Towards the Exploitation of Statistical Language Models for Plagiarism Detection with Reference

Alberto Barrón Cedeño and Paolo Rosso

Universidad Politécnica de Valencia

July, 2008

Overview

- Introduction
- LM approach
- Experiments
- Discussion
- Conclusions

Introduction

Plagiarise

To robe credit of another person’s work; in text it means including text fragments from an author without giving him the corresponding credit

Introduction

Plagiarise

To robe credit of another person’s work; in text it means including text fragments from an author without giving him the corresponding credit

In this work we describe our first attempt to detect plagiarised fragments in a text employing statistical Language Models (LMs) and perplexity.

Introduction

1. **Intrinsic plagiarism analysis**
   [Meyer zu Eissen and Stein, 2006, 2007]
   - No reference corpus is exploited
   - Idea: Search for variations (syntax, grammatical categories or text complexity) through the suspicious text

Introduction

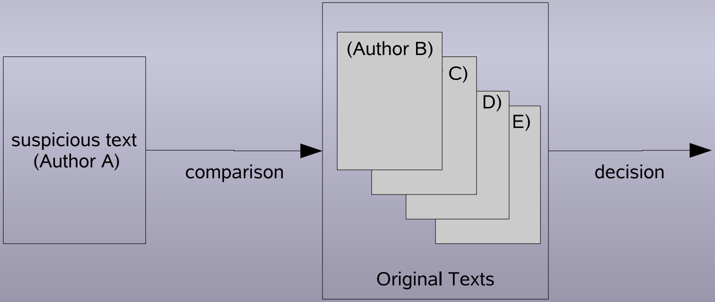
1. **Intrinsic plagiarism analysis**
   [Meyer zu Eissen and Stein, 2006, 2007]
   - No reference corpus is exploited
   - Idea: Search for variations (syntax, grammatical categories or text complexity) through the suspicious text

2. **Plagiarism analysis with reference**
   [Si et al., 1997, Iyer and Singh, 2005]
   - A reference corpus of original documents is needed
   - Idea: to compare fragments from the suspicious document with the original documents in the reference corpus

Introduction

We are interested in the second approach but...

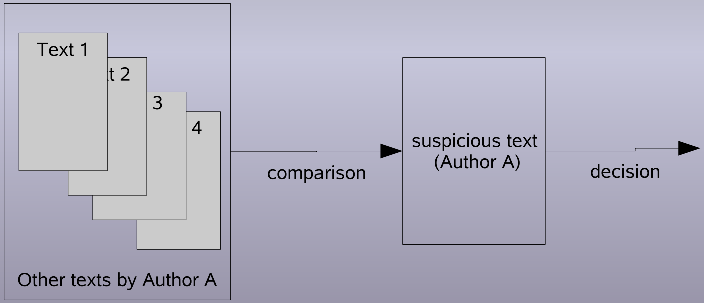
Usually The reference corpus is conformed by original documents



Introduction

We are interested in the second approach but...

Here The reference corpus is conformed by texts written by the author of the suspicious document



Introduction

Statistical Language Model (*LM*)

A LM "tries to predict a word given the previous words" [Manning and Schutze, 2000].

Ideal calculation:

$$P(W) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1 w_2) \cdots P(w_n|w_1 \cdots w_{n-1})$$

Introduction

Statistical Language Model (*LM*)

A LM "tries to predict a word given the previous words" [Manning and Schutze, 2000].

Ideal calculation:

$$P(W) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1 w_2) \cdots P(w_n|w_1 \cdots w_{n-1})$$

n-grams approach (case $n = 3$)

$$P_3(W) = P(w_{n-2}) \cdot P(w_{n-1}|w_{n-2}) \cdot P(w_n|w_{n-2} w_{n-1})$$

LM approach

Basic idea

- Computing the probability of n-grams in a corpus of texts from one author (representation of vocabulary, grammatical frequency and writing style)
- These representations can be compared to other texts in order to look for candidates to plagiarism

LM approach

Is a fragment f a plagiarism candidate?

LM approach

Is a fragment f a plagiarism candidate?

- Determine if a text is similar to another one based on perplexity, frequently used in order to evaluate how good a LM describes a language: “our author language“

$$PP_2 = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_i-1)}}$$

- The lower a text perplexity is, the more predictable its words are. In other words, the higher a perplexity is, the bigger the uncertainty about the following word in a sentence

LM approach

Hypothesis Given a LM m calculated over texts T written by author A. The perplexity of fragments $g, h \in T'$, given that g has been written by A and h has been "plagiarised" from an author B will be clearly different.

LM approach

Hypothesis Given a LM m calculated over texts T written by author A. The perplexity of fragments $g, h \in T'$, given that g has been written by A and h has been ”plagiarised” from an author B will be clearly different.

Specifically, $PP_m(g) \ll PP_m(h)$

Experiments: corpus

We have carried out experiments over two different kind of texts:

Specialised Corpus about Lexicography topics written by only one author

Literature A set of books written by Lewis Carroll and some passages from William Shakespeare texts

Experiments: corpus

Corpora preprocessing:

		vocabulary and syntactic richness	morphosyntactic style
i	original text	■	
ii	part-of-speech		■
iii	stemmed text	■	

Experiments: corpus

Training partition has been used for the LMs calculation

Test partition contains randomly inserted fragments written by a different author

Experiments: corpus

Training partition has been used for the LMs calculation

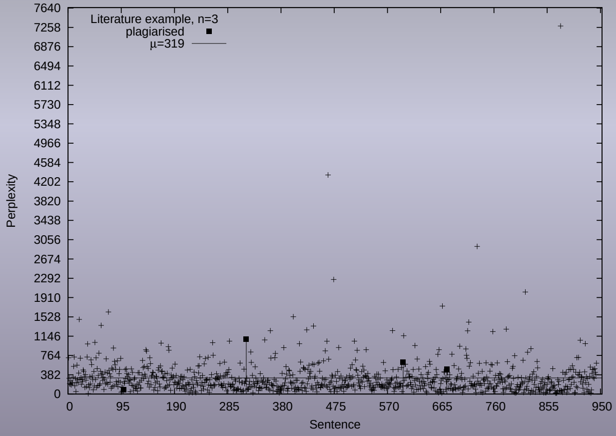
Test partition contains randomly inserted fragments written by a different author

In order to identify candidates, we calculate the perplexity of each sentence with respect to the LM associated to the author

Experiments: results

Results over the literature corpus

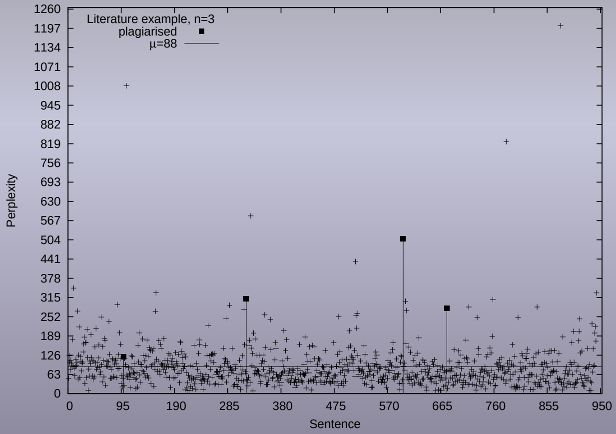
Considering the original text:



Experiments: results

Results over the literature corpus

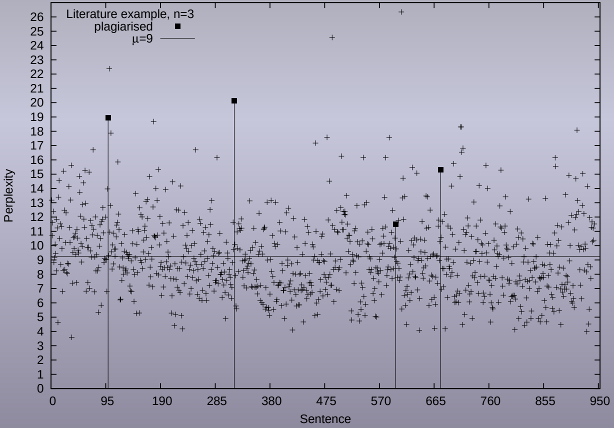
Considering the stemming of the text:



Experiments: results

Results over the literature corpus

Considering the POS of the text:



Discussion

This approach considers three of the five stylometric features categories useful for the plagiarism detection task [Meyer zu Eissen and Stein, 2006]:

Original and stemmed	*Syntactic features* (writing style)
	Special words counting (vocabulary richness)
POS	*Part-of-speech classes quantifycation*
Not considered	*Text statistics* (character level) *Structural features*

Discussion

Perplexity (as we applied it) is not enough to discriminate plagiarised from ”legal” fragments but...

Discussion

Perplexity (as we applied it) is not enough to discriminate plagiarised from ”legal” fragments but...

Is a good idea to consider it?

Discussion

Perplexity (as we applied it) is not enough to discriminate plagiarised from ”legal” fragments but...

Is a good idea to consider it?

What about original text, POS and stem versions?

Conclusions

1. We have considered perplexity on three different levels: word, part-of-speech and stem.
2. Unfortunately, there are non-plagiarised fragments that present high perplexity. However, plagiarised fragments seem to stand out in the highest positions when we consider these features.
3. We know that the perplexity feature space of plagiarised and non-plagiarised segments is not completely separable, but we believe that including perplexity among other features may improve the results.

References

- Iyer, P. and Singh, A. (2005).
  Document similarity analysis for a plagiarism detection system.
  2nd Indian Int. Conf. on Artificial Intelligence (IICAI-2005), pages 2534–2544.
- Manning, C. D. and Schutze, H. (2000).
  Foundations of Statistical Natural Language Processing.
  The MIT Press Publisher, Cambridge Massachusetts and London, England.
- Meyer zu Eissen, S. and Stein, B. (2006).
  Intrinsic plagiarism detection.
  Lalmas et. al. (Eds.): Advances in Information Retrieval Proc. of the 28th European Conf. on IR research, ECIR 2006, London, pages 565–569.
- Si, A., Leong, H. V., and Lau, R. W. H. (1997).
  Check: a document plagiarism detection system