

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

Alberto Barrón-Cedeño¹ and Paolo Rosso¹

Abstract. To plagiarise is to rob credit of another person's work. Particularly, plagiarism in text means including text fragments (and even an entire document) from an author without giving him the correspondent credit. In this work we describe our first attempt to detect plagiarised segments in a text employing statistical Language Models (LMs) and perplexity.

The preliminary experiments, carried out on two specialised and literary corpora (including original, part-of-speech and stemmed versions), show that perplexity of a text segment, given a Language Model calculated over an author text, could be a relevant feature in plagiarism detection.

1 INTRODUCTION

The Automatic Plagiarism Detection, a close related problem to the Automatic Authorship Attribution, has become a relevant task in Information Retrieval, scholar environments and even scientific circles.

There are some applications which try, for example, to detect whether a student report is plagiarised or not². Inside of specialised circles, there are cases when a person takes text fragments from other authors without making the corresponding citation and, in extreme cases, different authors claim for the authorship of a text and even an idea.

Language Models, commonly used in Speech Recognition [7] and Information Retrieval [11, 5], have been exploited in Automatic Authorship Attribution of text [10, 2] and even of source code [4]. In the first case, character level n-grams and perplexity are considered to determine the authorship of the analysed document. In the second case, frequencies of byte level n-grams are used to decide.

State of the art in Automatic Plagiarism Detection allows to detect word by word plagiarism, even when fragments have been modified [14, 6]. In this work we are trying to exploit lexical and grammatical level Language Models (n-grams and perplexity) to detect plagiarised fragments in a text.

The paper is organised as follows. Section 2 describes some of the current advances in the task of plagiarism detection with a reference corpus. Section 3 gives an overview of statistical Language Models and perplexity, in order to determine how well a Language Model could represent a language. Section 4 gives a description of the preliminary experiments we carried out over specialised and literary texts (Sections 4.1 and 4.2) and discusses the obtained results (Section 4.3). Finally, in Section 5 we draw some conclusions.

¹ Natural Language Engineering Lab., Dpto. Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Spain, email: {lbarron, proso}@dsic.upv.es

² See for instance <http://www.turnitin.com/static/plagiarism.html>

2 CURRENT APPROACHES IN AUTOMATIC PLAGIARISM DETECTION WITH REFERENCE

The automatic plagiarism detection can be mainly classified in two approaches based on the exploitation (or not) of a reference corpus.

In the case when no reference corpus is exploited [9, 16], the idea is to find variations through the text of the suspicious document (D_s), like syntax, grammatical categories, text complexity or the verbal form (*I play, she plays, we played*) used in the text. On the other hand, when a reference corpus is used [14, 6], the basic idea is to compare fragments (f) of the suspicious document (D_s) with the documents in a reference corpus (C). Of course, the reference corpus contains only non-plagiarised documents.

The reason for using a reference corpus in order to detect plagiarism in a given text is obvious. In order to decide if a text is plagiarised, we should compare it with other texts looking for common fragments.

In this way, the task could be reduced to make an exhaustive comparison to answer the question: *Is there a fragment $f \in D_s$ included in a document of C ?*

If this problem is approached directly, two difficulties appear immediately: the first one is the need of a huge big reference corpus in order to make a serious search of fragments $f \in D_s$ in C , and second, the processing cost of making all the necessary comparisons is, in a high level, $O(n \cdot m)$ being n and m the length of D_s and C in fragments respectively (the real cost of this kind of comparisons decreases dramatically using hash-based techniques [15]).

Trying to avoid these difficulties the CHECK system [14] pre-processes the documents to determine their "semantic meanings", considering factors like document structure or keywords. This system detects the subject of D_s in order to only compare it with the related documents in C , the original documents corpus. In those cases where paragraphs in D_s and C are semantically related, a per-sentence comparison is made.

The same CHECK architecture is used in [6], but the per-sentence comparison is made using the *dot plot* technique. The advantage is that each word in the analysed sentence is compared with all the words of the sentences in the reference corpus. Two sentences are considered similar if they pass a given threshold (based on the common occurrence of words), a reason to consider a sentence suspicious.

As we have said, the dimension of a corpus must be really big. For example, the plagiarism detection tool offered by *Turnitin* (see footnote in Section 1) not only searches fragments in a reference corpus, but also in the Web.

3 ON STATISTICAL LANGUAGE MODELS

A statistical Language Model (*LM*) “tries to predict a word given the previous words” [8]. LMs have been mainly used in speech and optical character recognition [1, 12], and statistical machine translation [3, 17] between other Natural Language Process applications, but are not limited to these tasks.

To predict which word is the next given its history, the best option should be to consider all the words before it in the text. The probability of a given sentence $w_1w_2 \dots w_n$, if we know $w_{\{1,2,\dots,n-1\}}$ but not w_n , would be given by the Bayes conditional probability, based on the chain rule, $P(W) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1w_2) \dots P(w_n|w_1 \dots w_{n-1})$. Unfortunately, the training set to correctly define these probabilities must be extremely big and, no matter the extension, we will never have a representation for all the possible sentences in a text.

The option is to consider LMs only of n -grams. Over this framework, the model is based on strings conformed by n words, including the analysed one (common values are $n = \{2, 3\}$). The n -gram probability definition considering, for example, $n = 3$ is $P_3(W) = P(w_{n-2}) \cdot P(w_{n-1}|w_{n-2}) \cdot P(w_n|w_{n-2}w_{n-1})$.

Our main idea is that if we compute the probability of n -grams in a corpus of texts from one author, we will have a representation of her vocabulary, grammatical frequency and even writing style. These representations can be compared to other texts in order to look for candidates for plagiarised segments.

The question now is how to determine if a text is similar to another one. Alike [10], we have opted for perplexity, one way to express language theory’s entropy, that is frequently used in order to evaluate how good a LM describes a language: “our author language”.

Formula 1 includes the equation of perplexity (PP), where N is the number of tokens in the analysed text and $P(w_i|w_{i-1})$ is the probability of word w_i given w_{i-1} . This is the case for the perplexity calculation for bigrams.

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

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 ([8, pp. 60-78] to further investigate this issue).

Two tools for LMs and perplexity calculation freely available are SRILM and Cambridge-CMU³. In the preliminary experiments we carried out the first one.

4 THE LANGUAGE MODEL APPROACH

No matter there exist works where multiple features are considered for the task of automatic plagiarism detection (such as [9]), we have opted for starting our explorations in this area working only with one feature: perplexity. It is just for this reason that our results cannot be directly compared with others obtained from more robust techniques.

At the moment our aim is not to improve current results in this field, but to determine whether or not this kind of characterisation of an author style could be useful for this task in order to possibly combine it with other features in the future.

As we have seen in Section 3, a low perplexity means that, given a sequence of words, we are prepared enough to predict, with a low

error rate, which will be the next one. Considering this, we define our main hypothesis:

Hypothesis Let k be the LM of a corpus composed by texts T written by an author A . The perplexity of fragments $g, h \in T'$, given that the fragment g has been written by A and the fragment h has been “plagiarised” from another author will be clearly different. Specifically, $PP_k(g) \ll PP_k(h)$.

Trying to prove (or reject) our hypothesis, we have carried out two main experiments: one over “specialised texts” (scientific papers) and another over “general literature texts” (novels, child literature), which we describe in Sections 4.1 and 4.2, respectively.

For these experiments, we have not only used the original documents. We have pre-processed all the texts in order to consider:

- i* the original text
- ii* the part-of-speech of the text
- iii* the stemmed text

We consider these three versions of the text in order to be able to represent the writer style. Specifically, we tried to recognise author’s vocabulary and syntactic richness, (*i*) and (*iii*), and morphosyntactic style (*ii*). Part-of-speech and stems have been obtained with Treetagger [13].

Independent LMs have been calculated over the three versions of the training corpus considering $\{2 - 4\}$ – *grams*.

With respect to the testing, we split the test corpus in sentences including those that were “artificially plagiarised” before applying to them the same pre-processing of the training set (we have considered the dot as the only delimiter among them).

4.1 Experiments over specialised texts

For this case, we have used a corpus about Lexicography topics written by only one author. One section of the corpus (composed of around 11,628 words), was used for the LM calculation and the other one for the test. In the test partition, we randomly inserted fragments about related topics, but written by other authors.

In order to identify the “plagiarised” fragments (in this case paragraphs), we calculated the perplexities of each sentence with respect to the LM of the author. Figure 1 shows the perplexity of each sentence in the test corpus based on trigrams⁴.

Due to the fact that it considers aspects such as singular/plural and verbal time, the perplexity values of the original text (a) are the highest of the three. The highest perplexities are $PP_{25} = 1132.15$ and $PP_1 = 980$, where 25 and 1 are the number of the sentence in the entire text. Sentence S_{25} has only seven words and contains a cite of the kind “*author_a (2001)*” and *author_a* did not appear in the training corpus, therefore, probability $P(author_a \in n - gram) \rightarrow 0$. In the case of sentence S_1 , it contains the title of the paper, author and author’s organisation, that is not English, so it contains words in another language.

The first plagiarised segment appears in the sixth place of the list of sentences sorted by perplexity. It is S_{27} : “*Such plain text representation is usually processed to add structure explicitly in a machine readable form.*” with $PP_{27} = 608.21$. This sentence contains six words that never appeared in the training corpus.

Working on the stemmed text (b) we consider only the richness of author’s vocabulary without caring about the additional features

³ See www.speech.sri.com/projects/srilm/ and www.speech.cs.cmu.edu/SLM_info.html respectively

⁴ In Figures 1 and 2 symbol “+” represents non-plagiarised sentences and a black square with a vertical bar plagiarised ones.

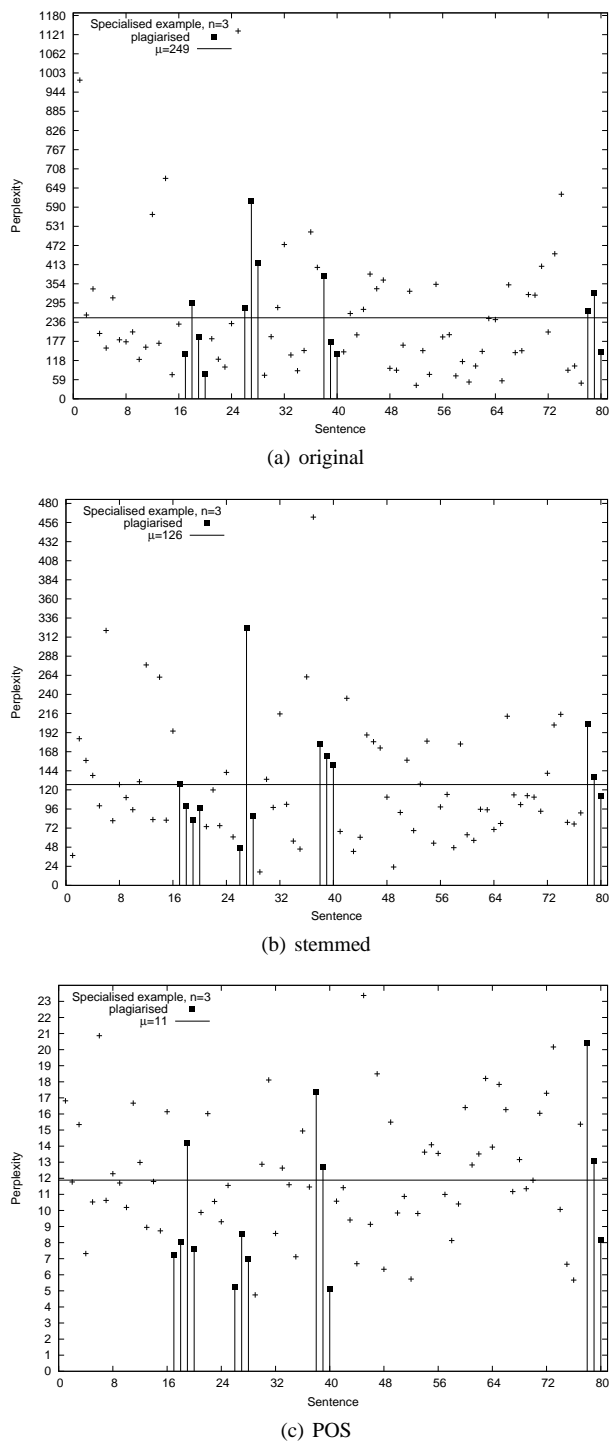


Figure 1. Perplexity on the specialised corpus (one point per sentence)

considered in the original text. The highest perplexities in this case are $PP_{37} = 462.78$ and $PP_{27} = 323.46$. Sentence S_{37} is a fragment copied by the author from another text in order to analyse it and, therefore, the result can be considered correct. Sentence S_{27} , as we have already said is plagiarised.

Finally, in the part-of-speech version of the text (c) the vocabulary is clearly smaller than in the other two cases (around 40 words given by the grammar categories⁵), resulting in a perplexity range much lower. In this case the three highest perplexities are $PP_{45} = 23.36$, $PP_6 = 20.87$ and $PP_{78} = 20.40$. Between the twenty tokens in S_{45} , three are non-frequent strings conformed by parenthesis and cardinal numbers, for example (2) which is tagged like (LS) (list item). S_{78} is plagiarised and contains the 3-gram *DT NN IN*, which is the third trigram with smaller probability and others that have not appeared in the training corpus, which is the case of 3-grams *RB VVZ DT* and *DT RBR JJ*⁶ and, therefore, their probability tends to 0.

These experiments have been carried out considering a small corpus. In Section 4.2 we describe the experiments we have carried out over a bigger corpus and, for this reason, a richer LM.

4.2 Experiments over general literature texts

In order to have a reference for our results, we have made the same experiments over a literary corpus. For these experiments we have taken a set of books written by the author Lewis Carroll and some passages from William Shakespeare texts to "plagiarise" the test section of Carroll's corpus⁷. The distribution of the training and test subcorpora is described in Table 1.

Table 1. Literary corpus

Author	Subcorpus	$ w $
Carroll	training	116,202
Carroll	test	26,626
Shakespeare	plagiarised	103

We have done the same pre-process, described at the Section 4.1, to the training and test corpora in order to obtain original, part-of-speech and stem versions of the texts. Figure 2 shows the results over the three versions of the test corpus. In this case we can see that the plagiarised sections, in general, obtain high values of perplexity with respect to non-plagiarised segments in POS and stem versions.

For example, in the case of the original text, the sentence with the highest perplexity, as it appears in the text, is "ALL PERSONS MORE THAN A MILE HIGH TO LEAVE THE COURT.". The words in bold have not appeared in the training corpus (at least with all the letters capitalised)⁸.

In the other two cases, the POS and stem versions, the reason for most of the cases is simple: there are errors in the part-of-speech and stems generated by the tagger (in some cases it is due to errors in the text). Let us consider the stemmed version of the test corpus to show some examples. Table 2 includes the sentences with the highest perplexities in the Carroll's plagiarised document.

⁵ See <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps>

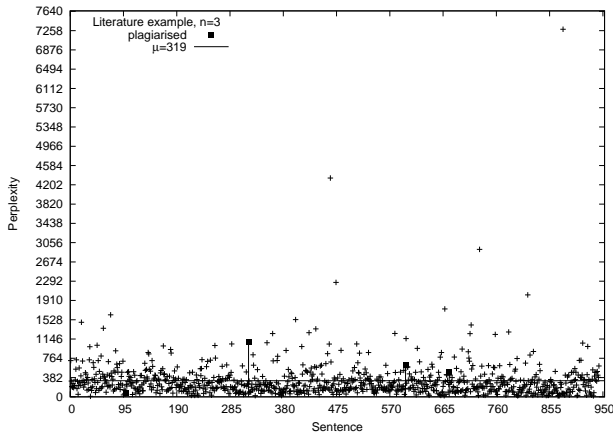
⁶ DT=determiner; NN=noun; IN=preposition; RB=adverb; VVZ=verb; RBR=comparative adverb; JJ=adjective.

⁷ Texts have been downloaded from Project Gutenberg, <http://www.gutenberg.org>

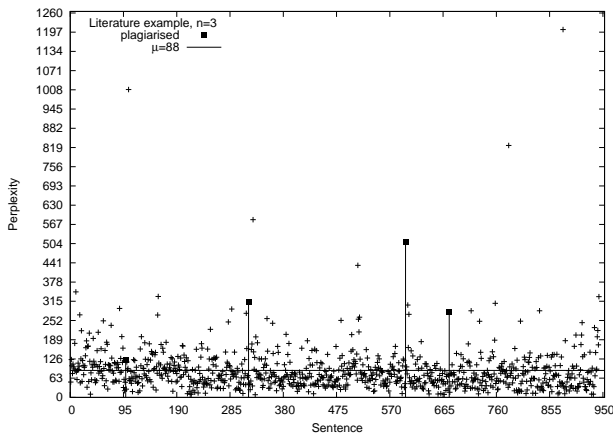
⁸ These kinds of "errors" could be avoided converting all the characters to lower case during the pre-processing of the corpus.

Table 2. Sentences with highest perplexities

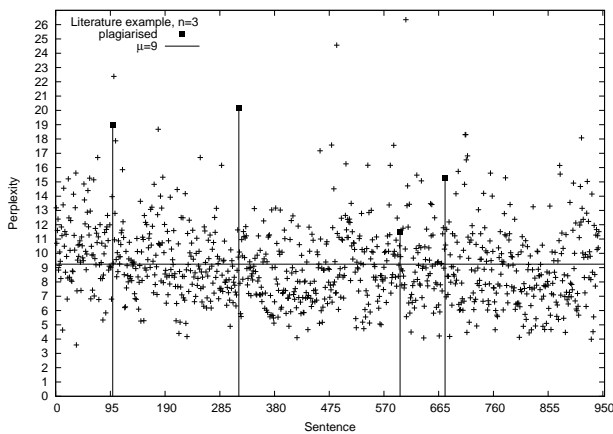
Perplexity	Sentence
1205.6	all Persons more Than A Mile High TO leave the court.
1009.1	William 's conduct at first be moderate.
825.6	the twelve juror be all write very busily on slate.
582.5	' oh , there go his precious nose ' ; as an unusually large saucepan fly close by it , and very nearly carry it off.
508.1	the hearing of my wife , with your approach : so humbly take my



(a) original



(b) stemmed



(c) POS

Figure 2. Perplexity on the literature corpus (one point per sentence)

The first sentence contains proper nouns (capitalised words *Persons*, *Than* and *Mile* were all considered proper nouns by the tagger), which are hard to occur in different texts and, in this case, have $P(w) \rightarrow 0$ because they do not appear in the LM. This is one of the weaknesses of the original and stemmed versions of texts: due to the fact that they use an open domain language, it is difficult that a LM contains all the "strange words" as it is the case of proper names and other "special" words that are commonly included in texts.

All the words in the second sentence were included in the training corpus and, therefore, the vocabulary in the two cases is not different. However, *William* never appeared at the beginning of a sentence and the trigram *William 's conduct* neither, just to give a couple examples. In the case of the third sentence, it contains the word *juror*, that the LM ignores, and *busily*, that has a low probability: $P('busily') = 0.0000191$ (in order to compare, $P('the') = 0.03869$).

The first plagiarised line is the fifth one. The interesting fact here is that the LM knows all the words in this phrase, but the words and 3-grams in it have a low probability.

In the case of the part-of-speech version, the sentence with the highest perplexity has $PP_{608} = 26.34$, and it contains, for example, the substring *DT NN RBR* (determiner, noun, comparative adverb). This POS trigram corresponds to the segment of three tokens (*that*)₁ (*is* - "*The*")₂ (*more*)₃, which, due to an error in words split was not correctly tagged and the resulting POS trigram has a really low probability.

In this case, the first plagiarised sentence in the sorted list is in the fourth place with $PP_{318} = 20.132$. This sentence has style and vocabulary completely different from the others in the text and corresponds to the sentence "*Mac. We will proceed no further in this Businesse: He hath Honour'd me of late, and I haue bought Golden Opinions from all sorts of people, Which would be worne now in their newest glosse, Not cast aside so*", written by W. Shakespeare.

4.3 Discussion

From the five categories of stylometric features useful for the plagiarism detection task [9], our LM approach considers just three of them. *Syntactic features* and *special words counting*, "which measure writing style at the sentence-level" [9] and vocabulary richness respectively, are considered with the perplexity calculation of the sentences over the original and stemmed test corpora. *Part-of-speech classes quantification* is implicitly considered with the POS version. The only two features that our approach does not consider are *text statistics* (at character level) and *structural features*, that deal with the organisation of the text.

It can be seen in the results of the experiments in Sections 4.1 and 4.2 that considering only the perplexity of a sentence is not good enough to discriminate it from a plagiarised or "legal" text fragment.

The perplexity calculations over the three versions of the text (original, POS and stem) have conducted to the detection of "non-

expected“ sentences that, in the most of the cases, include those that have been plagiarised. However, these three experiments do not detect the same sentences, but different ones, so we believe that we need to consider the three versions together in order to detect plagiarised sentences.

5 CONCLUSIONS AND FUTURE WORK

In this paper we have explored the utility of Language Models and perplexity, a measure to determine the coverage of a Language Model given a text, for the Automatic Detection of Plagiarism with a reference corpus. We have considered perplexity on three different levels: word, part-of-speech and stem.

In order to do that, we have calculated a Language Model over a reference corpus, written by one only author, and calculated perplexity of sentences on a test corpus (which contained plagiarised fragments) based on this model.

Our main hypothesis was that those segments with the highest perplexities with respect to the model, should be the plagiarised ones. Unfortunately, our hypothesis is not completely true because there are non-plagiarised fragments (in particular those with “strange segments” such as titles and bibliographic cites) that present high perplexity. However, plagiarised segments seem to stand out in the highest positions when we consider these features.

In the results that we have obtained, we have noted that in order to identify good candidates for plagiarised segments we should consider the three versions of the analysed text together (original, POS and stem).

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.

ACKNOWLEDGEMENTS

We would like to thank the MCyT TIN2006-15265-C06-04 research project for partially funding this work and the National Council for Science and Technology (CONACYT-MEXICO) for funding the research work of the first author (Id. 192021).

REFERENCES

- [1] Thomas M. Breuel, ‘The ocrpus open source ocr system’, *Proc. IS&T/SPIE 20th Annual Symposium 2008*, (2008).
- [2] Rosa Maria Coyotl-Morales, Luis Villaseñor Pineda, Manuel Montesy Gómez, and Paolo Rosso, ‘Authorship attribution using word sequences’, *Proc. of the 11th Iberoamerican Congress on Pattern Recognition, (CIARP 2006)*, **LNCS (4225)**, 844–853, (2006).
- [3] Josep M. Crego, Jos B. Mariño, and Adriá de Gispert, ‘An n-gram-based statistical machine translation decoder’, *Interspeech’2005 - Eurospeech, Lisbon, Portugal*, (2005).
- [4] Georgia Frantzeskou, Efsthios Stamatatos, and Stefanos Gritzalis, ‘Identifying authorship by byte-level n-grams: The source code author profile (scap) method’, *Int. Journal of Digital Evidence*, **6**(1), (2007).
- [5] Djoerd Hiemstra, ‘A linguistically motivated probabilistic model of information retrieval’, *Proc. of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, (ECDL 1998)*, **LNCS (1513)**, 569–584, (1998).
- [6] Parvati Iyer and Abhispita Singh, ‘Document similarity analysis for a plagiarism detection system’, *2nd Indian Int. Conf. on Artificial Intelligence (IICAI-2005)*, 2534–2544, (2005).
- [7] Frederick Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, Massachusetts, 1997.
- [8] Christopher D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press Publisher, Cambridge Massachusetts and London, England, 2000.
- [9] Sven Meyer zu Eissen and Benno Stein, ‘Intrinsic plagiarism detection’, *Lalmas et. al. (Eds.): Advances in Information Retrieval Proc. of the 28th European Conf. on IR research, ECIR 2006, London*, 565–569, (2006).
- [10] Fuchim Peng, Dale Schuurmans, Vlado Keselj, and Shaojun Wang, ‘Automated authorship attribution with character level language models’, *Proc. of the 10th Conf. of the European Chapter of the Association for Computational Linguistics (EACL 2003), Budapest, Hungary*, (2003).
- [11] Jay M. Ponte and W. Bruce Croft, ‘A language modeling approach to information retrieval’, *Croft, Moffat, van Rijsbergen, Wilkinson, and Zobel, Eds., 21st Annual Int. ACM SIGIR Conf.*, 275–281, (1998).
- [12] Verónica Romero, Vicente Alabau, and José-Miguel Benedí, ‘Combination of n-grams and stochastic context-free grammars in an offline handwritten recognition system’, *3rd Iberian Conf. on Pattern Recognition and Image Analysis, Springer-Verlag, Girona, Spain*, **LNCS (4477)**, 467–474, (2007).
- [13] Helmut Schmid, ‘Probabilistic part-of-speech tagging using decision trees’, *Proc. of the Int. Conf. on New Methods in Language Processing*, (1994).
- [14] Antonio Si, Hong Va Leong, and Rynson W. H. Lau, ‘Check: a document plagiarism detection system’, *Proc. of the 1997 ACM Symposium on Applied Computing, San Jose, California, United States*, 70–77, (1997).
- [15] Benno Stein, ‘Principles of hash-based text retrieval’, *Clarke, Fuhr, Kando, Kraaij, and de Vries, Eds., 30th Annual Int. ACM SIGIR Conf.*, 527–534, (2007).
- [16] Benno Stein and Sven Meyer zu Eissen, ‘Intrinsic plagiarism analysis with meta learning’, *B. Stein, M. Koppel, and E. Stamatatos, Eds., SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07), Amsterdam, Netherlands*, 45–50, (2007).
- [17] Richard Zens and Hermann Ney, ‘N-gram posterior probabilities for statistical machine translation’, *Human Language Technology Conf. of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proc. of the Workshop on Statistical Machine Translation, New York City, NY, United States*, (2006).