Experiments to investigate the utility of nearest neighbour metrics based on linguistically informed features for detecting textual plagiarism

Per Almquist and Jussi Karlgren
Swedish Institute of Computer Science (SICS), Stockholm
Royal Institute of Technology (KTH), Stockholm

Abstract

Plagiarism detection is a challenge for linguistic models — most current implemented models use simple occurrence statistics for linguistic items. In this paper we report two experiments related to plagiarism detection where we use a model for distributional semantics and of sentence stylistics to compare sentence by sentence the likelihood of a text being partly plagiarised. The result of the comparison are displayed for visual inspection by a plagiarism assessor.

1 Plagiarism detection

Plagiarism is the act of copying or including another author's ideas, language, or writing, without proper acknowledgment of the original source. Plagiarism analysis is a collective term for computer-based methods to identify plagiarism. (Stein et al., 2007a) Plagiarism analysis can be performed *intrinsically* — a text is examined for internal consistency, to detect suspicious passages that appear to diverge from the surrounding text, or *externally* — a text is inspected with respect to some known corpus to find passages with suspiciously similar content to other text.

In external plagiarism detection, it is assumed that the source document d_{src} for a given plagiarized document d_{plg} can be found in a target document collection D. Typically, plagiarism detection then proceeds in three stages:

1. candidate selection through retrieval of a set of candidate source documents D_{src} is retrieved from D_{plg};
2. candidates d_{src} from D_{src} is compared passage by passage with the suspicious document d_{plg} and every case where a passage from d_{plg} appears to be similar to some passage in some d_{src} is noted;
3. followed by some post-processing to remove false hits.(Stein et al., 2007b; Potthast et al., 2010)

2 PAN workshop series

A series of workshops on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, organised since 2007, have provided the field with a shared task and test materials in the form of gold standard text collections with manually and automatically constructed plagiarised sections marked for experimental purposes. Some of the plagiarised sections are *obfuscated* with word replacement, edits, and permutations. The research results from the workshops are comparable, since they are to a large extent performed on the same materials using the same starting points and same target measures.

Example results relevant to this study (and on the whole none too surprising) are that unobfuscated plagiarism can be detected with a reasonable accuracy by the top plagiarism detectors. The recall decreases slightly with increasing obfuscation and that longer stretches of plagiarised material are easier to detect than shorter segments.(Potthast et al., 2010)

Table 1: Stylometric features

Name	Description
arg	Sentence is argumentative (*merely, for sure, ...*)
cog	Sentence describes cognitive process (*remember, think, ...*)
com	Sentence is complex (average word length > 6 characters or sentence length > 25 words)
date	Sentence contains one or more date references
fin	Sentence contains a money symbol or a percentage sign
fpp	Sentence contains first person pronouns
le	Sentence refers to named entities such as a person or an organization
loc	Sentence mentions a location
neg	Sentence contains a grammatical negation
num	Sentence contains numbers
pa	Sentence contains place adverbials (*inside, outdoors ...*)
pun	Sentence contains punctuation in addition to its ending punctuation
se	Sentence contains split infinitives or stranded prepositions
spp	Sentence contains second person pronouns
sub	Sentence has subordinate clauses
ta	Sentence contains time adverbials (*early, presently, soon ...*)
tim	Sentence contains one or more time expression
tpp	Sentence contains third person pronouns
uni	Sentence contains symbols representing a unit of measurement

3 Our experimental set-up

The base of the experiment described here is to test a finer-grained analysis of plagiarised texts than other previous work. We use a sentence-by-sentence comparison of the suspicious text (d_{plg}) with all sentences of each target text (d_{src}) in D_{src} using two different similarity measures: one based on overall semantic similarity, the other on specific stylometric measures.

The experiment is not a full scale evaluation of our method but is intended to test the practicability of our approach. Given that we have a suspicious text and some reasonable number of candidate source texts (through some retrieval procedure) — can we detect the likelihood of plagiarism in a text by inspecting the sentence sequence of the suspicious text one by one? This paper reports a selected plot dry run of the methodology performed over a number of sample texts. A full scale evaluation is pending.

3.1 Data

The experiments are performed on the PAN-PC-09(Potthast et al., 2009)[1] corpus since it can be used free of charge for research and contains plagiarized passages which has previously been marked and labeled as plagiarism, so that we know beforehand which passages are plagiarism.

The corpus is divided in two sets, one for training and one for test. The training set is further divided into three parts (D_{plg}, D_{src}, and L). D_{plg} contain the documents that are suspicious and might plagiarize documents in D_{src}, where D_{src} contain only orginal documents that make out the sources of any plagiarism in D_{plg}, and L is the solution, the labeling that tells us which sentences in D_{plg} that plagiarize what sentence in D_{src}.

3.2 Nearest neighbour metrics

We use cosine similarity (as defined in equation 1) to represent how similar two vectors are.

$$sim_{COS}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

For every sentence $s \in d_{plg}$ its nearest neighbour score (as defined in equation 2) is calculated.

$$\max(sim_{COS}(s, x)) \text{ for all sentences } x \in D_{src} \quad (2)$$

[1] http://www.uni-weimar.de/cms/medien/webis/research/corpora/pan-pc-09.html

The nearest neighbour metric has the fortunate feature that a value of 1 describes identical or duplicate vectors. So if we were to find nearest neighbour values of 1 those two sentences would be very alike and therefore we would be able to assume that the newer sentence plagiarizes the older sentence.

In this experiment two settings for the experiment were used. In experiment one below we evaluate how well the nearest neighbour metric of two vectors in a semantic *word-space model* manage to detect plagiarism. In experiment two below we evaluate how well the nearest neighbour metric of two binary vectors based on 19 different stylometric features manage to detect plagiarism.

3.3 Target plots

As an example plagiarism inspection mechanism we plot the nearest neighbour metric with the sentences of a text along the x-axis against the score of the sentence. The objective is to find a stretch of material where several sentences have high nearest neighbour scores. As a comparison we will plot the gold standard plagiarism labeling of respective sentence and let the label for a sentence being plagiarism have value 1 and 0 otherwise. Now we can just plot our nearest neighbour scores and our modified labels against the sentences in the corpus.

3.4 Experiment 1: overall semantic similarity score

The sentences of d_{plg} were compared by semantic similarity using a word-space model (Schütze, 1993) as a base for computing similarity between sentences. Each sentence was represented by the centroid of its constituent words in a word-space trained on the entire test corpus. The implementation was based on previous work on effective word-space models.

> "The word-space model is a computational model of word meaning that utilizes the distributional patterns of words collected over large text data to represent semantic similarity between words in terms of spatial proximity." (Sahlgren, 2006).

The word-space model, from the work in (Kanerva, 1988) and (Sahlgren, 2006), models the meaning of words according to their distribution, creating a representation of their semantics based on where and how in the text the words appear.

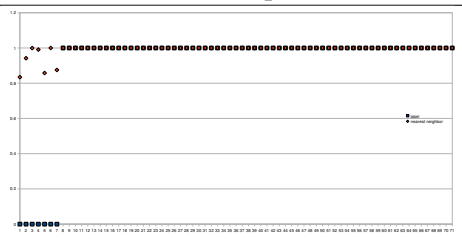
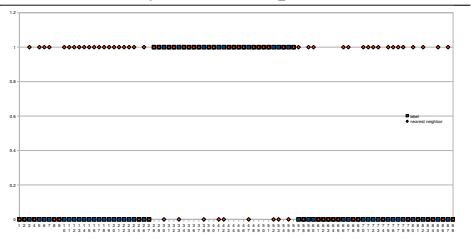
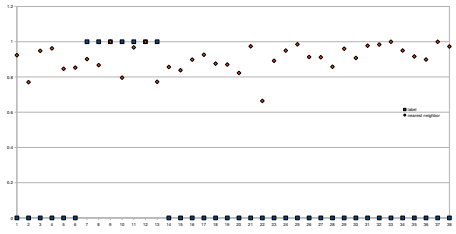
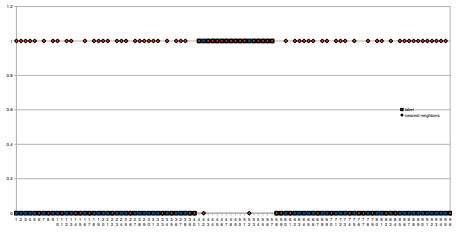
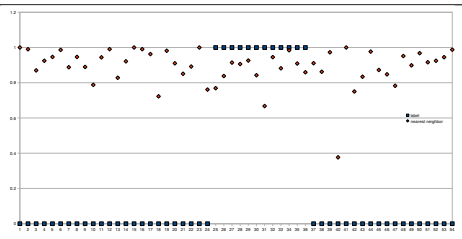
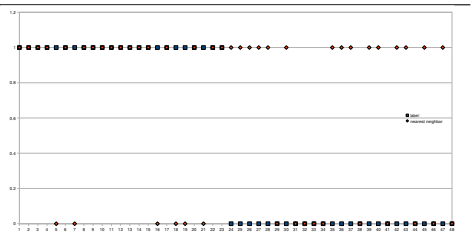
Obfuscation	Semantic space	Stylometric space
none		
low		
high		

Table 2: Plots of semantic word-space and stylometric sentence space neighbours for texts with known plagiarized sections

The word-space is a high dimensional vector space where every word is represented by a vector. Two words are semantically similar if their respective vectors are similar. For example the words "yellow" and "green" could be argued to have similar semantic meaning. So the vectors for yellow and green should be expected to be similar as seen in figure 1.

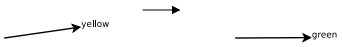


Figure 1: The vectors for the words "yellow" and "green" in a semantic space.

The word-space model is, as its name implies, mainly used to model words. It can however be used to model other linguistic entities such as sentences and documents using workarounds. A sentence can be represented by taking the centroid of the sentence's individual word's vectors. Therefore if the sentence "A yellow car." was changed to "A green car." the centroid ought not to change too much since the only change to the centroid would be one vector that in the first case represented the word "yellow" and in the second case the word "green" and these vectors should be fairly similar, as seen in figure 2. In our model we use a semantic word-space to model sentences under the hypothesis that if a sentence were to be obfuscated its semantic similarity would be kept. We build a semantic space for the corpus under consideration and assign each sentence a representative centroid vector of 3000 real dimensions for every sentence in the corpus. We then perform, for every sentence vector $\vec{s}$ from D_{plg}, the nearest neighbour search $nn(\vec{s}, \vec{s_{src}})$ against all the vectors $\vec{s_{src}}$ in D_{src}.

Figure 2: A centroid of a changed sentence.

3.5 Experiment 2: stylometric similarity score

The 19 stylometric features that were chosen can be seen in table 1, and were chosen based on the work in (Biber, 1988) and (Karlgren, 2000). Our intention was to capture the authors' writing styles. We tried to find features that would not change if another author were to copy the text and even obfuscate it. Therefore we chose features that

- binds the texts to its topic, such as numbers or units of measurement.
- anchors the text to its context, i.e. named entities, location or time.
- captures peculiarities in the author's writing style: split infinitives or stranded prepositions.
- indicates how complex the language is, such as long sentences or subordinate clauses.

For every sentence in D_{plg} we extracted the stylometric features into a 19th dimensional binary vector $\vec{f}$. We then extracted 470 unique 19th dimensional binary vectors $\mathbf{F}_{src}$, based on the same stylometric features, from D_{src}. Then we performed the nearest neighbour search $nn(\vec{f}, \vec{f_{src}})$ against all the vectors $\vec{f_{src}}$ in $\mathbf{F}_{src}$.

4 Results

Table 2 shows the results for the nearest neighbour scores for both experiments, run on a test text with a known plagiarized section with the corresponding source text. We have three plots representing different levels of obfuscation of plagiarism, namely; a high level of obfuscation, a low level of obfuscation, and no obfuscation. To determine the effectiveness of each nearness measure, the results (red rhomboids) are displayed together with an indication of which section of the text is plagiarized (blue squares) noted with a score of 1 and a score of 0 for the non-plagiarized sections.

5 Conclusions

5.1 Experiment 1: overall semantic similarity

We find that the semantic space model:

- is a good detector for no obfuscation;
- does not hold up for obfuscated materials, neither for low or high obfuscation since it is based on the presence of each word in the text; and consequentially
- needs tuning so that specifically topical terms are weighted up compared to less topical terms. This should be done specifically for the topic in the candidate document being examined, since presumably the topic under consideration is the most likely topic to be plagiarized.

5.2 Experiment 2: stylometric similarity

We find that the stylometric similarity score

- which is a dramatic dimensionality reduction unsurprisingly gives a large number of false positives for all levels of obfuscation;
- gives a comparatively high precision even for a high level of obfuscation.

5.3 Directions

Coming experiments will establish whether the combination of the two knowledge sources and the preservation of sequence information in the candidate source texts might provide effective results for a plagiarism detection task. Previous experiments on sequence encoding of stylistic information seem to indicate that sequential information can contain the right type of information to distinguish writing style. (Karlgren and Eriksson, 2007)

Acknowledgements

This work is performed at SICS, supported by the Swedish Research Council (Vetenskapsrådet) through the project "Distributionally derived grammatical analysis models".

References

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Pentti Kanerva. 1988. *Sparse Distributed Memory*. MIT Press, Cambridge, MA, USA.

Jussi Karlgren and Gunnar Eriksson. 2007. Authors, genre, and linguistic convention. In *SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*.

Jussi Karlgren. 2000. *Stylistic Experiments In Information Retrieval*. Ph.D. thesis, Department of Linguistics, Stockholm University.

Martin Potthast, Andreas Eiselt, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2009. PAN Plagiarism Corpus PAN-PC-09. http://www.webis.de/research/corpora.

Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. 2010. Overview of the 2nd international competition on plagiarism detection. In Martin Braschler and Donna Harman, editors, *Notebook Papers of CLEF 2010 LABs and Workshops*. CLEF, Padua, Italy.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Department of Linguistics, Stockholm University.

Hinrich Schütze. 1993. Word space. In *Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems, NIPS'93*, pages 895–902, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Benno Stein, Moshe Koppel, and Efstathios Stamatatos. 2007a. Plagiarism analysis, authorship identification, and near-duplicate detection (PAN 07). *SIGIR Forum*, 42(2):68–71.

Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. 2007b. Strategies for retrieving plagiarized documents. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen P. de Vries, editors, *30th Annual International ACM SIGIR Conference (SIGIR 07)*, pages 825–826. ACM, July.