Detecting Text Reuse with Modified and Weighted N-grams

Rao Muhammad Adeel Nawab*, **Mark Stevenson*** **and Paul Clough**†
*Department of Computer Science and †iSchool
University of Sheffield, UK.
{r.nawab@dcs, m.stevenson@dcs, p.d.clough@} .shef.ac.uk

Abstract

Text reuse is common in many scenarios and documents are often based, at least in part, on existing documents. This paper reports an approach to detecting text reuse which identifies not only documents which have been reused verbatim but is also designed to identify cases of reuse when the original has been rewritten. The approach identifies reuse by comparing word n-grams in documents and modifies these (by substituting words with synonyms and deleting words) to identify when text has been altered. The approach is applied to a corpus of newspaper stories and found to outperform a previously reported method.

1 Introduction

Text reuse is the process of creating new document(s) using text from existing document(s). Text reuse is standard practice in some situations, such as journalism. Applications of automatic detection of text reuse include the removal of (near-)duplicates from search results (Hoad and Zobel, 2003; Seo and Croft, 2008), identification of text reuse in journalism (Clough et al., 2002) and identification of plagiarism (Potthast et al., 2011).

Text reuse is more difficult to detect when the original text has been altered. We propose an approach to the identification of text reuse which is intended to identify reuse in such cases. The approach is based on comparison of word n-grams, a popular approach to detecting text reuse. However, we also account for synonym replacement and word deletion, two common text editing operations (Bell, 1991). The relative importance of n-grams is accounted for using probabilities obtained from a language model. We show that making use of modified n-grams and their probabilities improves identification of text reuse in an existing journalism corpus and outperforms a previously reported approach.

2 Related Work

Approaches for identifying text reuse based on word-level comparison (such as the SCAM copy detection system (Shivakumar and Molina, 1995)) tend to identify topical similarity between a pair of documents, whereas methods based on sentence-level comparison (e.g. the COPS copy detection system (Brin et al., 1995)) are unable to identify when text has been reused if only a single word has been changed in a sentence.

Comparison of word and character n-grams has proven to be an effective method for detecting text reuse (Clough et al., 2002; Cedeño et al., 2009; Chiu et al., 2010). For example, Cedeño et al. (2009) showed that comparison of word bigrams and trigrams are an effective method for detecting reuse in journalistic text. Clough et al. (2002) also applied n-gram overlap to identify reuse of journalistic text, combining it with other approaches such as sentence alignment and string matching algorithms. Chiu et al. (2010) compared n-grams to identify duplicate and reused documents on the web. Analysis of word n-grams has also proved to be an effective method for detecting plagiarism, another form of text reuse (Lane et al., 2006).

However, a limitation of n-gram overlap approach is that it fails to identify reuse when the original

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 54–58,
Montréal, Canada, June 7-8, 2012.

text has been altered. To overcome this problem we propose using modified n-grams, which have been altered by deleting or substituting words in the n-gram. The modified n-grams are intended to improve matching with the original document.

3 Determining Text Reuse with N-gram Overlap

3.1 N-grams Overlap (NG)

Following Clough et al. (2002), the asymmetric containment measure (eqn 1) was used to quantify the degree of text within a document (A) that is likely to have been reused in another document (B).

$$score_n(A,B) = \frac{\sum_{ngram \in B} count(ngram, A)}{\sum_{ngram \in B} count(ngram, B)} \quad (1)$$

where $count(ngram, A)$ is the number of times $ngram$ appears in document A. A score of 1 means that document B is contained in document A and a score of 0 that none of the n-grams in B occur in A.

3.2 Modified N-grams

N-gram overlap has been shown to be useful for measuring text reuse as derived texts typically share longer n-grams (≥ 3 words). However, the approach breaks down when an original document has been altered. To counter this problem we applied various techniques for modifying n-grams that allow for word deletions (Deletions) and word substitutions (WordNet and Paraphrases), two common text editing operations.

Deletions (Del) Assume that $w_1, w_2, ...w_n$ is an n-gram. Then a set of modified n-grams can be created by removing one of the w_2 ... w_{n-1}. The first and last words in the n-gram are not removed since they will also be generated as shorter n-grams. An n-gram will generate $n-2$ deleted n-grams and no deleted n-grams will be generated for unigrams and bigrams.

Substitutions Further n-grams can be created by substituting one of the words in an n-gram with one of its synonyms from **WordNet (WN)**. For words with multiple senses we use synonyms from *all senses*. Modified n-grams are created by substituting one of the words in the n-gram with one of its synonyms from WordNet.

Similarly to the WordNet approach, n-grams can be created by substituting one of the words with an equivalent term from a paraphrase lexicon, which we refer to as **Paraphrases (Para)**. A paraphrase lexicon was generated automatically (Burch, 2008) and ten lexical equivalents (the default setting) produced for each word. Modified n-grams were created by substituting one of the words in the n-gram with one of the lexical equivalents.

3.3 Comparing Modified N-grams

The modified n-grams are applied in the text reuse score by generating modified n-grams for the document that is suspected to contain reused text. These n-grams are then compared with the original document to determine the overlap. However, the techniques in Section 3.2 generate a large number of modified n-grams which means that the number of n-grams that overlap with document A can be greater than the total number of n-grams in B, leading to similarity scores greater than 1. To avoid this the n-gram overlap counts are constrained in a similar way that they are clipped in BLEU and ROUGE (Papineni et al., 2002; Lin, 2004).

For each n-gram in B, a set of modified n-grams, $mod(ngram)$, is created.[1] The count for an individual n-gram in B, $exp\_count(ngram, B)$, can be computed as the number of times any n-gram in $mod(ngram)$ occurs in A, see equation 2.

$$\sum_{ngram' \in mod(ngram)} count(ngram', A) \quad (2)$$

However, the contribution of this count to the text reuse score has to be bounded to ensure that the combined count of the modified n-grams appearing in A does not exceed the number of times the original n-gram occurs in B. Consequently the text reuse score, $score_n(A,B)$, is computed using equation 3.

$$\frac{\sum_{\substack{ngram \\ \in B}} min(exp\_count(ngram, A), count(ngram, B))}{\sum_{ngram \in B} count(ngram, B)} \quad (3)$$

3.4 Weighting N-grams

Probabilities of each n-gram, obtained using a language model, are used to increase the importance of

[1] This is the set of n-grams that could have been created by modifing an n-gram in B and includes the original n-gram itself.

rare n-grams and decrease the contribution of common ones. N-gram probabilities are computed using the SRILM language modelling toolkit (Stolcke, 2002). The score for each n-gram is computed as its Information Content (Cover and Thomas, 1991), ie. $-log(P)$. When the **language model (LM)** is applied the scores associated with each n-gram are used instead of counts in equations 2 and 3.

4 Experiments

4.1 METER Corpus

The METER corpus (Gaizauskas et al., 2001) contains 771 *Press Association (PA)* articles, some of which were used as source(s) for 945 news stories published by nine British newspapers.

These 945 documents are classified as *Wholly Derived (WD)*, *Partially Derived (PD)* and *Non Derived (ND)*. WD means that the newspaper article is likely derived entirely from the PA source text; PD reflects the situation where some of the newspaper article is derived from the PA source text; news stories likely to be written independently of the PA source fall into the category of ND. In our experiments, the 768 stories from court and law reporting were used (WD=285, PD=300, ND=183) to allow comparison with Clough et al. (2002). To provide a collection to investigate binary classification we aggregated the WD and PD cases to form a Derived set. Each document was pre-processed by converting to lower case and removing all punctuation marks.

4.2 Determining Reuse

The text reuse task aims to distinguish between levels of text reuse, i.e. WD, PD and ND. Two versions of a classification task were used: binary classification distinguishes between Derived (i.e. WD $\cup$ PD) and ND documents, and ternary classification distinguishes all three levels of reuse.

A Naive Bayes classifier (Weka version 3.6.1) and 10-fold cross validation were used for the experiments. Containment similarity scores between all PA source texts and news articles on the same story were computed for word uni-grams, bi-grams, tri-grams, four-grams and five-grams. These five similarity scores were used as features. Performance was measured using precision, recall and F_1 measures with the macro-average reported across all classes.

The language model (Section 3.4) was trained using 806,791 news articles from the Reuters Corpus (Rose et al., 2002). A high proportion of the news stories selected were related to the topics of entertainment and legal reports to reflect the subjects of the new articles in the METER corpus.

5 Results and Analysis

Tables 1 and 2 show the results of the binary and ternary classification experiments respectively. "NG" refers to the comparison of n-grams in each document (Section 3.1), while "Del", "WN" and "Para" refer to the modified n-grams created using deletions, WordNet and paraphrases respectively (Section 3.2). The prefix "LM" (e.g. "LM-NG") indicates that the n-grams are weighted using the language model probability scores (Section 3.4).

For the binary classification task (Table 1) it can be observed that including modified n-grams improves performance. This improvement is observed when each of the three types of modified n-grams is applied individually, with a greater increase being observed for the n-grams created using the WordNet and paraphrase approaches. Further improvement is observed when different types of modified n-grams are combined with the best performance obtained when all three types are used. All improvements over the baseline approach (NG) are statistically significant (Wilcoxon signed-rank test, $p < 0.05$). These results demonstrate that the various types of modified n-grams all contribute to identifying when text is being reused since they capture different types of rewrite operations.

In addition, performance consistently improves when n-grams are weighted using language model scores. The improvement is significant for all types of n-grams. This demonstrates that the information provided by the language model is useful in determining the relative importance of n-grams.

Several of the results are higher than those reported by Clough et al. (2002) (F_1=0.763), despite the fact their approach supplements n-gram overlap with additional techniques such as sentence alignment and string search algorithms.

Results of the ternary classification task are shown in Table 2. Results show a similar pattern to those observed for the binary classification task

Approach	P	R	F_1
NG	0.836	0.706	0.732
LM-NG	0.846	0.722	0.746
Del	0.851	0.745	0.767
LM-Del	0.858	0.765	0.785
WN	0.876	0.801	0.817
LM-WN	0.879	0.810	0.825
Para	0.884	0.821	0.834
LM-Para	0.888	0.831	0.843
Del+WN	0.889	0.835	0.847
LM-Del+WN	0.884	0.848	0.855
Del+Para	0.892	0.841	0.853
LM-Del+Para	0.896	0.849	0.860
WN+Para	0.894	0.848	0.858
LM-WN+Para	0.896	0.865	0.871
Del+WN+Para	0.897	0.856	0.865
LM-Del+WN+Para	**0.903**	**0.876**	**0.882**
(Clough et al., 2002)	—	—	0.763

Table 1: Results for binary classification

Approach	P	R	F_1
NG	0.596	0.557	0.551
LM-NG	0.615	0.579	0.574
Del	0.612	0.584	0.579
LM-Del	0.633	0.611	0.606
WN	0.644	0.636	0.631
LM-WN	0.649	0.640	0.635
Para	0.662	0.653	0.647
LM-Para	0.669	0.659	0.654
Del+WN	0.655	0.649	0.643
LM-Del+WN	0.668	0.656	0.650
Del+Para	0.665	0.658	0.652
LM-Del+Para	0.661	0.662	0.655
WN+Para	0.668	0.661	0.655
LM-WN+Para	0.680	0.675	0.668
Del+WN+Para	0.669	0.666	0.660
LM-Del+WN+Para	**0.688**	**0.689**	**0.683**
(Clough et al., 2002)	—	—	0.664

Table 2: Results for ternary classification

and the best result is also obtained when all three types of modified n-grams are included and n-grams are weighted with probability scores. Once again weighting n-grams with language model scores improves results for all types of n-gram and this improvement is significant. Results for several types of n-gram are also better than those reported by Clough et al. (2002) (F_1=0.664).

Results for all approaches are lower for the ternary classification. This is because the binary classification task involves distinguishing between two classes of documents which are relatively distinct (derived and non-derived) while the ternary task divides the derived class into two (WD and PD) which are more difficult to separate (see Table 3 showing confusion matrix for the approach which gave best results for ternary classification).

Classified as	WD	PD	ND
WD	139	94	14
PD	57	206	54
ND	1	13	191

Table 3: Confusion matrix when "LM-Del+WN+Para" approach used for ternary classification

6 Conclusion

This paper describes an approach to the analysis of text reuse which is based on comparison of n-grams. This approach is augmented by modifying the n-grams in various ways and weighting them with probabilities derived from a language model. Evaluation is carried out on a standard data set containing examples of reused journalistic texts. Making use of modified n-grams with appropriate weights is found to improve performance when detecting text reuse and the approach described here outperforms an existing approach. In future we plan to experiment with other methods for modifying n-grams and also to apply this approach to other types of text reuse.

Acknowledgments

This work was funded by the COMSATS Institute of Information Technology, Islamabad, Pakistan under the Faculty Development Program (FDP) and a Google Research Award.

References

Alberto B. Cedeño, Paolo Rosso, and Jose M. Bened 2009. *Reducing the Plagiarism Detection Search Space on the basis of the Kullback-Leibler Distance* Proceedings of CICLing-09, 523–534.

Allan Bell 1991. *The Language of News Media*. Blackwell.

Andreas Stolcke. 2002. *SRILM - An Extensible Language Modeling Toolkit*. In Proceedings of the International Conference on Spoken Language Processing, 901–904.

Chin-Yew Lin. 2004. *Rouge: A Package for Automatic Evaluation of Summaries*. In Proceedings of the ACL-04 Workshop, 74–81.

Chris Callison-Burch. 2008. *Syntactic Constraints on Paraphrases Extracted from Parallel Corpora*. In Proceedings of EMNLP'08, 196–205.

Jangwon Seo and W. Bruce Croft. 2008. *Local Text Reuse Detection*. In Proceedings of SIGIR'08, 571–578. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 571–578.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei J. Zhu. 2002. *Bleu: A Method for Automatic Evaluation of Machine Translation*. In Proceedings of ACL'02, 311–318.

Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein and Paolo Rosso. 2011. *Overview of the 3rd International Competition on Plagiarism Detection*. Notebook Papers of CLEF 11 Labs and Workshops.

Narayanan Shivakumar and Hector G. Molina. 1995. *SCAM: A Copy Detection Mechanism for Digital Documents*. Proceedings of the 2nd Annual Conference on the Theory and Practice of Digital Libraries, Texas, USA.

Paul Clough, Robert Gaizauskas, Scott S.L. Piao, and Yorick Wilks. 2002. *Measuring Text Reuse*. In Proceedings of ACL'02, Philadelphia, USA, 152–159.

Peter C. R. Lane, Caroline M. Lyon, and James A. Malcolm. 2006. *Demonstration of the Ferret plagiarism detector*. Proceedings of the 2nd International Plagiarism Conference, Newcastle, UK.

Robert Gaizauskas, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough, and Scott S.L. Piao. 2001. *The METER Corpus: A Corpus for Analysing Journalistic Text Reuse*. In Proceedings of the Corpus Linguistics Conference, 214–223.

Sergey Brin, James Davis and Hector G. Molina. 1995. *Copy Detection Mechanisms for Digital Documents*. Proceedings ACM SIGMOD'95, 398–409.

Stanford Chiu, Ibrahim Uysal, Bruce W. Croft. 2010. *Evaluating text reuse discovery on the web*. In Proceedings of the third symposium on Information interaction in context, 299–304.

Thomas M. Cover, Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley, New York, USA.

Timothy C. Hoad and Justin Zobel. 2003. *Methods for Identifying Versioned and Plagiarized Documents*. Journal of the American Society for Information Science and Technology, 54(3):203–215.

Tony Rose, Mark Stevenson, Miles Whitehead. 2002. *The Reuters Corpus Volume 1 - from Yesterday's news to tomorr ow's language resources*. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02), 827–832.