Text Reuse Detection Using a Composition of Text Similarity Measures

Daniel Bär[1] *Torsten Zesch*[1,2] *Iryna Gurevych*[1,2]
(1) Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
(2) Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
www.ukp.tu-darmstadt.de

ABSTRACT

Detecting text reuse is a fundamental requirement for a variety of tasks and applications, ranging from journalistic text reuse to plagiarism detection. Text reuse is traditionally detected by computing similarity between a source text and a possibly reused text. However, existing text similarity measures exhibit a major limitation: They compute similarity only on features which can be derived from the *content* of the given texts, thereby inherently implying that any other text characteristics are negligible. In this paper, we overcome this traditional limitation and compute similarity along three characteristic dimensions inherent to texts: *content*, *structure*, and *style*. We explore and discuss possible combinations of measures along these dimensions, and our results demonstrate that the composition consistently outperforms previous approaches on three standard evaluation datasets, and that text reuse detection greatly benefits from incorporating a diverse feature set that reflects a wide variety of text characteristics.

TITLE AND ABSTRACT IN GERMAN

Erkennung von Textwiederverwendung durch Komposition von Textähnlichkeitsmaßen

Die Frage, ob und in welcher Weise Texte in abgewandelter Form wiederverwendet werden, ist ein zentraler Aspekt bei einer Reihe von Problemstellungen, etwa im Rahmen journalistischer Tätigkeit oder als Mittel zur Plagiatserkennung. Textwiederverwendung wird traditionell ermittelt durch Berechnen von Textähnlichkeit zwischen einem Ursprungstext und einem potentiell wiederverwendeten Text. Bestehende Textähnlichkeitsmaße haben jedoch die starke Einschränkung, dass sie Ähnlichkeit nur anhand von Eigenschaften berechnen, die vom *Inhalt* der gegebenen Texte abgeleitet werden können, und somit implizieren, dass jegliche andere Textcharacteristika vernächlässigbar sind. In dieser Arbeit berechnen wir Textähnlichkeit anhand von drei Dimensionen: *Inhalt*, *Struktur* und *Stil*. Wir untersuchen mögliche Kombinationen von Maßen entlang dieser Dimensionen, und zeigen deutlich anhand der Ergebnisse auf drei etablierten Evaluationsdatensätzen, dass die Komposition generell bessere Ergebnisse liefert als bestehende Ansätze, und dass die Bestimmung von Textwiederverwendung stark von einem breiten Spektrum an Textcharacteristika profitiert.

KEYWORDS: text similarity, text reuse, plagiarism, paraphrase.

KEYWORDS IN GERMAN: Textähnlichkeit, Textwiederverwendung, Plagiat, Paraphrase.

1 Introduction

Text reuse is a common phenomenon and arises, for example, on the Web from mirroring texts on different sites or reusing texts in public blogs. In other text collections such as content authoring systems of communities or enterprises, text reuse arises from keeping multiple versions, copies containing customizations or reformulations, or the use of template texts (Broder et al., 1997).

Problems with text reuse particularly arise in settings where systems are extensively used in a collaborative manner. For example, *wikis* are web-based, collaborative content authoring systems which offer fast and simple means for adding and editing content (Leuf and Cunningham, 2001). At any time, users can modify content already present in the wiki, augment existing texts with new facts, ideas, or thoughts, or create new texts from scratch. However, when users contribute to wikis, they need to avoid content duplication. This requires comprehensive knowledge of what content is already present in the wiki, and what is not. As wikis are traditionally growing fast, this is hardly feasible, though. To remedy this issue, we aim at supporting authors of collaborative text collections by means of automatic text reuse detection. We envision a semi-supervised system that informs a content author of potentially pre-existing instances of text reuse, and then lets the author decide how to proceed, e.g. to merge both texts.

Detecting text reuse has been studied in a variety of tasks and applications, e.g. the detection of journalistic text reuse (Clough et al., 2002), the identification of rewrite sources for ancient literary texts (Lee, 2007), or the analysis of text reuse in blogs and web pages (Abdel-Hamid et al., 2009). Another common instance of text reuse is plagiarism, with the additional constraint that the reuse needs to be unacknowledged. Near-duplicate detection is also a broad field of related work where the detection of text reuse is crucial, e.g. in the context of web search and crawling (Hoad and Zobel, 2003; Henzinger, 2006; Manku et al., 2007). Prior work, however, mainly utilizes fingerprinting and hashing techniques (Charikar, 2002) for text comparison rather than methods from natural language processing.

A common approach to text reuse detection is to compute similarity between a source text and a possibly reused text. A multitude of text similarity measures have been proposed for computing similarity based on surface-level and/or semantic features (Mihalcea et al., 2006; Landauer et al., 1998; Gabrilovich and Markovitch, 2007). However, existing similarity measures typically exhibit a major limitation: They compute similarity only on features which can be derived from the *content* of the given texts. By following this approach, they inherently imply that the similarity computation process does not need to take any other text characteristics into account.

In contrast, we propose that text reuse detection indeed benefits from also assessing similarity along other text characteristics (*dimensions*, henceforth). We follow empirical evidence by Bär et al. (2011) and focus on three characteristic similarity dimensions inherent to texts: *content*, *structure*, and *style*. Figure 1 shows an example of text reuse taken from the Wikipedia Rewrite Corpus (see Section 3.1) where parts of a given source text have been reused either verbatim or by using similar words or phrases. As the example illustrates, the process of creating reused text includes a revision step in which the editor has a certain degree of freedom on how to reuse the source text. This kind of similarity is detectable by content-centric text similarity measures. However, the editor has further split the source text into two individual sentences and changed the order of the reused parts. For detecting the degree of similarity of such a revision, text similarity measures for *structural similarity* are necessary. Additionally, the given texts exhibit a certain degree of similarity with respect to stylistic features, e.g. vocabulary richness.[1] In

[1]The type-token ratio (Templin, 1957) of the texts is .79 and .71, respectively.

Source Text. *PageRank is a link analysis algorithm used by the Google Internet search engine that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set.*

Text Reuse. *The PageRank algorithm is used to designate every aspect of a set of hyperlinked documents with a numerical weighting. It is used by the Google search engine to estimate the relative importance of a web page according to this weighting.*

Figure 1: Example of text reuse taken from the Wikipedia Rewrite Corpus (Clough and Stevenson, 2011). Various parts of the source text have been reused, either verbatim (underlined) or using similar words or phrases (wavy underlined). However, the editor has split the source text into two individual sentences and changed the order of the reused parts.

order to use such features as indicators of text reuse, we propose to further include measures of *stylistic similarity*.

In this paper, we thus overcome the traditional limitation of text similarity measures to *content* features. In contrast, we adopt ideas of seminal studies by cognitive scientists (Tversky, 1977; Goodman, 1972; Gärdenfors, 2000) and discuss the role of three similarity dimensions for the task of text reuse detection: *content*, *structure*, and *style*, as proposed in our previous work (Bär et al., 2011). In Section 2, we report on a multitude of text similarity measures from these dimensions that we used for our experiments. In Section 3, we demonstrate empirically that text reuse can be best detected if measures are combined across dimensions, so that a wide variety of text characteristics are taken into consideration. Our approach consistently outperforms previous work on three standard evaluation datasets, and demonstrates the advantage of integrating text characteristics other than *content* into the similarity computation process.

2 Text Similarity Measures

In this section, we report on a variety of similarity measures which we used to compute similarity along characteristic dimensions inherent to texts.[2] We classify them into measures for *content similarity*, *structural similarity*, and *stylistic similarity*, as proposed by Bär et al. (2011).

2.1 Content Similarity

Probably the easiest way to reuse text is verbatim copying. It can be detected by using string measures which operate on substring sequences. The *longest common substring* measure (Gusfield, 1997) compares the length of the longest contiguous sequence of characters between two texts, normalized by the text lengths. However, the editorial process in journalistic text reuse or the attempt to obfuscate copying in plagiarism may shorten the longest common substring considerably, e.g. when words are inserted or deleted, or parts of reused text appear in a different order. The *longest common subsequence* measure (Allison and Dix, 1986) drops the contiguity requirement and allows to detect text reuse in case of word insertions/deletions. *Greedy String Tiling* (Wise, 1996) further allows to deal with reordered parts of reused text as it determines a set of shared contiguous substrings between two given documents, each substring thereby being a match of maximal length. A multitude of other string similarity measures have been proposed which view texts as sequences of characters and compute their degree of

[2] In addition, we release an open-source framework which contains implementations of all discussed measures in order to stimulate the development of novel measures: `http://code.google.com/p/dkpro-similarity-asl`

distance according to a given metric. We used the following measures in our experiments: *Jaro (1989)*, *Jaro-Winkler* (Winkler, 1990), *Monge and Elkan (1997)*, and *Levenshtein (1966)*.

Starting from the observation that not all words in a document are of equal importance, we further employed a similarity measure which weights all words by a *tfidf* scheme (Salton and McGill, 1983) and computes text similarity as the cosine between two document vectors.

Comparing *word n-grams* (Lyon et al., 2001) is a popular means for comparing lexical patterns between two texts. The more similar the patterns, the more likely is it that text reuse has occurred. After compiling two sets of n-grams, we compared them using the Jaccard coefficient, following Lyon et al. (2001), as well as using the containment measure (Broder, 1997). We tested n-gram sizes for $n = 1, 2, \ldots, 15$, and will use the original system name *Ferret* (Lyon et al., 2004) to refer to the variant with $n = 3$ using the Jaccard coefficient, henceforth.

Following the idea of comparing lexical patterns, we also used a measure which has not yet been considered for assessing content similarity: *character n-gram profiles* (Keselj et al., 2003).[3] We follow the implementation by Barrón-Cedeño et al. (2010) and discard all characters (case insensitive) which are not in the alphabet $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$, then generate all n-grams on character level, weight them by a *tfidf* scheme, and finally compare the feature vectors of both the rewritten and the source text using the cosine measure. While in the original implementation only $n = 3$ was used, we generalize the measure to $n = 2, 3, \ldots, 15$.

In cases where the editor replaced content words by synonyms, string measures typically fail due to the vocabulary gap. We thus used similarity measures which are capable of measuring semantic similarity between words. We used the following word similarity measures with WordNet (Fellbaum, 1998): *Jiang and Conrath (1997)*, *Lin (1998)*, and *Resnik (1995)*. In order to scale these pairwise word similarity scores to the document level, we follow the aggregation strategy by Mihalcea et al. (2006): First, a directional similarity score $sim_d(T_i, T_j)$ is computed from a text T_i to a second text T_j (Eq. 1). Therefore, for each word w_i in T_i, its best-matching counterpart in T_j is sought ($maxSim(w_i, T_j)$). The similarity scores of all these matches are summed up and weighted according to their inverse document frequency *idf* (Spärck Jones, 1972), then normalized. The final document-level similarity figure is the average of applying this strategy in both directions, from T_i to T_j and vice-versa (Eq. 2).

$$sim_d(T_i, T_j) = \frac{\sum_{w_i} maxSim(w_i, T_j) \cdot idf(w_i)}{\sum_{w_i} idf(w_i)} \quad (1) \qquad sim(T_i, T_j) = \frac{1}{2}\left(sim_d(T_i, T_j) + sim_d(T_j, T_i)\right) \quad (2)$$

We also tested text expansion mechanisms with the semantic word similarity measures described above: We used the Moses SMT system (Koehn et al., 2007), trained on Europarl (Koehn, 2005), to translate the original English texts via a bridge language (Dutch) back to English. Thereby, the idea was that in the translation process additional lexemes are introduced which alleviate potential lexical gaps. We computed pairwise word similarity with the measures described above and aggregated according to Mihalcea et al. (2006).

Furthermore, we used the statistical technique *Latent Semantic Analysis (LSA)* (Landauer et al.,

[3] Traditionally, character n-gram profiles have rather been shown successful for authorship attribution. However, the similarity scores of word n-grams and those of character n-gram profiles are highly correlated: Assuming 5 characters per word on average for English texts (Shannon, 1951), we set $n = 3$ for word n-grams and $n = 15$ for character n-grams, and computed Pearson's correlation r between the corresponding similarity scores. We obtained $r = .93$ and $r = .86$ on the datasets introduced in Sections 3.1 and 3.2, respectively, and thus conclude that this measure captures *content similarity* rather than *stylistic similarity*.

1998) for comparing texts. The construction of the semantic space was done using the evaluation corpora (see Section 3). We also used the vector space model *Explicit Semantic Analysis* (ESA) (Gabrilovich and Markovitch, 2007). Besides WordNet, we used two additional lexical-semantic resources for the construction of the ESA vector space: Wikipedia[4] and Wiktionary[5].

2.2 Structural Similarity

As discussed above, we presume that content similarity alone is not a reliable indicator of text reuse. Two independently written texts about the same topic are likely to make use of a common vocabulary to a certain extent. We thus propose to also use measures of structural similarity which compute similarity based on structural aspects inherent to the compared texts.

Stopword n-grams (Stamatatos, 2011) are based on the idea that text reuse often preserves syntactic similarities while exchanging content words. Thus, the measure removes all content words while preserving only stopwords. All n-grams of both texts are then compared using the containment measure (Broder, 1997). We tested n-gram sizes for $n = 2, 3, \ldots, 15$.

For the same reason, we also included *part-of-speech n-grams* in our feature set. Disregarding the actual words that appear in two given texts, computing n-grams along part-of-speech tags allows to detect syntactic similarities between these texts. Again, we tested n-gram sizes for $n = 2, 3, \ldots, 15$, and compared the two sets using the containment measure (Broder, 1997).

We also employed two similarity measures between pairs of words (Hatzivassiloglou et al., 1999). The *word pair order* measure assumes that a similar syntactical structure in reused texts may cause two words to occur in the same order in both texts (with any number of words in between). The complementary *word pair distance* measure counts the number of words which lie between those of a given pair. For each measure, we computed feature vectors for both texts along all shared word pairs and compared the vectors using Pearson's correlation.

2.3 Stylistic Similarity

Measures of stylistic similarity adopt ideas from authorship attribution (Mosteller and Wallace, 1964) or use statistical properties of texts to compute text similarity. The *type-token ratio (TTR)* (Templin, 1957), for example, compares the vocabulary richness of two texts. However, it suffers from sensitivity to variations in text length and the assumption of textual homogeneity (McCarthy and Jarvis, 2010): As a text gets longer, the increase of tokens is linear, while the increase of types steadily slows down. In consequence, lexical repetition causes the TTR value to vary, while it does not necessarily entail that a reader perceives changes in the vocabulary usage. Secondly, textual homogeneity is the assumption of the existence of a single lexical diversity level across a whole text, which may be violated by different rhetorical strategies. *Sequential TTR* (McCarthy and Jarvis, 2010) alleviates these shortcomings. It iteratively computes a TTR score for a dynamically growing text segment until a point of saturation – i.e. a fixed TTR score of .72 – is reached, then starts anew from that position in the text for a new segment. The final lexical diversity score is computed as the number of tokens divided by the number of segments.

Inspired by Yule (1939) who discussed sentence length as a characteristic of style, we also used two simple measures, *sentence length* and *token length*, in our system. These measures compute the average number of tokens per sentence and the average number of characters per token.

[4] www.wikipedia.org
[5] www.wiktionary.org

Text Similarity Feature	WP Rewrite		METER		Webis CPC	
	Acc.	$\vec{F}_1$	Acc.	$\vec{F}_1$	Acc.	$\vec{F}_1$
Majority Class Baseline	.400	.143	.715	.417	.517	.341
Ferret Baseline	.642	.517	.684	.535	.794	.789
Content Similarity						
Character 5-gram Profiles	.642	.537	.715	.417	.753	.742
ESA (Wikipedia)	.474	.323	.711	.484	.760	.753
Greedy String Tiling	.558	.457	.755	.645	.805	.800
Longest Common Substring	.621	.524	.719	.467	.743	.736
Resnik	.632	.500	.715	.417	.666	.656
Word 2-grams Containment	.747	.683	.727	.692	.801	.797
Structural Similarity						
Lemma Pair Distance	.611	.489	.715	.417	.775	.767
Lemma Pair Ordering	.642	.494	.715	.417	.785	.780
POS 3-grams Containment	.642	.554	.731	.701	.787	.783
Stopword 3-grams	.632	.515	.715	.417	.778	.776
Stopword 7-grams	.653	.527	.652	.482	.753	.750
Stylistic Similarity						
Function Word Frequencies	.453	.296	.715	.417	.727	.719
Sequential TTR	.400	.220	.715	.417	.667	.638
Sentence Ratio	.389	.268	.755	.625	.657	.653
Token Ratio	.432	.222	.755	.619	.778	.774
Type-Token Ratio	.379	.197	.715	.417	.723	.712

Table 1: Performance of selected similarity measures on the Wikipedia Rewrite Corpus, the METER Corpus, and the Webis Crowd Paraphrase Corpus, grouped by similarity dimension

Additionally, we compared the average sentence and token lengths between the reused text and the original source. We refer to these measures as *sentence ratio* and *token ratio*, respectively.

Finally, we compare texts by their *function word frequencies* (Dinu and Popescu, 2009) which have shown to be good style indicators in authorship attribution studies. Following the original work, this measure uses a set of 70 function words identified by Mosteller and Wallace (1964) and computes feature vectors of their frequencies for each possibly reused document and the source text. The comparison of the vectors is then performed using Pearson's correlation.

3 Experiments & Results

We utilized three datasets for the evaluation of our system which originate in the fields of plagiarism detection, journalistic text reuse detection, and paraphrase recognition: the **Wikipedia Rewrite Corpus** (Clough and Stevenson, 2011), the **METER Corpus** (Gaizauskas et al., 2001), and the **Webis Crowd Paraphrase Corpus** (Burrows et al., 2012), described below.

We carried out the same evaluation procedure for each of the three datasets: First, we computed text similarity scores between all pairs of possibly reused texts and their original sources using all the measures introduced in Section 2. We then used these scores as features for two machine learning classifiers in order to combine them across the three dimensions *content*, *structure*, and *style*. We experimented with two classifiers from the WEKA toolkit (Hall et al., 2009): a Naive Bayes classifier and a C4.5 decision tree classifier (J48 implementation).

In a 10-fold cross-validation setup, we ran three sets of experiments as follows: (i) First, we tested only the text similarity scores of one single measure at a time as single feature for the classifiers, in order to determine the individually best-performing measures per similarity

System	Acc.	$\bar{F}_1$
Majority Class Baseline	.400	.143
Ferret Baseline	.642	.517
Chong et al. (2010)[6]	*.705*	*.641*
Clough and Stevenson (2011)		
- our re-implementation[7]	.726	.658
- as reported in their work	*.800*	*.757*
Our Approach	**.842**	**.811**

exp. \ class.	cut&paste	light rev.	heavy rev.	no plag.
cut&paste	**15**	1	1	2
light rev.	3	**13**	3	0
heavy rev.	2	2	**15**	0
no plag.	0	0	1	**37**

Table 2: Results and confusion matrix (expected class vs. classification result) for the best classification on the Wikipedia Rewrite Corpus for the original 4-way classification

dimension. (ii) We then combined the measures per dimension by using multiple text similarity scores as feature set, in order to determine the performance of multiple measures within a single dimension. (iii) Finally, we combined the measures across dimensions to determine the best overall configuration. We compare our results with two baselines: the majority class baseline and the word trigram similarity measure *Ferret* (Lyon et al., 2004) (see Section 2.1). Additionally, we report the best results from the literature for comparison.

Evaluation was carried out in terms of accuracy and $\bar{F}_1$ score. By accuracy, we refer to the number of correctly predicted texts divided by the total number of texts. As the class distributions in both datasets are skewed, we report the overall $\bar{F}_1$ score as the arithmetic mean across the F_1 scores of all classes in order to account for the class imbalance.

3.1 Wikipedia Rewrite Corpus

Dataset The dataset contains 100 pairs of short texts (193 words on average). For each of 5 questions about topics of computer science (e.g. "What is dynamic programming?"), a reference answer (*source text*, henceforth) has been manually created by copying portions of text from a suitable Wikipedia article. Text reuse now occurs between a source text and an answer given by one of 19 participants. The participants were asked to provide short answers, each of which should comply to one of 4 rewrite levels and hence reuse the source text to a varying extent. According to the degree of rewrite, the dataset is 4-way classified as *cut & paste* (38 texts; simple copy of text portions from the Wikipedia article), *light revision* (19; synonym substitutions and changes of grammatical structure allowed), *heavy revision* (19; rephrasing of Wikipedia excerpts using different words and structure), and *no plagiarism* (19; answer written independently from the Wikipedia article). An example of a *heavy revision* was given in Figure 1.

Results We summarize the results on this dataset in Table 2.[8] In the best configuration, when combining similarity measures across dimensions, our system achieves a performance of

[6]Chong et al. (2010) report $\bar{F}_1 = .698$ in their original work. This figure, however, reflects the *weighted* arithmetic mean over all four classes of the dataset where one class is twice as prominent as each of the others. As dicussed in Section 3, we report all $\bar{F}_1$ scores as the *unweighted* arithmetic mean in order to account for the class imbalance.

[7]While we were able to reproduce the results of the *Ferret* baseline as reported by Chong et al. (2010), our re-implementation of the system by Clough and Stevenson (2011) (Naive Bayes classifier, same feature set) resulted in a much lower overall performance. We observed the largest difference for the *longest common subsequence* measure, even though we used a standard implementation (Allison and Dix, 1986) and normalized as described by Clough and Stevenson (2011).

[8]Figures in italics are taken from the literature, while we (re-)implemented the remaining systems. This applies to all result tables in this paper.

Text Similarity Dimension	Acc.	$\bar{F}_1$
Combinations within dimensions		
Content	.747	.693
Structure	.716	.660
Style	.442	.398
Combinations across dimensions		
Content + Style	.800	.757
Content + Structure	**.842**	**.811**
Structure + Style	.632	.569
Content + Structure + Style	.832	.798

Table 3: Results of the best combinations of text similarity measures within and across dimensions on the Wikipedia Rewrite Corpus

$\bar{F}_1 = .811$. It outperforms the best reference system by Clough and Stevenson (2011) by 5.4% points in terms of $\bar{F}_1$ score compared to their reported numbers, and by 15.3% points compared to our re-implementation of this system[7]. Their system uses a Naive Bayes classifier with only a very small feature set: *word n-gram containment* ($n = 1, 2, \ldots, 5$) and *longest common subsequence*. For comparison, we re-implemented their system and also applied it to the two datasets in the remainder of this paper. We report our findings in Sections 3.2 and 3.3.

In Table 1, we further report the detailed results for a selected set of individual text similarity measures, listed by similarity dimension.[9] Due to space limitations, we only report a selected set of best-performing measures per dimension and compare them with the baselines: While the majority class baseline performs very poor on this dataset ($\bar{F}_1 = .143$), the *Ferret* baseline achieves $\bar{F}_1 = .517$. Some content similarity measures such as *word 2-grams containment* show a reasonable performance ($\bar{F}_1 = .683$), while structural measures cannot exceed $\bar{F}_1 = .554$, and stylistic measures perform only slightly better than the majority class baseline ($\bar{F}_1 = .296$).

In Table 3, we report the best results for the combinations of text similarity measures within and across dimensions. When we combine the measures within their respective dimensions, *content* outperforms structural and stylistic similarity. However, all combinations of measures across dimensions in addition to content similarity improve the results. The best performance is achieved by combining the three similarity measures *longest common subsequence*, *stopword* 10-*grams*, and *character* 5-*gram profiles* from the two dimensions *content* and *structure*. This supports our hypothesis that the similarity computation process indeed profits from dimensions other than *content*. The effects of dimension combination held true regardless of the classifier used, even though the decision tree classifier performed consistently better than Naive Bayes.

Error Analysis We present the confusion matrix for our best configuration in Table 2. In total, 15 texts out of 95 have been classified with the wrong label. While all texts except a single one in the class *no plagiarism* have been classified correctly, 67% of errors (10 texts) are due to misclassifications in the *light* and *heavy revision* classes. We assume that these errors are due to questionable gold standard annotations as the annotation guidelines for these two classes are highly similar (Clough and Stevenson, 2011). For the *light revision* class, the annotators "could alter the text in some basic ways", thereby "altering the grammatical structure (i.e. paraphrasing)." Likewise, for the *heavy revision* class, the annotation manual expected the

[9]Table 1 also lists the detailed results for the METER Corpus and the Webis Crowd Paraphrase Corpus. We will discuss the numbers in the corresponding Sections 3.2 and 3.3.

System	Acc.	$\bar{F}_1$
Majority Class Baseline	.400	.190
Ferret Baseline	.768	.745
Clough and Stevenson (2011)[13]	.821	.788
Our Approach	**.884**	**.859**

exp. \ class.	cut&paste	potential	no plag.
cut&paste	**14**	3	2
potential	5	**33**	0
no plag.	0	1	**37**

Table 4: Results and confusion matrix on the Wikipedia Rewrite Corpus for the folded 3-way classification

System	Acc.	$\bar{F}_1$
Majority Class Baseline	.600	.375
Ferret Baseline	.937	.935
Clough and Stevenson (2011)		
- our re-implementation	.958	.957
- as reported	*.947*	*n/a*
Our Approach	**.968**	**.967**

exp. \ class.	plagiarism	no plag.
plagiarism	**55**	2
no plag.	1	**37**

Table 5: Results and confusion matrix on the Wikipedia Rewrite Corpus for the folded binary classification

annotators to "rephrase the text to generate an answer with the same meaning as the source text, but expressed using different words and structure."

As each text of this dataset was written by only a single person for a given rewrite category, we decided to conduct an annotation study, in which we were mostly interested in the inter-rater agreement of the subjects. We asked 3 participants to rate the degree of text reuse and provided them with the original annotation guidelines. We used a generalization of Scott's (1955) π-measure for calculating a chance-corrected inter-rater agreement for multiple raters, which is known as Fleiss' (1971) κ and Carletta's (1996) K.[10] In summary, the results[11] of our study support our hypothesis that the annotators mostly disagree for the *light* and *heavy revision* classes, with fair[12] agreements of $\kappa = .34$ and $\kappa = .28$, respectively. For the *cut & paste* and *no plagiarism* classes, we observe moderate[12] agreements, $\kappa = .53$ and $\kappa = .56$, respectively.

Based on these insights, we decided to fold the *light* and *heavy revision* classes into a single class *potential plagiarism*. This approach was also briefly discussed by Clough and Stevenson (2011), though not carried out in their work. We report the corresponding results and the confusion matrix in Table 4. As the classification task gets easier by the reduction to three classes, the results for the Ferret baseline improve, from $\bar{F}_1 = .517$ to $\bar{F}_1 = .745$. The re-implementation of the system by Clough and Stevenson (2011) achieves $\bar{F}_1 = .788$. Our system again outperforms all other systems with $\bar{F}_1 = .859$.

In our envisioned semi-supervised application scenario, potentially reused texts are presented to users in an informative manner. Here, fine-grained distinctions are not necessary, and we decided to go even one step further and fold all potential cases of text reuse. This variant of the dataset results in a binary classification of plagiarized/non-plagiarized texts. We present

[10] An exhaustive discussion of inter-rater agreement measures is given by Artstein and Poesio (2008).

[11] `http://www.ukp.tu-darmstadt.de/data/text-similarity/text-reuse-annotations`

[12] Strength of agreement for κ values according to Landis and Koch (1977)

[13] We report the results for our re-implementation of the system by Clough and Stevenson (2011). In their original work, they did not evaluate on this dataset.

System	Acc.	$\bar{F}_1$
Majority Class Baseline	.715	.417
Ferret Baseline	.684	.535
Clough and Stevenson (2011)[13]	.692	.680
Sánchez-Vega et al. (2010)	*.783*	*.705*
Our Approach	**.802**	**.768**

exp. \ class.	reuse	no reuse
reuse	**151**	30
no reuse	20	**52**

Table 6: Results and confusion matrix for the best classification on the METER Corpus

the results and the corresponding confusion matrix in Table 5. In this simplified setting, even the Ferret baseline achieves an excellent performance of $\bar{F}_1 = .935$. Our approach still slightly outperforms ($\bar{F}_1 = .967$) the re-implementation of the system by Clough and Stevenson (2011).

An interesting observation across all three variants of the dataset is that the same three texts always constitute severe error instances where e.g. a *cut & paste* text is falsely labeled as *no plagiarism*, which is more severe than mislabeling a *light revision* as a *heavy revision*. Two of the three cases account for the texts which describe the PageRank algorithm. One of these instances was falsely labeled as *cut & paste* while it is non-plagiarized, and the other one vice-versa. We attribute the misclassifications to the model built up in the classifier's training phase.

In the envisioned semi-supervised setting, the remaining less severe error instances, where e.g. a *light revision* was classified as a *heavy revision*, can be reviewed by a user of the system. We suppose it is even hard for users to draw a strict line between possibly reused and non-reused texts, as this heavily depends on external effects such as user intentions and the task at hand.

3.2 METER Corpus

Dataset The dataset contains news sources from the UK Press Association (PA) and newspaper articles from 9 British newspapers that reused the PA source texts to generate their own texts. The complete dataset contains 1,716 texts from two domains: law & court and show business. All newspaper articles have been annotated whether they are *wholly derived* from the PA sources (i.e. the PA text has been used exclusively as text reuse source), *partially derived* (the PA text has been used in addition to other sources), or *non-derived* (the PA text has not been used at all).

Several newspaper texts, though, have more than a single PA source in the original dataset where it is unclear which (if not all) of the source stories have been used to generate the rewritten story. However, for text reuse detection it is important to have aligned pairs of reused texts and source texts. Therefore, we followed Sánchez-Vega et al. (2010) and selected a subset of texts where only a single source story is present in the dataset. This leaves 253 pairs of short texts (205 words on average). We further followed Sánchez-Vega et al. (2010) and folded the annotations to a binary classification of 181 *reused* (wholly/partially derived) and 72 *non-reused* instances in order to carry out a comparable evaluation study.

Results We summarize the results on this dataset in Table 6. In the best configuration, our system achieves an overall performance of $\bar{F}_1 = .768$. It outperforms the best reference system by Sánchez-Vega et al. (2010) by 6.3% points in terms of $\bar{F}_1$ score. Their system uses a Naive Bayes classifier with two custom features which compare texts based on the length and frequency of common word sequences and the relevance of individual words. As in Section 3.1, we further report the detailed results for a selected set of individual text similarity measures

Text Similarity Dimension	Acc.	$\bar{F}_1$
Combinations within dimensions		
Content	.759	.712
Structure	.731	.701
Style	.755	.672
Combinations across dimensions		
Content + Style	.779	.733
Content + Structure	.739	.713
Structure + Style	.767	.739
Content + Structure + Style	**.802**	**.768**

Table 7: Results of the best combinations of text similarity measures within and across dimensions on the METER Corpus

in Table 1. From these figures, we learn that many text similarity measures cannot exceed the simple majority class baseline ($\bar{F}_1 = .417$) when applied individually.

In Table 7, we show that the performance of text reuse detection always improves over individual measures (cf. Table 1) when we combine the measures within their respective dimensions. An exception is the combination of structural similarity measures, which only performs on the same level as the best individual measure *part-of-speech* 3-*grams containment*. Combinations of content similarity measures show a better performance than combinations of structural or stylistic measures. Our system achieves its best performance on this dataset when text similarity measures are combined across all three dimensions *content*, *structure*, and *style*. The best configuration resulted from using a Naive Bayes classifier with the following measures: *Greedy String Tiling*, *stopword* 12-*grams*, and *Sequential TTR*. As for the previous dataset, the effects of dimension combination held true regardless of the classifier used.

The influence of the stylistic similarity measures is particularly interesting to note. In contrast to the Wikipedia Rewrite Corpus, including these measures in the composition improves the results on this dataset: Our classifier is able to detect similarity even for reused texts by expert journalists. This is due to the fact that a journalistic text which reuses the original press agency source most likely also shows stylistic similarity in terms of e.g. vocabulary richness.

Error Analysis We present the confusion matrix for our best configuration in Table 6. In total, 50 texts out of 253 have been classified incorrectly: 30 instances of text reuse have not been identified by the classifier, and 20 non-reused texts have been mistakenly labeled as such. However, the original annotations have been carried out by only a single annotator (Gaizauskas et al., 2001) which may have resulted in subjective judgments. Thus, as for the previous dataset in Section 3.1, we conducted an annotation study with three annotators to gain further insights into the data. The results[11] show that for 61% of all texts the annotators fully agree. The chance-corrected Fleiss' (1971) agreement $\kappa = .47$ is moderate[12].

For the 30 instances of text reuse which have not been identified by the classifier, it is particularly interesting to note that many errors are due to the fact that a lower overall text similarity between the possibly reused text and the original source does not necessarily entail the label *no reuse*. The newspaper article about the English singer-songwriter Liam Gallagher, for example, is originally labeled as text reuse. However, our classifier falsely assigned the label *no reuse*. It turns out, though, that the reused text is about four times as long as the original press agency source, with lots of new facts being introduced there. Consequently, only a low similarity score

System	Acc.	$\bar{F}_1$
Majority Class Baseline	.517	.341
Ferret Baseline	.794	.789
Clough and Stevenson (2011)[13]	.798	.795
Burrows et al. (2012)	*.839*	*.837*
Our Approach	**.853**	**.852**

exp. \ class.	paraphrase	no para.
paraphrase	**3,654**	413
no para.	759	**3,033**

Table 8: Results and confusion matrix for the best classification on the Webis Crowd Paraphrase Corpus

can be computed between the additional material in the newspaper article and the original source, and the overall similarity score decreases.

We conclude that applications will benefit from an improved classifier which better deals with theses instances. For example, similarity features could be computed per section, not per document, which would allow to also identify potential instances of text reuse for only partially matching texts. The currently achieved performance (see Table 6) of text reuse detection, though, is sufficient for our envisioned semi-supervised application scenario where content authors are provided only with suggestions of potential instances of text reuse and then are free to decide how to proceed, e.g. to merge both texts. The final decision probably also depends on external factors such as user intentions and the task at hand.

3.3 Webis Crowd Paraphrase Corpus

Dataset The dataset was originally introduced as part of the PAN 2010 international plagiarism detection competition (Potthast et al., 2010). It contains 7,859 pairs of original texts along with their paraphrases (28 to 954 words in length) with 4,067 (52%) positive and 3,792 (48%) negative samples. The original texts are book excerpts from Project Gutenberg[14], and the corresponding paraphrases were acquired in a crowdsourcing process using Amazon Mechanical Turk (Callison-Burch and Dredze, 2010). In the manual filtering process[15] of all acquired paraphrases, Burrows et al. (2012) hereby follow the paraphrase definition by Boonthum (2004), where a *good* paraphrase exhibits patterns such as synonym use, changes between active and passive voice, or changing word forms and parts of speech, and a *bad* paraphrase is rather e.g. a (near-)duplicate or an automated one-for-one word substitution. This definition implies that a more sophisticated interpretation of text similarity scores needs to be learned, where e.g. (near-)duplicates with very high similarity scores are in fact negative samples.

Results We summarize the results on this dataset in Table 8. Even though the Ferret baseline is a strong competitor ($\bar{F}_1 = .789$), our approach achieves the best results on this dataset with $\bar{F}_1 = .852$. The results reported by Burrows et al. (2012) are slightly worse ($\bar{F}_1 = .837$). Their best score was achieved by using a k-nearest neighbor classifier with a feature set of 10 similarity measures. They exclusively used similarity measures that operate on the texts' string sequences and thus capture the content dimension of text similarity only, e.g. *Levenshtein (1966)* distance and a *word n-gram* similarity measure. As in the previous sections, we report the detailed results for a selected set of individual text similarity measures in Table 1. These figures show that

[14] `www.gutenberg.org`

[15] Burrows et al. (2012) do not report any inter-annotator agreements for the filtering process, as the task was split across two annotators and each text pair was labeled by only a single annotator.

Text Similarity Dimension	Acc.	$\bar{F}_1$
Combinations within dimensions		
Content	.840	.839
Structure	.816	.814
Style	.819	.817
Combinations across dimensions		
Content + Style	.844	.843
Content + Structure	.838	.838
Structure + Style	.831	.830
Content + Structure + Style	**.853**	**.852**

Table 9: Results of the best combinations of text similarity measures within and across dimensions on the Webis Crowd Paraphrase Corpus

regardless of the similarity dimension many measures achieve a very reasonable performance when applied individually, with the measures *Greedy String Tiling* and *word 2-grams containment* performing best.

As for the previous datasets, our hypothesis holds true that the combination of similarity dimensions improves the results: When we combine the similarity features within each of the respective dimensions, the performance numbers increase (see Table 9 as compared to Table 1). The combination of content similarity measures is stronger than the combination of structural and stylistic similarity measures, and performs on the same level as the original results reported by Burrows et al. (2012). This is to be expected, as their system uses a feature set which also addresses the content dimension exclusively.

When we combine measures across dimensions, the results improve even further. An exception is the combination of content and structural measures, which performs slightly worse than content measures alone due to the lower performance of structural measures on this dataset. The best configuration of our system resulted from combining all three dimensions *content*, *structure*, and *style* in a single classification model using the decision tree classifier, resulting in $\bar{F}_1 = .852$. The final feature set contains 16 text similarity features which are listed in Table 10.

Error Analysis We present the confusion matrix for our best classification in Table 8. In total, 1,172 (15%) out of 7,859 text pairs have been classified incorrectly. Out of these, our classifier mistakenly labeled 759 instances of negative samples as true paraphrases, while 413 cases of true paraphrases were not recognized. However, in our opinion the 759 false positives are less severe errors in our envisioned semi-supervised application setting, as user intentions and the current task at hand may highly influence a user's decision to consider texts as reused or not.

In general, we attribute the errors to the particular properties of this dataset, which differ from those of the Wikipedia Rewrite Corpus and the METER Corpus (see Sections 3.1 and 3.2). For those two datasets, the more similar two texts are, the higher their degree of text reuse. For the Webis Crowd Paraphrase Corpus, however, a different interpretation needs to be learned by the classifier: Here, (near-)duplicates and texts with automated word-by-word substitutions, which will receive high similarity scores by any of our content similarity measures, are in fact annotated as *bad* paraphrases, i.e. negative samples. Unrelated texts, empty samples, or texts alike also belong to the class of negative samples. In consequence, positive samples are only those in the medium similarity range. We assume that the more elaborate definition of positive and negative cases makes it more difficult to learn a proper model for the given data.

Content	ESA (WordNet, with + w/o stopwords), Greedy String Tiling, Jaro, Longest Common Substring, Longest Common Subseq. (2 norm.), n-gram Jaccard ($n = \{6, 14, 15\}$), Resnik (SMT wrapper)
Structure	Lemma Pair Ordering, POS 2-grams Jaccard, Stopword 6-grams
Style	Function Word Frequencies, Sequential TTR, Token Ratio

Table 10: Feature set used to achieve the best results on the Webis Crowd Paraphrase Corpus

4 Conclusions and Future Work

The motivation for this work stemmed from the hypothesis that *content* features alone are not a reliable indicator for text reuse detection. As illustrated in Figure 1, a reused text may also contain modifications such as split sentences, changed order of reused parts, or stylistic variance. We thus devised an architecture which composes diverse text similarity measures in a supervised classification model. In this model, we overcome the traditional limitation of text similarity measures to *content* features and compute similarity along three characteristic dimensions inherent to texts: *content*, *structure*, and *style*.

We evaluated our classification model on three standard datasets where text reuse is prevalent and which originate in the fields of plagiarism detection, journalistic text reuse detection, and paraphrase recognition: the *Wikipedia Rewrite Corpus* (Clough and Stevenson, 2011), the *METER Corpus* (Gaizauskas et al., 2001), and the *Webis Crowd Paraphrase Corpus* (Burrows et al., 2012). Based on the evaluation results, we discussed the influence of each of the similarity dimensions, and demonstrated empirically that text reuse can be best detected if measures are combined across dimensions, so that a wide variety of text features are taken into consideration. The composition consistently outperforms previous approaches across all datasets.

As we showed, similarity computation works best if the similarity dimensions are chosen well with respect to the type of text reuse at hand. For the Wikipedia Rewrite Corpus, for example, the stylistic similarity features perform only poorly, which is why the composition of all three dimensions performs slightly worse than than the combination of only content and structural features. For the other two datasets, however, stylistic similarity is a strong dimension within the composition, and consequently the best performance is reached when combining all three dimensions. Based on these insights, we conclude that for novel datasets it is essential to address the dimensions explicitly in the annotation process, so that text reuse detection approaches can be evaluated precisely against particular characteristics of different kinds of data.

For future work, we expect that considering a dimensional representation of text similarity features will also benefit any other task where text similarity computation is fundamental and which is yet limited to *content* features, e.g. paraphrase recognition or automatic essay grading. For the latter, we see great potential for improvements by including, for example, measures for grammar analysis, lexical complexity, or measures assessing text organization with respect to the discourse elements. However, each task exhibits particular characteristics which influence the choice of a suitable set of similarity dimensions. As discussed above, a particular dimension may or may not contribute to an overall improvement based on the nature of the data.

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Klaus Tschira Foundation under project No. 00.133.2008. We thank Chris Biemann for his inspirations, as well as Carolin Deeg, Andriy Nadolskyy, and Artem Vovk for their participation in the annotation studies.

References

Abdel-Hamid, O., Behzadi, B., Christoph, S., and Henzinger, M. (2009). Detecting the Origin of Text Segments Efficiently. In *Proceedings of the 18th International Conference on World Wide Web*, pages 61–70, Madrid, Spain.

Allison, L. and Dix, T. I. (1986). A bit-string longest-common-subsequence algorithm. *Information Processing Letters*, 23:305–310.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Bär, D., Zesch, T., and Gurevych, I. (2011). A Reflective View on Text Similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 515–520, Hissar, Bulgaria.

Barrón-Cedeño, A., Rosso, P., Agirre, E., and Labaka, G. (2010). Plagiarism Detection across Distant Language Pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 37–45, Beijing, China.

Boonthum, C. (2004). iSTART: Paraphrase Recognition. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 31–36, Barcelona, Spain.

Broder, A. Z. (1997). On the resemblance and containment of documents. *Proceedings of Compression and Complexity of Sequences*, pages 21–29.

Broder, A. Z., Glassman, S. C., Manasse, M. S., and Zweig, G. (1997). Syntactic clustering of the Web. In *Proceedings of the 6th International World Wide Web Conference*, pages 1157–1166, Santa Clara, CA, USA.

Burrows, S., Potthast, M., and Stein, B. (2012). Paraphrase Acquisition via Crowdsourcing and Machine Learning. *Transactions on Intelligent Systems and Technology*, V(January):1–22.

Callison-Burch, C. and Dredze, M. (2010). Creating Speech and Language Data With Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, CA, USA.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Charikar, M. S. (2002). Similarity Estimation Techniques from Rounding Algorithms. In *Proceedings of the 34th Annual Symposium on Theory of Computing*, pages 380–388, Montreal, Canada.

Chong, M., Specia, L., and Mitkov, R. (2010). Using Natural Language Processing for Automatic Detection of Plagiarism. In *Proceedings of the 4th International Plagiarism Conference*, Newcastle upon Tyne, UK.

Clough, P., Gaizauskas, R., Piao, S. S., and Wilks, Y. (2002). METER: MEasuring TExt Reuse. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 152–159, Philadelphia, PA, USA.

Clough, P. and Stevenson, M. (2011). Developing a Corpus of Plagiarised Short Answers. *Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis*, 45(1):5–24.

Dinu, L. P. and Popescu, M. (2009). Ordinal measures in authorship identification. In *Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 62–66, San Sebastian, Spain.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., and Piao, S. (2001). The METER Corpus: A corpus for analysing journalistic text reuse. In *Proceedings of the Corpus Linguistics 2001 Conference*, pages 214–223.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Goodman, N. (1972). Seven strictures on similarity. In Goodman, N., editor, *Problems and projects*, pages 437–446. Bobbs-Merrill.

Gusfield, D. (1997). *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Hatzivassiloglou, V., Klavans, J. L., and Eskin, E. (1999). Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, College Park, MD, USA.

Henzinger, M. (2006). Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 284–291, Seattle, WA, USA.

Hoad, T. C. and Zobel, J. (2003). Methods for identifying versioned and plagiarized documents. *Journal of the American Society of Information Science and Technology*, 54(3):203–215.

Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Association*, 84(406):414–420.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*.

Keselj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pages 255–264.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket Island, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2):259–284.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lee, J. (2007). A Computational Model of Text Reuse in Ancient Literary Texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479.

Leuf, B. and Cunningham, W. (2001). *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley Professional.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*, pages 296–304.

Lyon, C., Barrett, R., and Malcolm, J. (2004). A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector. In *In Plagiarism: Prevention, Practice and Policies Conference*.

Lyon, C., Malcolm, J., and Dickerson, B. (2001). Detecting short passages of similar text in large document collections. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 118–125.

Manku, G. S., Jain, A., and Sarma, A. D. (2007). Detecting Near-Duplicates for Web Crawling. In *Proceedings of the 16th International World Wide Web Conference*, pages 141–149, Banff, AB, Canada.

McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 775–780, Boston, MA, USA.

Monge, A. and Elkan, C. (1997). An efficient domain-independent algorithm for detecting approximately duplicate database records. In *Proceedings of the SIGMOD Workshop on Data Mining and Knowledge Discovery*, pages 23–29, Tucson, AZ, USA.

Mosteller, F. and Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. (2010). Overview of the 2nd International Competition on Plagiarism Detection. In *Notebook Papers of CLEF 10 Labs and Workshops*, Padua, Italy.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Sánchez-Vega, F., Villaseñor-Pineda, L., Montes-y-Gómez, M., and Rosso, P. (2010). Towards Document Plagiarism Detection Based on the Relevance and Fragmentation of the Reused Text. In *Proceedings of the 9th Mexican International Conference on Artificial Intelligence*, pages 24–31, Pachuca, Mexico.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal*, 30:50–64.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Stamatatos, E. (2011). Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, 62(12):2512–2527.

Templin, M. C. (1957). *Certain language skills in children*. University of Minnesota Press.

Tversky, A. (1977). Features of Similarity. In *Psychological Review*, volume 84, pages 327–352.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research Methods*, pages 354–359.

Wise, M. J. (1996). YAP3: Improved detection of similarities in computer program and other texts. In *Proceedings of the 27th SIGCSE technical symposium on Computer science education*, pages 130–134, Philadelphia, PA, USA.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4):363–390.