Overview of the 1st International Competition on Plagiarism Detection[*]

Martin Potthast Benno Stein Andreas Eiselt

Web Technology & Information Systems Group
Bauhaus-Universität Weimar
<first name>.<last name>@uni-weimar.de

Alberto Barrón-Cedeño Paolo Rosso

Natural Language Engineering Lab, ELiRF
Universidad Politécnica de Valencia
{lbarron|prosso}@dsic.upv.es

Abstract: The 1st International Competition on Plagiarism Detection, held in conjunction with the 3rd PAN workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, brought together researchers from many disciplines around the exciting retrieval task of automatic plagiarism detection. The competition was divided into the subtasks external plagiarism detection and intrinsic plagiarism detection, which were tackled by 13 participating groups.

An important by-product of the competition is an evaluation framework for plagiarism detection, which consists of a large-scale plagiarism corpus and detection quality measures. The framework may serve as a unified test environment to compare future plagiarism detection research. In this paper we describe the corpus design and the quality measures, survey the detection approaches developed by the participants, and compile the achieved performance results of the competitors.

Keywords: Plagiarism Detection, Competition, Evaluation Framework

1 *Introduction*

Plagiarism and its automatic retrieval have attracted considerable attention from research and industry: various papers have been published on the topic, and many commercial software systems are being developed. However, when asked to name the best algorithm or the best system for plagiarism detection, hardly any evidence can be found to make an educated guess among the alternatives. One reason for this is that the research field of plagiarism detection lacks a controlled evaluation environment. This leads researchers to devise their own experimentation and methodologies, which are often not reproducible or comparable across papers. Furterhmore, it is unknown which detection quality can at least be expected from a plagiarism detection system.

To close this gap we have organized an international competition on plagiarism detection. We have set up, presumably for the first time, a controlled evaluation environment for plagiarism detection which consists of a large-scale corpus of artificial plagiarism and detection quality measures. In what follows we overview the corpus, the quality measures, the participants' detection approaches, and their obtained results.

1.1 Related Work

Research on plagiarism detection has been surveyed by Maurer, Kappe, and Zaka (2006) and Clough (2003). Particularly the latter provides well thought-out insights into, even today, "[...] *new challenges in automatic plagiarism detection*", among which the need for a standardized evaluation framework is already mentioned.

With respect to the evaluation of commercial plagiarism detection systems, Weber-Wulff and Köhler (2008) have conducted a manual evaluation: 31 handmade cases of plagiarism were submitted to 19 systems. The sources for the plagiarism cases were selected from the Web and the systems were judged by their capability to retrieve them. Due to the use of the Web, the experiment is not controlled which limits reproducibility, and since each case is only about two pages long there are concerns with respect to the study's representativeness. However, com-

[*] `http://www.webis.de/research/workshopseries/pan-09`

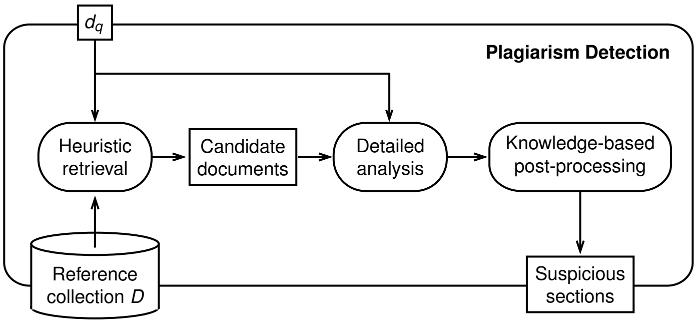

Figure 1: Generic retrieval process for external plagiarism detection.

mercial systems are usually not available for a close inspection which may leave no other choice to evaluate them.

1.2 Plagiarism Detection

The literature on the subject often puts plagiarism detection on a level with the identification of highly similar sections in texts or other objects. But this does not show the whole picture. From our point of view plagiarism detection divides into two major problem classes, namely external plagiarism detection and intrinsic plagiarism detection. Both of which include a number of subproblems and the frequently mentioned step-by-step comparison of two documents is only one of them.

For external plagiarism detection Stein, Meyer zu Eissen, and Potthast (2007) introduce a generic three-step retrieval process. The authors consider that the source of a plagiarism case may be hidden in a large reference collection, as well as that the detection results may not be perfectly accurate. Figure 1 illustrates this retrieval process. In fact, all detection approaches submitted by the competition participants can be explained in terms of these building blocks (cf. Section 4).

The process starts with a suspicious document d_q and a collection D of documents from which d_q's author may have plagiarized. Within a so-called heuristic retrieval step a small number of candidate documents D_x, which are likely to be sources for plagiarism, are retrieved from D. Note that D is usually very large, e.g., in the size of the Web, so that it is impractical to compare d_q one after the other with each document in D. Then, within a so-called detailed analysis step, d_q is compared section-wise with the retrieved candidates. All pairs of sections (s_q, s_x) with $s_q \in d_q$ and $s_x \in d_x$, $d_x \in D_x$, are to be retrieved such that s_q and s_x have a high similarity under some retrieval model. In a knowledge-based post-processing step those sections are filtered for which certain exclusion criteria hold, such as the use of proper citation or literal speech. The remaining suspicious sections are presented to a human, who may decide whether or not a plagiarism offense is given.

Intrinsic plagiarism detection has been studied in detail by Meyer zu Eissen and Stein (2006). In this setting one is given a suspicious document d_q but no reference collection D. Technology that tackles instances of this problem class resembles the human ability to spot potential cases of plagiarism just by reading d_q.

1.3 Competition Agenda

We have set up a large-scale corpus (D_q, D, S) of "artificial plagiarism" cases for the competition, where D_q is a collection of suspicious documents, D is a collection of source documents, and S is the set of annotations of all plagiarism cases between D_q and D. The competition divided into two tasks and into two phases for which the corpus was split up into 4 parts; one part for each combination of tasks and phases. For simplicity the sub-corpora are not denoted by different symbols.

Competition tasks and phases:

- *External Plagiarism Detection Task.* Given D_q and D the task is to identify the sections in D_q which are plagiarized, and their source sections in D.
- *Intrinsic Plagiarism Detection Task.* Given only D_q the task is to identify the plagiarized sections.

- *Training Phase.* Release of a training corpus (D_q, D, S) to allow for the development of a plagiarism detection system.
- *Competition Phase.* Release of a competition corpus (D_q, D) whose plagiarism cases were to be detected and submitted as detection annotations, R.

Participants were allowed to compete in either of the two tasks or both. After the competition phase the participants' detections were evaluated, and the winner of each task as well as an overall winner was determined as that participant whose detections R best matched S in the respective competition corpora.

2 Plagiarism Corpus

The PAN plagiarism corpus, PAN-PC-09, comprises 41 223 text documents in which 94 202 cases of artificial plagiarism have been inserted automatically (Webis at Bauhaus-Universität Weimar and NLEL at Universidad Politécnica de Valencia, 2009). The corpus is based on 22 874 book-length documents from the Project Gutenberg.[1] All documents are, to the best of our knowledge, public domain; therefore the corpus is available free of charge to other researchers. Important parameters of the corpus are the following:

- *Document Length.* 50% of the documents are small (1-10 pages), 35% medium (10-100 pages), and 15% large (100-1000 pages).
- *Suspicious-to-Source Ratio.* 50% of the documents are designated as suspicious documents D_q, and 50% are designated as source documents D (see Figure 2).
- *Plagiarism Percentage.* The percentage θ of plagiarism per suspicious document $d_q \in D_q$ ranges from 0% to 100%, whereas 50% of the suspicious documents contain no plagiarism at all. Figure 3 shows the distribution of the plagiarized documents for the external test corpus. For the intrinsic test corpus applies the hashed part of the distribution.
- *Plagiarism Length.* The length of a plagiarism case is evenly distributed between 50 words and 5000 words.

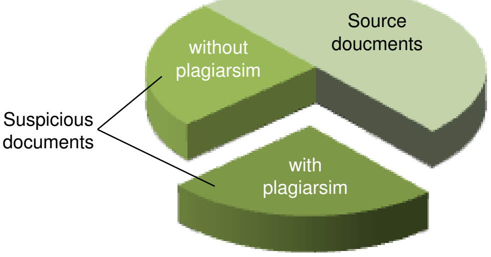

Figure 2: Distribution of suspicious documents (with and without plagiarism) and source documents.

- *Plagiarism Languages.* 90% of the cases are monolingual English plagiarism, the remainder of the cases are cross-lingual plagiarism which were translated automatically from German and Spanish to English.
- *Plagiarism Obfuscation.* The monolingual portion of the plagiarism in the external test corpus was obfuscated (cf. Section 2.1). The degree of obfuscation ranges evenly from none to high.

Note that for the estimation of the parameter distributions one cannot fall back on large case studies on real plagiarism cases. Hence, we decided to construct more simple cases than complex ones, where "simple" refers to short lengths, a small percentage θ, and less obfuscation. However, complex cases are overrepresented to allow for a better judgement whether a system detects them properly.

2.1 Obfuscation Synthesis

Plagiarists often modify or rewrite the sections they copy in order to obfuscate the plagiarism. In this respect, the automatic synthesis of plagiarism obfuscation we applied when constructing the corpus is of particular interest. The respective synthesis task reads

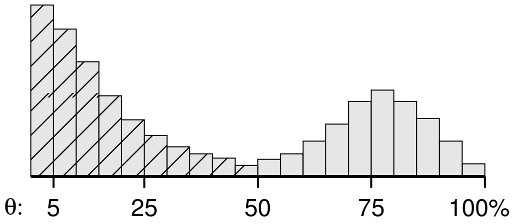

Figure 3: Distribution of the plagiarism percentage θ in the external test corpus. For the intrinsic test corpus applies the hashed part only.

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

as follows: given a section of text s_x, create a section s_q which has a high content similarity to s_x under some retrieval model but with a (substantially) different wording than s_x.

An optimal obfuscation synthesizer, i.e., an automatic plagiarist, takes an s_x and creates an s_q which is human-readable and which creates the same ideas in mind as s_x does when read by a human. Today, such a synthesizer cannot be constructed. Therefore, we approach the task from the basic understanding of content similarity in information retrieval, namely the bag-of-words model. By allowing our obfuscation synthesizers to construct texts which are not necessarily human-readable they can be greatly simplified. We have set up three heuristics to construct s_q from s_x:

- *Random Text Operations.* Given s_x, s_q is created by shuffling, removing, inserting, or replacing words or short phrases at random. Insertions and replacements are, for instance, taken from the document d_q, the new context of s_q.
- *Semantic Word Variation.* Given s_x, s_q is created by replacing each word by one of its synonyms, antonyms, hyponyms, or hypernyms, chosen at random. A word is retained if neither are available.
- *POS-preserving Word Shuffling.* Given s_x its sequence of parts of speech (POS) is determined. Then, s_q is created by shuffling words at random while the original POS sequence is maintained.

2.2 Critical Remarks

The corpus has been conceived and constructed only just in time for the competition so that there may still be errors in it. For instance, the participants pointed out that there are a number of unintended overlaps between unrelated documents. These accidental similarities do not occur frequently, so that an additional set of annotations solves this problem.

The obfuscation synthesizer based on random text operations produces anomalies in some of the obfuscated texts, such as sequences of punctuation marks and stop words. These issues were not entirely resolved so that it is possible to find some of the plagiarism cases by applying a kind of anomaly detection. Nevertheless, this was not observed during the competition.

Finally, by construction the corpus does not accurately simulate a heuristic retrieval situation in which the Web is used as reference collection. The source documents in the corpus do not resemble the Web appropriately. Note, however, that sampling the Web is also a problem for many ranking evaluation frameworks.

3 *Detection Quality Measures*

A measure that quantifies the performance of a plagiarism detection algorithm will resemble concepts in terms of precision and recall. However, these concepts cannot be transferred one-to-one from the classical information retrieval situation to plagiarism detection. This section explains the underlying connections and introduces a reasonable measure that accounts for the particularities.

Let d_q be a plagiarized document; d_q defines a sequence of characters each of which is either labeled as plagiarized or non-plagiarized. A plagiarized section s forms a contiguous sequence of plagiarized characters in d_q. The set of all plagiarized sections in d_q is denoted by S, where $\forall s_i, s_j \in S : i \neq j \rightarrow (s_i \cap s_j = \emptyset)$, i.e., the plagiarized sections do not intersect. Likewise, the set of all sections $r \subset d_q$ found by a plagiarism detection algorithm is denoted by R. See Figure 4 for an illustration.

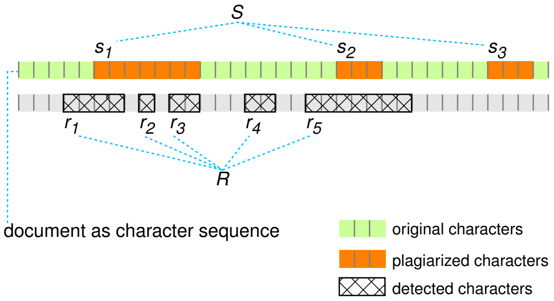

Figure 4: A document as character sequence, including plagiarized sections S and detections R returned by a plagiarism detection algorithm. The figure is drawn at scale $1 : n$ chars, $n \gg 1$.

If the *characters* in d_q are considered as basic retrieval units, precision and recall for a given $\langle d_q, S, R \rangle$ compute straightforwardly. This view may be called *micro-averaged* or system-oriented. For the situation shown in Figure 4 the micro-averaged precision is 8/16, likewise, the micro-averaged recall is 8/13. The advantage of a micro-averaged view is its clear computational semantics, which comes at a price: given an imbalance in the lengths of the elements in S—which usually correlates with the detection difficulty of a plagiarized section—the explanatory power of the computed measures is limited.

It is more natural to treat the contiguous sequences of plagiarized characters as basic retrieval units. In this sense each $s_i \in S$ defines a query q_i for which a plagiarism detection algorithm returns a result set $R_i \subseteq R$. This view may be called *macro-averaged* or user-oriented. The recall of a plagiarism detection algorithm, rec_{PDA}, is then defined as the mean of the returned fractions of the plagiarized sections, averaged over all sections in S:

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

where $\sqcap$ computes the positionally overlapping characters.

Problem 1. The *precision* of a plagiarism detection algorithm is not defined under the macro-averaged view, which is rooted in the fact that a detection algorithm does not return a unique result set for each plagiarized section $s \in S$. This deficit can be resolved by switching the reference basis. Instead of the plagiarized sections, S, the algorithmically determined sections, R, become the targets: the precision with which the queries in S are answered is identified with the recall of R under S.[2] 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:

$$prec_{PDA}(S,R) = \frac{1}{|R|} \sum_{r \in R} \frac{|r \sqcap \bigcup_{s \in S} s|}{|r|}, \qquad (2)$$

where $\sqcap$ computes the positionally overlapping characters. The domain of $prec_{PDA}$ is $[0,1]$; in particular it can be shown that this definition quantifies the necessary properties of a precision statistic.

Problem 2. Both the micro-averaged view and the macro-averaged view 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 $s \in S : |C_s| > 0$. The granularity of R given S is defined as follows:

$$gran_{PDA}(S,R) = \frac{1}{|S_R|} \sum_{s \in S_R} |C_s|, \qquad (3)$$

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

The measures (1), (2), and (3) 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}. To smooth the influence of the granularity on the overall score we take its logarithm.

4 Survey of Detection Approaches

For the competition, 13 participants developed plagiarism detection systems to tackle one or both of the tasks external plagiarism detection and intrinsic plagiarism detection. The questions that naturally arise: how do they work and how well? To give an answer, we survey the approaches in a unified way and report on their detection quality in the competition.

4.1 External Plagiarism Detection

Most of the participants competed in the external plagiarism detection task of the competition; detection results were submitted for 10 systems. As it turns out, all systems are based on common approaches—although they perform very differently.

As explained at the outset, external plagiarism detection divides into three steps (cf. Figure 1): the heuristic retrieval step, the detailed analysis step, and the post-processing step. Table 1 summarizes the participants' detection approaches in terms of these steps. However, the post-processing step was omitted here since neither of the participants applied noteworthy post-processing. Each row of the table summarizes one system; we restrict the survey to the top 6 systems since

[2] In (Stein, 2007) this idea is mathematically derived as "precision stress" and "recall stress".

Table 1: Unified summary of the detection approaches of the participants.

External Plagiarism Detection Approach		
Heuristic Retrieval	**Detailed Analysis**	**Participant**
Retrieval Model. Character-16-gram VSM (frequency weights, cosine similarity) *Comparison of D_q and D.* Exhaustive *Candidates $D_x \subset D$ for a d_q.* The 51 documents most similar to d_q.	*Exact Matches of d_q and $d_x \in D_x$.* Character-16-grams *Match Merging Heuristic to get (s_q, s_x).* Computation of the distances of adjacent matches. Joining of the matches based on a Monte Carlo optimization. Refinement of the obtained section pairs, e.g., by discarding too small sections.	Grozea, Gehl, and Popescu (2009)
Retrieval Model. Word-5-gram VSM (boolean weights, Jaccard similarity) *Comparison of D_q and D.* Exhaustive *Candidates $D_x \subset D$ for a d_q.* Documents which share at least 20 n-grams with d_q.	*Exact Matches of d_q and $d_x \in D_x$.* Word-5-grams *Match Merging Heuristic to get (s_q, s_x).* Extraction of the pairs of sections (s_q, s_x) of maximal size which share at least 20 matches, including the first and the last n-gram of s_q and s_x, and for which 2 adjacent matches are at most 49 not-matching n-grams apart.	Kasprzak, Brandejs, and Křipač (2009)
Retrieval Model. Word-8-gram VSM (frequency weights, custom distance) *Comparison of D_q and D.* Exhaustive *Candidates $D_x \subset D$ for a d_q.* The 10 documents nearest to d_q.	*Exact Matches of d_q and $d_x \in D_x$.* Word-8-grams *Match Merging Heuristic to get (s_q, s_x).* Extraction of the pairs of sections (s_q, s_x) which are obtained by greedily joining consecutive matches if their distance is not too high.	Basile et al. (2009)
Using the commercial system Plagiarism Detector (`http://plagiarism-detector.com`)		Palkovskii, Belov, and Muzika (2009)
Retrieval Model. Word-1-gram VSM (frequency weights, cosine similarity) *Comparison of D_q and D.* Clustering-based data-partitioning of D's sentences. Comparison of D_q's sentences with each partitions' centroid. *Candidates $D_x \subset D$ for a d_q.* For each sentence of d_q, the documents from the 2 most similar partitions which share similar sentences.	*Exact Matches of d_q and $d_x \in D_x$.* Sentences *Match Merging Heuristic to get (s_q, s_x).* Extraction of the pairs of sections (s_q, s_x) which are obtained by greedily joining consecutive sentences. Gaps are allowed if the respective sentences are similar to the corresponding sentences in the other document.	Muhr et al. (2009)
Retrieval Model. Winnowing fingerprinting 50 char chunks with 30 char overlap *Comparison of D_q and D.* Exhaustive *Candidates $D_x \subset D$ for a d_q.* Documents whose fingerprints share at least one value with d_q's fingerprint.	*Exact Matches of d_q and $d_x \in D_x$.* Fingerprint chunks *Match Merging Heuristic to get (s_q, s_x).* Extraction of the pairs of sections (s_q, s_x) which are obtained by enlarging matches and joining adjacent matches. Gaps must be below a certain Levenshtein distance.	Scherbinin and Butakov (2009)

the overall performance of the remaining systems is negligible. Nevertheless, these systems also implement the generic three-step process. The focus of this survey is on describing algorithmic and retrieval aspects rather than implementation details. The latter are diverse in terms of applied languages, software, and their runtime efficiency; descriptions can be found in the respective references.

The heuristic retrieval step (column 1 of Table 1) involves the comparison of the corpus' suspicious documents D_q with the source documents D. For this, each participant em-

Table 2: Performance results for the external plagiarism detection task.

			External Detection Quality			
Rank	**Overall**	**F**	**Precision**	**Recall**	**Granularity**	**Participant**
1	0.6957	0.6976	0.7418	0.6585	1.0038	Grozea, Gehl, and Popescu (2009)
2	0.6093	0.6192	0.5573	0.6967	1.0228	Kasprzak, Brandejs, and Křipač (2009)
3	0.6041	0.6491	0.6727	0.6272	1.1060	Basile et al. (2009)
4	0.3045	0.5286	0.6689	0.4370	2.3317	Palkovskii, Belov, and Muzika (2009)
5	0.1885	0.4603	0.6051	0.3714	4.4354	Muhr et al. (2009)
6	0.1422	0.6190	0.7473	0.5284	19.4327	Scherbinin and Butakov (2009)
7	0.0649	0.1736	0.6552	0.1001	5.3966	Pereira, Moreira, and Galante (2009)
8	0.0264	0.0265	0.0136	0.4586	1.0068	Vallés Balaguer (2009)
9	0.0187	0.0553	0.0290	0.6048	6.7780	Malcolm and Lane (2009)
10	0.0117	0.0226	0.3684	0.0116	2.8256	Allen (2009)

ploys a specific retrieval model, a comparison strategy, and a heuristic to select the candidate documents D_x from the D. Most of the participants use a variation of the well-known vector space model (VSM) as retrieval model, whereas, the tokens are often character- or word-n-grams instead of single words. As comparison strategy, the top 3 approaches perform an exhaustive comparison of D_q and D, i.e., each $d_q \in D_q$ is compared with each $d_x \in D$ in time $O(|D_q| \cdot |D|)$, while the remaining approaches employ data partitioning and space partitioning technologies to achieve lower runtime complexities. To select the candidate documents D_x for a d_q either its k nearest neighbors are selected or the documents which exceed a certain similarity threshold.

The detailed analysis step (column 2 of Table 1) involves the comparison of each $d_q \in D_q$ with its respective candidate documents D_x in order to extract pairs of sections (s_q, s_x), where $s_q \in d_q$ and $s_x \in d_x$, $d_x \in D_x$, from them which are highly similar, if any. For this, each participant first extracts all exact matches between d_q and d_x and then merges the matches heuristically to form suspicious sections (s_q, s_x). While each participant uses the same type of token to extract exact matches as his respective retrieval model of the heuristic retrieval step, the match merging heuristics differ largely from one another. However, it can be said that in most approaches a kind of distance between exact matches is measured first, and then a custom algorithm is employed which clusters them to sections.

Table 2 lists the detection performance results of all approaches, computed with the quality measures introduced in Section 3. Observe that the approach with top precision is the one on rank 6 which is based on fingerprinting, the approach with top recall is the one on rank 2, and the approach with top granularity is the one on rank 1. The latter is also the winner of this task since it provides the best trade off between the three quality measures.

4.2 Intrinsic Plagiarism Detection

The intrinsic plagiarism detection task has gathered less attention than external plagiarism detection; detection results were submitted for 4 systems. Table 3 lists their detection performance results. Unlike in external plagiarism detection, in this task the baseline performance is not 0. The reason for this is that intrinsic plagiarism detection is a one-

Table 3: Performance results for the intrinsic plagiarism detection task.

			Intrinsic Detection Quality				
Rank	**Overall**	**F**	**Precision**	**Recall**	**Granularity**	**Participant**	
1	0.2462	0.3086	0.2321	0.4607	1.3839	Stamatatos (2009)	
2	0.1955	0.1956	0.1091	0.9437	1.0007	Hagbi and Koppel (2009)	(Baseline)
3	0.1766	0.2286	0.1968	0.2724	1.4524	Muhr et al. (2009)	
4	0.1219	0.1750	0.1036	0.5630	1.7049	Seaward and Matwin (2009)	

Table 4: Overall plagiarism detection performance.

	Overall Detection Quality					
Rank	**Overall**	**F**	**Precision**	**Recall**	**Granularity**	**Participant**
1	0.4871	0.4884	0.5193	0.4610	1.0038	Grozea, Gehl, and Popescu (2009)
2	0.4265	0.4335	0.3901	0.4877	1.0228	Kasprzak, Brandejs, and Křipač (2009)
3	0.4229	0.4544	0.4709	0.4390	1.1060	Basile et al. (2009)
4	0.2131	0.3700	0.4682	0.3059	2.3317	Palkovskii, Belov, and Muzika (2009)
5	0.1833	0.4001	0.4826	0.3417	3.5405	Muhr et al. (2009)
6	0.0996	0.4333	0.5231	0.3699	19.4327	Scherbinin and Butakov (2009)
7	0.0739	0.0926	0.0696	0.1382	1.3839	Stamatatos (2009)
8	0.0586	0.0587	0.0327	0.2831	1.0007	Hagbi and Koppel (2009)
9	0.0454	0.1216	0.4586	0.0701	5.3966	Pereira, Moreira, and Galante (2009)
10	0.0366	0.0525	0.0311	0.1689	1.7049	Seaward and Matwin (2009)
11	0.0184	0.0185	0.0095	0.3210	1.0068	Vallés Balaguer (2009)
12	0.0131	0.0387	0.0203	0.4234	6.7780	Malcolm and Lane (2009)
13	0.0081	0.0157	0.2579	0.0081	2.8256	Allen (2009)

class classification problem in which it has to be decided for each section of a document whether it is plagiarized, or not. The baseline performance in such problems is commonly computed as the naive assumption that everything belongs to the target class, which is also what Hagbi and Koppel (2009) did who classified almost everything as plagiarized. Interestingly, the baseline approach is on rank 2 while two approaches perform worse than the baseline. Only the approach of Stamatatos (2009) performs better than the baseline.

4.3 Overall Detection Results

To determine the overall winner of the competition, we have computed the combined detection performance of each participant on the competition corpora of both tasks. Table 4 shows the results. Note that the competition corpus of the external plagiarism detection task is a lot bigger than the one for the intrinsic plagiarism detection task, which is why the top ranked approaches are those who performed best in the former task. Overall winner of the competition is the approach of Grozea, Gehl, and Popescu (2009).

5 *Summary*

The 1st International Competition on Plagiarism Detection fostered research and brought a number of new insights into the problems of automatic plagiarism detection and its evaluation. An important by-product of the competition is a controlled large-scale evaluation framework which consists of a corpus of artificial plagiarism cases and new detection quality measures. The corpus contains more than 40 000 documents and about 94 000 cases of plagiarism.

Furthermore, in this paper we give a comprehensive overview about the competition and in particular about the plagiarism detection approaches of the competition's 13 participants. It turns out that all of the detection approaches follow a generic retrieval process scheme which consists of the three steps heuristic retrieval, detailed analysis, and knowledge-based post-processing. To ascertain this fact we have compiled a unified summary of the top approaches in Table 1.

The competition divided into the two tasks external plagiarism detection and intrinsic plagiarism detection. The winning approach for the former task achieves 0.74 precision at 0.65 recall at 1.00 granularity. The winning approach for the latter task improves 26% above the baseline approach and achieves 0.23 precision at 0.46 recall at 1.38 granularity.

Acknowledgements

We thank Yahoo! Research and the University of the Basque Country for their sponsorship. This work was also partially funded by the Text-Enterprise 2.0 TIN2009-13391-C04-03 project and the CONACYT-MEXICO 192021 grant. Our special thanks go to the participants of the competition for their devoted work.

References

Allen, James. 2009. Submission to the 1st International Competition on Plagiarism Detection. From the Southern Methodist University in Dallas, USA.

Basile, Chiara, Dario Benedetto, Emanuele Caglioti, Giampaolo Cristadoro, and Mirko Degli Esposti. 2009. A Plagiarism Detection Procedure in Three Steps: Selection, Matches and "Squares". In Stein et al. (Stein et al., 2009).

Clough, Paul. 2003. Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service, `http://ir.shef.ac.uk/cloughie/papers/pas_plagiarism.pdf`.

Grozea, Cristian, Christian Gehl, and Marius Popescu. 2009. ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In Stein et al. (Stein et al., 2009).

Hagbi, Barak and Moshe Koppel. 2009. Submission to the 1st International Competition on Plagiarism Detection. From the Bar Ilan University, Israel.

Kasprzak, Jan, Michal Brandejs, and Miroslav Křipač. 2009. Finding Plagiarism by Evaluating Document Similarities. In Stein et al. (Stein et al., 2009).

Malcolm, James A. and Peter C. R. Lane. 2009. Tackling the PAN'09 External Plagiarism Detection Corpus with a Desktop Plagiarism Detector. In Stein et al. (Stein et al., 2009).

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

Meyer zu Eissen, Sven and Benno Stein. 2006. Intrinsic plagiarism detection. In Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, *Proceedings of the European Conference on Information Retrieval (ECIR 2006)*, volume 3936 of *Lecture Notes in Computer Science*, pages 565–569. Springer.

Muhr, Markus, Mario Zechner, Roman Kern, and Michael Granitzer. 2009. External and Intrinsic Plagiarism Detection Using Vector Space Models. In Stein et al. (Stein et al., 2009).

Palkovskii, Yurii Anatol'yevich, Alexei Vitalievich Belov, and Irina Alexandrovna Muzika. 2009. Submission to the 1st International Competition on Plagiarism Detection. From the Zhytomyr State University, Ukraine.

Pereira, Rafael C., V. P. Moreira, and R. Galante. 2009. Submission to the 1st International Competition on Plagiarism Detection. From the Universidade Federal do Rio Grande do Sul, Brazil.

Scherbinin, Vladislav and Sergey Butakov. 2009. Using Microsoft SQL Server Platform for Plagiarism Detection. In Stein et al. (Stein et al., 2009).

Seaward, Leanne and Stan Matwin. 2009. Intrinsic Plagiarism Detection Using Complexity Analysis. In Stein et al. (Stein et al., 2009).

Stamatatos, Efstathios. 2009. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In Stein et al. (Stein et al., 2009).

Stein, Benno. 2007. Principles of hash-based text retrieval. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen de Vries, editors, *30th Annual International ACM SIGIR Conference*, pages 527–534. ACM, July.

Stein, Benno, 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, July.

Stein, Benno, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors. 2009. *Proceedings of the SEPLN Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, PAN'09, September 10 2009, Donostia-San Sebastián, Spain.* Universidad Polytécnica de Valencia.

Vallés Balaguer, Enrique. 2009. Putting Ourselves in SME's Shoes: Automatic Detection of Plagiarism by the WCopyFind tool. In Stein et al. (Stein et al., 2009).

Weber-Wulff, Debora and Katrin Köhler. 2008. Plagiarism detection softwaretest 2008. `http://plagiat.htw-berlin.de/software/2008/`.

Webis at Bauhaus-Universität Weimar and NLEL at Universidad Politécnica de Valencia. 2009. PAN Plagiarism Corpus PAN-PC-09. `http://www.webis.de/research/corpora`. Martin Potthast, Andreas Eiselt, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso (editors).