Information Theoretical and Statistical Features for Intrinsic Plagiarism Detection

Rashedur Rahman
IRT-SystemX & LIMSI-CNRS
Paris-Sud University
rashedur.rahman@limsi.fr

Abstract

In this paper we present some information theoretical and statistical features including function word skip n-grams for detecting plagiarism intrinsically. We train a binary classifier with different feature sets and observe their performances. Basically, we propose a set of 36 features for classifying plagiarized and non-plagiarized texts in suspicious documents. Our experiment finds that entropy, relative entropy and correlation coefficient of function word skip n-gram frequency profiles are very effective features. The proposed feature set achieves F-Score of 85.10%.

1 Introduction

Extrinsic plagiarism detection attempts to detect whether a document is plagiarised relative to reference documents. IPD (intrinsic plagiarism detection), which is relatively new, detects the plagiarised section(s) in a suspicious document *without* using any reference document. The basic hypothesis behind IPD is different writers have their own styles and they maintain these in their writings consciously or subconsciously. Sometimes it is very difficult to define the reference set for the task of external plagiarism detection. Additionally, the source of the plagiarized text may not be available in digitized format. Therefore, researchers are trying to answer whether it is possible to detect plagiarism without using any reference.

In this paper, we investigate some information theoretical and statistical measurements for IPD as a binary classification task. A set of 36 features has been proposed for classifying plagiarized and non-plagiarized segments in the suspicious documents. We use the PAN-PC-11 (Potthast et al., 2010) corpus compiled for IPD task. The PAN corpus is artificially plagiarised and it provides a meta-file mentioning the offsets of plagiarised and non-plagiarized parts for each suspicious document. We consider that each suspicious document is written by single author and it is either partially plagiarised or not plagiarised and we try to identify the text-segments that differ in writing style compared to the whole document. We train an SMO (Platt, 1998) classifier in Weka3.6 (Hall et al., 2009) by using 10 fold cross-validation. Then the classification performances are observed with different feature sets according to the standard precision, recall and F-score.

The next sections are organized as follows: section 2 discusses related works and section 3 briefly describes information theoretical and statistical features. The text segmentation and windowing process is summarized in section 4 while the experimental framework and baseline feature sets are discussed in section 5. Section 6 compares the classification performances with different feature sets and finally, the paper concludes in section 7.

2 Related Work

A series of regular studies on plagiarism detection were started following the first international competition for plagiarism detection, the PAN[1] workshop in 2009. Potthast et al. (2009) provides an overview on PAN'09 including the corpus design for plagiarism detection, quality measurements and the methods of plagiarism detection developed by the participants.

Zu Eissen and Stein (2006) proposed the first method for IPD and presented a taxonomy of plagiarism with methods for analysis. They also proposed some features including *average sentence length*, *part-of-speech* features, *average stopword number* and *averaged word frequency class* for quantifying the writing style. Some researchers used character *n*-gram profiles for the task of IPD

[1] http://pan.webis.de/

Proceedings of the SIGDIAL 2015 Conference, pages 144–148,
Prague, Czech Republic, 2-4 September 2015.

(Stamatatos, 2009; Kestemont et al., 2011). Oberreuter et al. (2011) proposed word n-gram based method and they assumed that different writers use different sets of words that they repeat frequently. Tschuggnall and Specht (2012) proposed the *Plag-Inn* algorithm that finds plagiarized sentences in a suspicious document by comparing grammar trees of the sentences.

Stamatatos (2009) introduced sliding window and proposed a *distance function* for calculating the dissimilarity between two texts based on a character tri-gram profile. Stamatatos (2011) employed n-grams of function word sequence with different lengths and found significant impact to distinguish between plagiarised and non-plagiarized texts. We employ function words differently as skip n-gram profiles for measuring entropy, relative entropy and correlation coefficient as discussed in Section 5.2. Stein et al. (2011) employed unmasking technique and proposed a set of features of different types for example POS, function words etc for intrinsic plagiarism analysis.

Seaward and Matwin (2009) and Chudá and Uhlík (2011) proposed compression based methods for IPD. They measured the *Kolmogorov complexity* of the distributions of different *parts-of-speech* and word classes in the sentences. For calculating the complexity a binary string is generated for each distribution and later the string is compressed by a compression algorithm.

3 Information Theoretical and Statistical Features

Shannon Entropy (Shannon, 1948) has a great impact on communication theory or theory of information transmission, it measures the uncertainty of a random variable. Mathematically, entropy is defined as in equation (1).

$$H(X) = -\sum_{i=1}^{n} p(x_i)\ log_2(p(x_i)) \qquad (1)$$

$$KLD_{(p||q)} = \sum_{x \in X} p(x)\ log_2\left(\frac{p(x)}{q(x)}\right) \qquad (2)$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right) \qquad (3)$$

We measure entropy of n-gram frequency profile generated from each text-window (X) for quantifying the writing style. Manning and Schütze (1999) measured the distance between two probability distributions by using *Relative entropy* or *Kullback-Leibler divergence* (KLD) which is calculated by using the equation (2). The *Pearson correlation coefficient* (Pearson, 1920) or simply *correlation coefficient* measures the linear correlation between two samples that is calculated by the equation (3). Since the task of IPD does not use any reference document we require a robust method for comparing small sections of the document relative to the whole document under question. Measuring the relative entropy and correlation coefficient between a small section and the rest of the document are possible methods. We use the frequency profiles of n-grams generated from the individual text-window (X) and the complete suspicious document (Y) separately for calculating relative entropy and correlation coefficient. The probability distributions of n-gram frequencies (P and Q) is calculated from n-gram frequency profiles (from X and Y) for measuring the relative entropy.

4 Text Segmentation and windowing

To define the small sections of text for comparison to the rest of the document, we experiment with window of different lengths (1000, 2000, 5000 characters). To prepare the corpus for training and testing to support this additional experimentation, we separate plagiarised and non-plagiarized sections of the documents in the corpus according to the offsets (as indicated in the meta-file). By doing this we can guarantee that the smaller texts we generate are still accurately annotated as to whether the content is plagiarised or not. The whole procedure is illustrated in figure 1.

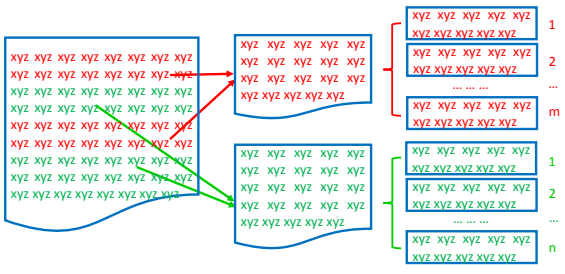

Figure 1: Text segmentation and windowing

5 Experimental Framework and Feature Sets

This section illustrates the experimental framework of IPD task by combining the preprocessing and classification tools, the framework is graphically described in figure 2. After extracting and windowing the corpus, we calculate different feature values for generating the feature vectors. Before calculating the features, several text preprocessing tasks, for example, tokenizing, sentence detection and POS-tagging are employed. We gen-

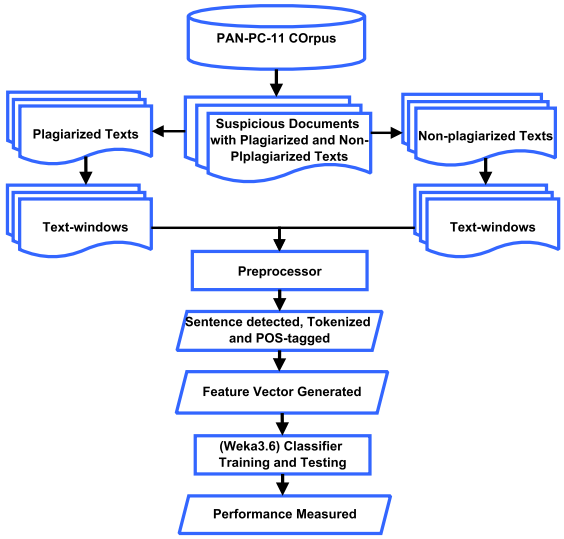

Figure 2: Experimental framework

erate several feature vectors for different baseline feature sets and proposed feature set. Then a classifier model is trained with the feature sets, we train SMO classifier with 10 fold cross validation in *Weka 3.6* explorer interface. Equal number of plagiarized and non-plagiarized text samples are trained with the classifier. We train the classifier with $8,100$ text segments from each class where each segment initially contains $5,000$ characters. Finally, the classification performances are observed for different feature sets.

5.1 Baseline feature sets

We used three different baseline feature sets for the experiment which are listed below:

- Baseline-1 (feature set used by Stein et al. (2011)): used 30 features that includes lexical and syntactical features, surface features, vocabulary richness and readability measurement-based features, n-gram-based features, POS-based features etc.
- Baseline-2 (feature set used by Seaward and Matwin (2009)): calculated the *Kolmogorov complexity* of function words and different parts-of-speech.
- Baseline-3 (*distance function* proposed by Stamatatos (2009)): measured *distance function* or *style-change score* of the text-windows with respect to the whole suspicious document by using their character tri-gram profiles.

5.2 Proposed feature set

We propose 36 features for IPD including entropy, relative entropy, correlation coefficient, skip n-grams of function words etc. Lavergne et al. (2008) and Zhao et al. (2006) used relative entropy for fake content detection and authorship attribution accordingly. Islam et al. (2012) classified readability levels of texts by using both entropy and relative entropy. Stamatatos (2011) used function word n-grams for exterinsic plagiarism detection but here we generate several skip n-grams of function words instead of simple n-grams. Guthrie et al. (2006) used 1 to 4 skip n-grams for modelling unseen sequences of words in the text. Here we summarize the proposed feature set:

- **Character *tri-gram* frequency profile**: we measure entropy for text windows and relative entropy and the correlation coefficient of the character tri-gram frequency profile for the text windows and documents. Additionally, we calculate *average n-gram frequency class* by using the equation of *average word frequency class* proposed by Zu Eissen and Stein (2006). Here we have 4 features: entropy, relative entropy, correlation coefficient and n-gram frequency class calculated from character tri-gram frequency profiles of text-windows and complete document.
- ***bi-gram* and *tri-gram* frequency profile with 1, 2, 3 and 4 skips** : we measure entropy, relative entropy, correlation coefficient of function-word bi-gram and tri-gram frequency profile with 1, 2, 3 and 4 skips. Additionally, we calculate the *style change scores* with these frequency profiles using the *distance function* proposed by Stamatatos (2009). For generating the skip n-gram profiles of function-words we extract the function words sequentially from each sentence.

We generate function-word skip n-gram profiles of the text segments by considering only the function words at sentence level instead of passage level as Stamatatos (2011) used. Here we have 32 features: entropy, relative entropy, correlation coefficient and style-change score calculated from 8 function-word skip n-gram frequency profiles.

6 Experimental Results

We observe that the proposed feature set achieves the highest F-Score compared to the baseline feature sets as illustrated in figure 3. All the feature sets together obtain a promising F-Score of 91% while the three baselines combined result in an F-Score around 89%. The proposed feature set achieves an 85% F-Score which is the highest compared to the three baseline feature sets. Baseline-1 and baseline-2 obtain F-Score around 68% and 62% while baseline-3 surprisingly results in an 84% F-Score as a single feature. We pair feature sets and observe their performances, figure 4 shows that the proposed feature set increases the F-Score with the combination of baseline feature sets.

Figure 5 depicts separate observations of entropy, relative entropy, correlation coefficient and distance function of function word skip n-gram frequency profiles. Here we notice that relative entropy achieves a very good F-Score of 72%, entropy and correlation coefficient also obtain better F-Scores than the distance function. Though distance function results in very good F-Score with the character tri-gram frequency profile it does not perform good enough with the function word skip n-gram frequency profile. Distance function with function word skip n-gram frequency profile obtains around a 35% F-Score which is the lowest compared to other functions with function word skip n-gram frequency profile. We also observe the effect of different window lengths (discussed in section 4) on classification performance, the classification performance increases for each feature set if the window length is increased. All the feature sets combined result in F-Score of 82% and 87% for window lengths of 1000 and 2000 characters accordingly while a 91% F-Score is achieved with the window length of 5000 characters.

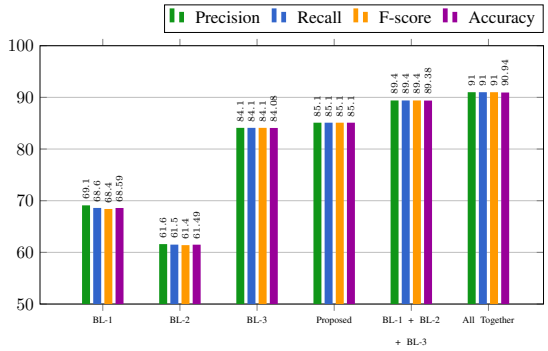

Figure 3: Performance observation of the baseline and proposed feature sets

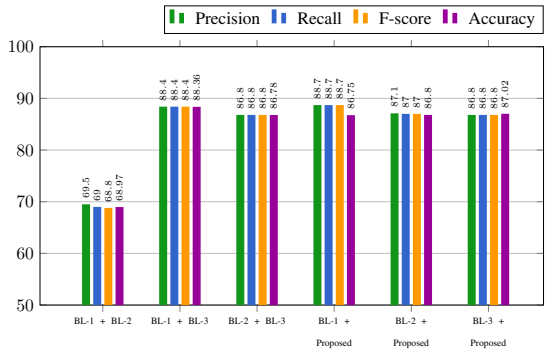

Figure 4: Performance observation of the coupled feature sets

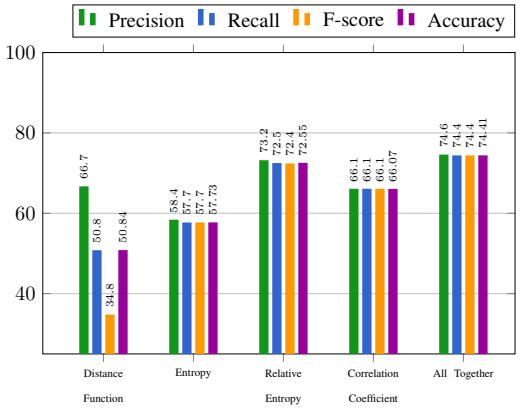

Figure 5: Performance observation of function word skip n-gram based features

7 Conclusion

In this paper we proposed a set of new features for intrinsic plagiarism detection that support arguments for continued research on IPD. In the future we would like to evaluate these features on human-plagiarised and different domain corpora. We are also interested in expanding the IPD task by considering the case that a suspicious document is written by multiple authors.

Acknowledgement

This paper is a part of my master thesis work while studied at Frankfurt University of Applied Sciences. I am very thankful to my thesis supervisor Dr. Alexander Mehler and my especial thanks to IRT-SystemX for ensuring me to attend at SIGdial conference. I also thank my SIGdial mentor and reviewers for their feedback and guidance.

References

Daniela Chudá and Martin Uhlík. The plagiarism detection by compression method. In *Proceedings of the 12th International Conference on Computer Systems and Technologies*, pages 429–434. ACM, 2011.

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4, 2006.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

Zahurul Islam, Alexander Mehler, Rashedur Rahman, and AG Texttechnology. Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation.(Accepted)*, 2012.

Mike Kestemont, Kim Luyckx, and Walter Daelemans. Intrinsic plagiarism detection using character trigram distance scores. *Proceedings of the PAN*, 2011.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. Detecting fake content with relative entropy scoring. In *PAN*, 2008.

Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

Gabriel Oberreuter, Gaston LâĂŹHuillier, Sebastián A Ríos, and Juan D Velásquez. Approaches for intrinsic and external plagiarism detection. *Proceedings of the PAN*, 2011.

Karl Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.

John C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.

Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeno, and Paolo Rosso. Overview of the 1st international competition on plagiarism detection. In *3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE*, 2009.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, August 2010. Association for Computational Linguistics.

Leanne Seaward and Stan Matwin. Intrinsic plagiarism detection using complexity analysis. In *Proc. SEPLN*, pages 56–61, 2009.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1): 3–55, 1948.

Efstathios Stamatatos. Intrinsic plagiarism detection using character n-gram profiles. *Proceedings of the PAN*, pages 38–46, 2009.

Efstathios Stamatatos. Plagiarism detection based on structural information. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1221–1230. ACM, 2011.

Benno Stein, Nedim Lipka, and Peter Prettenhofer. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82, 2011.

Michael Tschuggnall and Günther Specht. Plagiinn: intrinsic plagiarism detection using grammar trees. In *Natural Language Processing and Information Systems*, pages 284–289. Springer, 2012.

Ying Zhao, Justin Zobel, and Phil Vines. Using relative entropy for authorship attribution. In *Information Retrieval Technology*, pages 92–105. Springer, 2006.

Sven Meyer Zu Eissen and Benno Stein. Intrinsic plagiarism detection. In *Advances in Information Retrieval*, pages 565–569. Springer, 2006.