Corpus and Evaluation Measures for Automatic Plagiarism Detection

Alberto Barrón-Cedeño[1], Martin Potthast[2], Paolo Rosso[1], Benno Stein[2]

[1]Natural Language Engineering Lab — ELiRF
Department of Information Systems and Computation
Universidad Politécnica de Valencia, Spain
http://users.dsic.upv.es/grupos/nle
{lbarron, prosso}@dsic.upv.es

[2] Web Technology and Information Systems
Faculty of Media / Media Systems
Bauhaus-Universität Weimar, Germany
http://www.webis.de
{martin.potthast, benno.stein}@uni-weimar.de

Abstract

The simple access to texts on digital libraries and the WWW has led to an increased number of plagiarism cases in recent years, which renders manual plagiarism detection infeasible at large. Various methods for automatic plagiarism detection have been developed whose objective is to assist human experts to analyze documents for plagiarism. Unlike other tasks in natural language processing and information retrieval, it is not possible to publish a collection of real plagiarism cases for evaluation purposes since they cannot be properly anonymized. Therefore, current evaluations found in the literature are incomparable and often not even reproducible. Our contribution in this respect is a newly developed large-scale corpus of *artificial* plagiarism and new detection performance measures tailored to the evaluation of plagiarism detection algorithms.

1. Introduction

Plagiarism is the use of text written by a third party in one's own writing without permission or acknowledgment (Clough, 2000). The goal of automatic plagiarism detection is to identify the plagiarized sections in a suspicious document d_q. Two approaches exist to tackle this task: *intrinsic* plagiarism detection and *external* plagiarism detection.

In intrinsic plagiarism detection, features that indicate writing style are used to detect style irregularities caused by the insertion of text from a different author into d_q. The writing style of a text can be quantified, by measuring a text's readability, vocabulary richness, or by the use of basic statistics, such as the average sentence length and the average word length (Meyer zu Eißen and Stein, 2006). Other approaches apply character n-grams profiles to characterize an author's style and search for irregularities across d_q (Stamatatos, 2009).

External plagiarism detection has attracted more attention because of its close relation to information retrieval. A document d_q and a collection of potential source documents D are given, and the task is to identify the plagiarized sections in d_q (if there are any), and their respective source sections in D (Potthast et al., 2009). Two issues render this task difficult: the number of potential source documents, $|D|$, and the fact that plagiarizing a text often includes paraphrasing, summarizing, and sometimes even translation.

To deal with these problems it has been proposed to compile D into a fingerprint index using text fingerprinting schemes, such as SPEX (Bernstein and Zobel, 2004) and Winnowing (Schleimer et al., 2003). The index then can be queried with d_q's fingerprint in order to retrieve documents from D with overlapping or near-duplicate contents. However, since D is often considered to be the whole Web, a more practical solution is the automatic retrieval of a small number of candidate source documents using a Web search engine, and to compare them with d_q on the basis of the vector space model (Broder, 1997; Maurer et al., 2006).

We observe in this connection that the evaluation of plagiarism detection algorithms is not standardized, i.e., most of the time the algorithms are evaluated on homemade corpora using various different performance measures.[1] This situation renders the existing research almost incomparable. The contributions of the paper in hand address this problem. We propose:

1. a plagiarism corpus, called PAN-PC-09 (Section 2.)
2. tailored performance measures based on the idea of precision, recall, and granularity (Section 3.)

The corpus and the measures form the first controlled evaluation environment dedicated to plagiarism detection. They were used, among others, in the First International Competition on Plagiarism Detection (Section 4.). Final remarks are provided in Section 5.

This work has been partially supported by the TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 research project as well as the CONACyT-Mexico 192021 grant.

[1]To our knowledge, the only corpus which can be used to evaluate plagiarism detection is the METER corpus (Clough et al., 2002); it consists of a small number of annotated cases of text reuse in news articles but is not specifically designed to support plagiarism detection evaluation.

2. The PAN-PC-09 Plagiarism Corpus

The PAN Plagiarism Corpus (PAN-PC-09) is a collection of 41 223 documents in which 94 202 artificial plagiarism cases have been inserted.[2] It is the first corpus which allows for large-scale evaluations of both intrinsic and external plagiarism detection methods. During its construction a number of parameters have been varied so that the corpus features a wide cross-section of different plagiarism cases:

- *Document Length.* 50% of the documents are small (1-10 pages), 35% medium (10-100 pages), and 15% large (100-1,000 pages).
- *Suspicious-to-Source Ratio.* 50% of the documents are designated as suspicious documents D_q, and 50% as source documents D.
- *Plagiarism Percentage.* The percentage of plagiarism per suspicious document $d_q \in D_q$ ranges from 0% to 100% (cf. Figure 1). In order to compose a realistic framework, 50% of the suspicious documents contain no plagiarism at all.
- *Plagiarism Length.* The length of a plagiarism case is uniformly distributed between 50 and 5,000 words.
- *Plagiarism Languages.* 90% of the cases are monolingual English plagiarism. The remaining 10% are cross-language plagiarism, i.e., the source document is written either in German or in Spanish, and the plagiarism is translated into English.
- *Plagiarism Obfuscation.* The monolingual portion of the plagiarism in the external test corpus is obfuscated. The degree of obfuscation ranges uniformly from none to high.

With respect to plagiarism obfuscation further explanations are necessary. Plagiarists often paraphrase or summarize the text they plagiarize in order to *obfuscate* it, i.e., to hide their offense. A synthesizer, that simulates the obfuscation of a section of text s_x in order to generate a different text section s_q to be inserted into d_q, has been designed on the basis of the following basic operations:

- *Random Text Operations.* Given s_x, s_q is created by shuffling, removing, inserting, or replacing words or short phrases at random.
- *Semantic Word Variation.* Given s_x, s_q is created by replacing each word by one of its synonyms, hyponyms, hypernyms, or even antonyms.
- *POS-preserving Word Shuffling.* s_q is created by shuffling words while maintaining the original sequence of parts of speech in s_x.

These operations do not guarantee the generation of human-readable text. However, automatic text generation is still a largely unsolved problem which is why we have approached the task from the basic understanding of content similarity in information retrieval, namely the bag-of-words model.

[2] http://www.webis.de/research/corpora

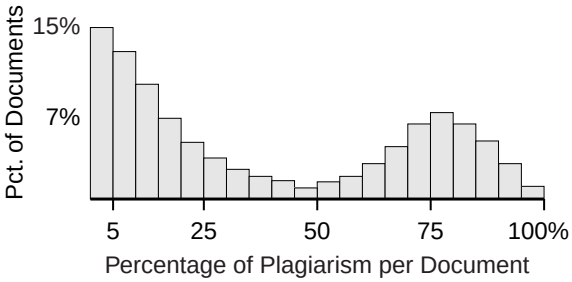

Figure 1: Distribution of plagiarism per document.

As mentioned before, a weakness of corpora containing actual cases of plagiarism is that they cannot be published due to ethical reasons. One of the aims of the PAN-PC-09 corpus was to avoid such weaknesses. Therefore, suspicious and source documents were generated on the basis of 22 874 documents from Project Gutenberg.[3] To the best of our knowledge, such documents are public domain. Hence, the PAN-PC-09 corpus is available free of charge for research purposes.

3. Evaluation Measures

In order to evaluate the performance of a plagiarism detection algorithm, precision and recall cannot be applied directly. For intrinsic plagiarism detection it is necessary to evaluate if a plagiarized section has been properly identified as such. Additionally, for external plagiarism detection it is also necessary to evaluate if the source section has been accurately retrieved.

Let d_q be a document including plagiarized sections; d_q defines a sequence of characters labeled as plagiarized or non-plagiarized. A plagiarized section s forms a contiguous sequence of characters in d_q. The set of all plagiarized sections in d_q is denoted by S; the plagiarized sections do not intersect, i.e., $\forall s_i, s_j \in S : i \neq j \rightarrow (s_i \cap s_j = \emptyset)$. Likewise, the set of all sections $r \subset d_q$ found by a plagiarism detection algorithm is denoted by R. See Figure 2 for an illustration.

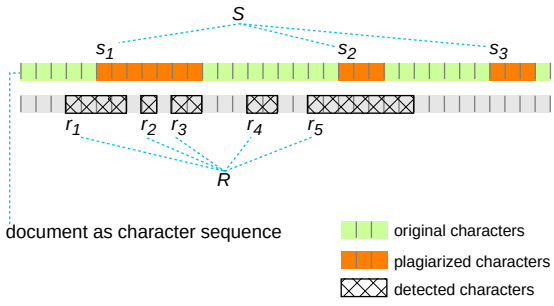

Figure 2: A document as character sequence, including plagiarized sections S and detections R returned by a plagiarism detection algorithm.

The plagiarized sections are treated as basic retrieval units. In this sense, each $s_i \in S$ defines a query for which a plagiarism detection algorithm returns a result set $R_i \subseteq R$. The recall of a plagiarism detection algorithm, rec_{PDA}, is

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

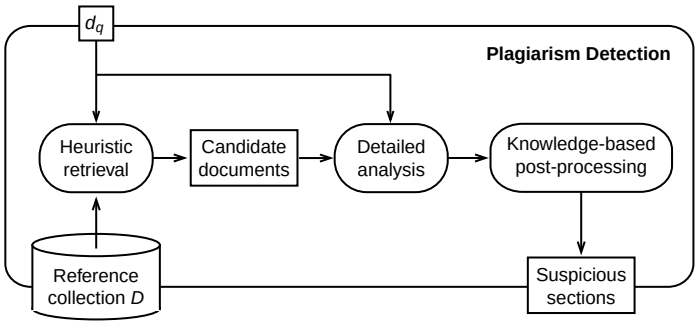

Figure 3: Retrieval process of external plagiarism detection, derived from (Stein et al., 2007).

then defined as the mean of the returned fractions of the plagiarized sections, averaged over all sections in S:

$$rec_{PDA}(S,R) = \frac{1}{|S|} \sum_{s \in S} \frac{|s \sqcap \bigcup_{r \in R} r|}{|s|} \quad ,$$

where $\sqcap$ computes the positionally overlapping characters. However, the precision of a plagiarism detection algorithm is not defined under this view, which is rooted in the fact that a detection algorithm does not return a unique result set for each plagiarized section $s \in S$ but for the whole of S. This deficit can be resolved by switching the reference basis. Instead of the plagiarized sections S, the algorithmically determined detections R become the targets: the precision with which the queries in S are answered is then measured as the recall of R under S. By computing the mean average over the $r \in R$ one obtains a definite computation rule that captures the concept of retrieval precision for S:

$$prec_{PDA}(S,R) = \frac{1}{|R|} \sum_{r \in R} \frac{|r \sqcap \bigcup_{s \in S} s|}{|r|} \quad .$$

rec_{PDA} and $prec_{PDA}$ are insensitive to the number of times an $s \in S$ is detected in a detection result R, i.e., the *granularity* of R. We define the granularity of R for a set of plagiarized sections S by the average size of the existing covers: a detection $r \in R$ belongs to the cover C_s of an $s \in S$ iff s and r overlap. Let $S_R \subseteq S$ denote the set of cases so that for each $s \in S : |C_s| > 0$. Then, the granularity of R given S is defined as:

$$gran_{PDA}(S,R) = \frac{1}{|S_R|} \sum_{s \in S_R} |C_s| \quad ,$$

where $C_s = \{r \mid r \in R \wedge s \cap r \neq \emptyset\}$ and $S_R = \{s \mid s \in S \wedge \exists r \in R : s \cap r \neq \emptyset\}$. The domain of the granularity is $[1, |R|]$, where 1 marks the desirable one-to-one correspondence between R and S, and $|R|$ marks the worst case, when a single $s \in S$ is detected over an over again.

In order to allow for an absolute ranking among plagiarism detection algorithms, the three measures are combined to an overall score:

$$overall_{PDA}(S,R) = \frac{F}{\log_2(1 + gran_{PDA})},$$

where F denotes the F-Measure, i.e., the harmonic mean of the precision $prec_{PDA}$ and the recall rec_{PDA}. We take the logarithm of the granularity to smooth its influence on the overall score.

4. Competition on Plagiarism Detection

The PAN-PC-09 corpus was formerly constructed for employment in the First International Competition on Plagiarism Detection (Potthast et al., 2009).[4] Here, 13 research teams from all over the world submitted detection results obtained with a variety of different plagiarism detection algorithms. 10 teams attempted to solve the external task, 4 teams competed in the intrinsic task, and one of them in both tasks.

In the intrinsic plagiarism detection task unexpected variations through a text were measured in order to determine whether a document contained plagiarized fragments. Such variations were analysed on the basis of character n-grams profiles (Stamatatos, 2009), word frequency class and text frequencies (Zechner et al., 2009), and Kolmogorov complexity measures (Seaward and Matwin, 2009). The former strategy was applied by the winner in this subtask.

For the external plagiarism detection task, all systems were based on common approaches, following the three-stage plagiarism detection process illustrated in Figure 3. The teams carried out the heuristic retrieval on the basis of the vector space model using character-16-grams (Grozea et al., 2009) or word-1-grams (Kasprzak et al., 2009), word-5-grams (Basile et al., 2009), or word-8-grams (Zechner et al., 2009). Only one team approached the task on the basis of fingerprinting techniques (Scherbinin and Butakov, 2009). The detailed analysis was carried out by searching for exact matches of character-n-grams (Grozea et al., 2009; Kasprzak et al., 2009), or word-n-grams (Basile et al., 2009), and sentences (Zechner et al., 2009). The winning approach of this task, as well as the overall winner of the competition was the team of Grozea et al. (2009).

Interestingly, no team tried to detect cross-language plagiarism; presumably because this type of plagiarism detection is still in its infancy and has attracted attention only recently (Potthast et al., 2010 in press).

[4] For further reading on the competition and the results obtained by the participants see `http://www.webis.de/research/workshopseries/pan-09/competition.html`

5. Conclusions

The PAN-PC-09 corpus is the first standardized corpus dedicated to the evaluation of automatic plagiarism detection and was successfully employed in the First International Competition on Plagiarism Detection. We believe that our corpus and the performance measures will become an effective means to evaluate future plagiarism detection research. Currently, an improved version of the corpus is being constructed. This corpus will be used in the Second International Competition on Plagiarism Detection, held in conjunction with the evaluation conference CLEF 2010.[5]

6. References

Chiara Basile, Dario Benedetto, Giampaolo Caglioti, and Mirko Degli Esposti. 2009. A Plagiarism Detection Procedure in Three Steps: Selection, Matches and Squares. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pan, 2009), pages 19–23.

Yaniv Bernstein and Justin Zobel. 2004. A Scalable System for Identifying Co-Derivative Documents. In *Proceedings of the Symposium on String Processing and Information Retrieval*, pages 55–67. Springer.

Andrei Z. Broder. 1997. On the Resemblance and Containment of Documents. In *Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29. IEEE Computer Society.

Paul Clough, Robert Gaizauskas, and Scott Piao. 2002. Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume V, pages 1678–1691, Las Palmas, Spain.

Paul Clough. 2000. Plagiarism in Natural and Programming Languages: an Overview of Current Tools and Technologies. Research Memoranda: CS-00-05, Department of Computer Science. University of Sheffield, UK.

Cristian Grozea, Christian Gehl, and Marius Popescu. 2009. ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pan, 2009), pages 10–18.

Jan Kasprzak, Michal Brandejs, and Miroslav Kriǐpač. 2009. Finding Plagiarism by Evaluating Document Similarities. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pan, 2009), pages 24–28.

Hermann Maurer, Frank Kappe, and Bilal Zaka. 2006. Plagiarism - A Survey. *Journal of Universal Computer Science*, 12(8):1050–1084.

Sven Meyer zu Eißen and Benno Stein. 2006. Intrinsic plagiarism detection. *Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research (ECIR 2006)*, LNCS (3936):565–569.

2009. San Sebastian, Spain. CEUS-WS.org.

Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. 2009. Overview of the 1st International Competition on Plagiarism Detection. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pan, 2009), pages 1–9.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2010 (in press). Cross-Language Plagiarism Detection.

Vladislav Scherbinin and Sergey Butakov. 2009. Using Microsoft SQL Server Platform for Plagiarism Detection. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pan, 2009), pages 36–37.

Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. 2003. Winnowing: Local Algorithms for Document Fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, New York, NY. ACM.

Leanne Seaward and Stan Matwin. 2009. Intrinsic Plagiarism Detection Using Complexity Analysis. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pan, 2009), pages 56–61.

Efstathios Stamatatos. 2009. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pan, 2009), pages 38–46.

Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. 2007. Strategies for Retrieving Plagiarized Documents. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen de Vries, editors, *30th Annual International ACM SIGIR Conference*, pages 825–826. ACM.

Michael Zechner, Markus Muhr, Roman Kern, and Michael Granitzer. 2009. External and Intrinsic Plagiarism Detection Using Vector Space Models. In *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (pan, 2009), pages 47–55.

[5]For further details, confer `http://pan.webis.de` as well as `http://www.clef2010.org`