Example 1: Correctly classified pair of cases by SVM-rank and SVM-tree kernels from the *Simulated Dataset*

Original: But a better idea of the journal can perhaps be given, by stating what it lacked than what it then contained. It had no leaders, no parliamentary reports, and very little indeed, in any shape, that could be termed political news.

Plagiarised: The journal could better be described by what was missing than what it contained. It lacked leaders, had no parliamentary reports and in no way could be described as political news.

In this example, we speculate that in addition to the strong features throughout all instances (the language model features), others contributed to classify this pair. They include the average sentence length, number of characters, and independent clause rate. For example, the average sentence length for the plagiarised text is lower than the original text. Also, the proportion of nouns is higher in the original text and the lexical richness is lower in the plagiarised text. These clues suggest that the simplification traits were good indicators in this particular case.

Example 2: Incorrectly classified pair of cases by SVM-rank and SVM-tree kernels from the *Simulated Dataset*

Original: There is a great gain in time of acceleration and for stopping, and for the Boston terminal it was estimated that with electricity 50 per cent, more traffic could be handled, as the headway could be reduced from three to two minutes.

Plagiarised: There is a huge profit in time of speeding up and for slowing down, and for the Boston extremity it was guessed that with current 50 percent, more movement could be lifted, as the headway could be minimised from three to two minutes.

Example 2 does not contain any simplification traits but only synonym substitution. The shallow statistical features failed to identify any differences between the two segments. The length of both texts is virtually the same and they are both equally fluent. Morphological and syntactic features did not perform well either. The proportion of grammatical and lexical words remains the same, and the word order and syntactic structure in both texts is the same.

Example 3: Incorrectly classified pair of cases by SVM-rank, but correctly classified by SVM-tree kernels from the *Artificial Dataset*

Original: "Giulietta," at last said the young man, earnestly, when he found her accidentally standing alone by the parapet, "I must be going to-morrow." "Well, what is that to me?" said Giulietta, looking wickedly from under her eyelashes.

Plagiarised: "well, what is that to me?" said Giulietta, standing alone under the parapet, earnestly, when he found her were accidentally looking wickedly from by her eyelashes. "Giulietta," at last young the man, "I must be going to-morrow.

Example 3 involves shuffling of sequences of words. As both texts kept the same words and length, none of the statistical, morphological and simplification features were able to distinguish the two. On the other hand, SVM-tree kernels correctly classified these cases according to their subtrees structure. This suggests that syntactic clues should be considered especially when all other features fail.

Conclusions

This paper presented a study on the underexplored area of distinguishing original from reused text segments, with application to plagiarism detection. A number of statistical and linguistic indicators were explored using a supervised machine learning approach to distinguish between original and plagiarised texts, as well as to rank pairs of original-plagiarised texts according to the order in which they were created. Overall, the study showed that original and reused texts exhibit distinguishable traits. It thus confirms our hypothesis that original texts and plagiarised texts exhibit significant differences and that these are measurable via computational means.

The findings of this study can be directly used to improve the filtering performed prior to more complex comparisons in plagiarism detection approaches. It can also be used to improve intrinsic plagiarism detection and authorship attribution. In addition, this study lays the foundation for further research on text reuse, as it can be expanded to cover multiple versions of the same text, as well as cross-lingual text reuse.

We plan to further investigate this problem in a number of directions, including the use of other types of rewritten texts, such as news, with potentially more than one version for each original text, as well as different levels of text reuse (as in (Clough et al., 2002)).

References

Baker, M. (1993). *Corpus Linguistics and Translation Studies: Implications and Applications*. John Benjamins, Amsterdam.

Baker, M. (1996). *Corpus-based Translation Studies: The Challenges that Lie Ahead*. John Benjamins, Amsterdam.

Baroni, M. and Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Clough, P., Gaizauskas, R., Piao, S., and Wilks, Y. (2002). Meter: Measuring text reuse. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 152–159. Association for Computational Linguistics.

Gellerstam, M. (1986). *Translationese in Swedish novels translated from English*. CWK Gleerup.

Grozea, C. and Popescu, M. (2010). Who's the thief ? automatic detection of the direction of plagiarism. *Lecture Notes in Computer Science*, 6008:700–710.

Halteren, H. V. (2008). Source language markers in europarl translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, number August, pages 937–944, Manchester, UK.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, number 2009, pages 187–197, Edinburgh, UK.

Ilisei, I. and Inkpen, D. (2011). Translationese traits in romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, 2.

Ilisei, I., Inkpen, D., Pastor, G. C., and Mitkov, R. (2010). Identification of translationese: A machine learning approach. In *11th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 503–511, Iaşi, Romania.

Joachims, T. (2006). Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, pages 217–226, Philadelphia, USA. ACM Press.

Klein, D. and Manning, C. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japa.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1318–1326, Portland, USA.

Lembersky, G., Ordan, N., and Wintner, S. (2011). Language models for machine translation: original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 363–374, Edinburgh, Scotland.

Lembersky, G., Ordan, N., and Wintner, S. (2012). Adapting translation models to translationese improves smt. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 255–265, Avignon, France.

Moschitti, A. (2006). Making tree kernels practical for natural language learning. In *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Pastor, G., Mitkov, R., Afzal, N., and Pekar, V. (2008). Translation universals: do they exist? a corpus-based nlp study of convergence and simplification. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08)*, number October, pages 21–25, Waikiki, Hawaii.

Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 997–1005, Beijing, China. Association for Computational Linguistics.

Volansky, V., Ordan, N., and Wintner, S. (2012). More human or more translated ? original texts vs . human and machine translations. In *11th Bar-Ilan Symposium on the Foundations of Artificial Intelligence*, Ramat Gan, Israel.