Text Reuse Detection Using a Composition of Text Similarity Measures

Bär, Zesch, Gurevych 2012

HS Computational study of linguistic differences

Sabrina Galasso
sabrina.galasso@student.uni-tuebingen.de

1. Introduction

What is meant by "text reuse"?
How and why should text reuse be detected?

2. Text Similarity Measures

How can text similarity be measured?
What types of measures do exist?

3. Experiments & Results

How do the measures perform on different datasets?
How do individual measure perform?
How can they be combined?

4. Summary

What can we conclude from the experiments?
What can be done as future work?

INTRO

MEASURES

EXPERIMENTS

SUMMARY

What is text reuse?

- Examples for text reuse:
  - Mirroring texts on different websites
  - Reusing texts in public blogs
- Problems with text reuse:
  - Using systems in a collaborative manner
  - e.g., Wikipedia
  - Users should avoid content duplication
- Idea: Supporting authors of collaborative text collections by means of automatic text reuse detection

Text reuse detection

- Applications:
  - Detection of journalistic text reuse
  - Identification of rewrite sources for ancient texts
  - Analysis of text reuse in blogs or web pages
  - Plagiarism detection
  - Near-duplicate detection of websites (web search and crawling)
- Few NLP used so far

Text reuse detection

- Common approach:
  - Computation of similarity based on surface-level or semantic features

    → only consider the text's content
- Idea: investigation of three similarity dimensions:
  - content
  - structure
  - style

Text reuse detection

- Verbatim reuse vs. use of similar words or phrases

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.*

→ detectable by content-centric measures

→ But: What about structural and stylistic similarity?

- Source text was split into two sentences
- Similar vocabulary richness

Text Similarity Measures
Content Similarity

- Detecting verbatim copying: using string measures on substring sequences:
  - ***Longest Common Substring***
    length of longest **contiguous** sequence of characters, normalized by the text length
  - ***Longest Common Subsequence:***
    allows for insertions/deletions
  - ***Greedy String Tiling:***
    determines a set of shared contiguous substrings → allows to deal with reordered parts
  - Other string similarity measures, e.g. ***Levenshtein***

Text Similarity Measures
Content Similarity

- ***tfidf:***
  Measuring similarity based on the importance of individual words
- ***word n-grams***
- ***character n-grams***
- ***Semantic similarity measures***
  using WordNet
- ***Latent Semantic Analysis (LSA)***
- ***Explicit Semantic Analysis (ESA)***
  using WordNet, Wikipedia and Wiktionary

Text Similarity Measures
Structural Similarity

- Assumption: "*Two **independently written** texts about the same topic are likely to make use of a common vocabulary to a certain extent.*"
  →content similarity is not sufficient
  → inclusion of structural aspects
- often only content words are exchanged:
  → comparison of ***stopword n-grams***
  → comparison of ***part-of-speech n-grams***
- two words are likely to occur again in the same order (with any number of words in between)
  - ***word pair order***
  - ***word pair distance***

Text Similarity Measures
Stylistic Similarity

Stylistic similarity:
- Ideas partly adopted from authorship attribution
- Investigation of statistical properties of a text

- ***Type-token ratio (TTR)***
  → no sensitivity to text length
  → assumes textual homogeneity
- ***Sequential TTR***
  computation of the mean length of a string sequence, which maintains a TTR above a default threshold

Text Similarity Measures
Stylistic Similarity

- ***sentence length ratio***
- ***token length ratio***
- ***function word frequencies***
  - makes use of a set of 70 function words identified by Mosteller and Wallace (1964)

Experiments & Results

Experimental Setup

- Three datasets:
  - Wikipedia Rewrite Corpus (Clough and Stevenson, 2011) → plagiarism detection
  - METER Corpus (Gaizauskas et al., 2001) → journalistic text reuse
  - Webis Crowd Paraphrase Corpus (Burrows et al., 2012) → paraphrase recognition

Experiments & Results
Experimental Setup

- Computation of text similarity scores
- Machine learning classifiers: Naive Bayes and decision tree classifier
- Three sets of experiments using 10-fold cross-validation:
  - Performance of individual features
  - Performance of feature combinations within dimensions
  - Performance of feature combinations across dimensions
- Comparison baselines:
  - Majority class baseline
  - Word trigram similarity measure (Ferret)
- Evaluation in terms of accuracy and $\bar{F}_1$ score (arithmetic mean across the F_1 scores of all classes)

Wikipedia Rewrite Corpus

Dataset

- 100 pairs of short texts (193 words)
- Topics of computer science
- Source texts: manually created out of Wikipedia texts
- Reused texts: generated by participants according to 4 rewrite levels:
  - Cut & paste
  - Light revision
  - Heavy revision
  - No plagiarism

INTRO
MEASURES
EXPERIMENTS
SUMMARY

Wikipedia Rewrite Corpus
Comparison to other approaches

- Results for the best classification (combining measures across dimensions):

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**

Features used in Clough and Stevenson (2011):
- word n-gram containment (n= 1,2,...,5)
- longest common subsequence

INTRO MEASURES EXPERIMENTS SUMMARY

Wikipedia Rewrite Corpus

Consideration of individual measures

Text Similarity Feature	WP Rewrite Acc.	WP Rewrite $\bar{F}_1$
Majority Class Baseline	.400	.143
Ferret Baseline	.642	.517
Content Similarity		
Character 5-gram Profiles	.642	.537
ESA (Wikipedia)	.474	.323
Greedy String Tiling	.558	.457
Longest Common Substring	.621	.524
Resnik	.632	.500
Word 2-grams Containment	.747	.683
Structural Similarity		
Lemma Pair Distance	.611	.489
Lemma Pair Ordering	.642	.494
POS 3-grams Containment	.642	.554
Stopword 3-grams	.632	.515
Stopword 7-grams	.653	.527
Stylistic Similarity		
Function Word Frequencies	.453	.296
Sequential TTR	.400	.220
Sentence Ratio	.389	.268
Token Ratio	.432	.222
Type-Token Ratio	.379	.197

- Reasonable performance of some content measures
- Structural measures at most $\bar{F}_1 = 0.554$
- Stylistic measures only slightly better than baseline

Wikipedia Rewrite Corpus

Performance within and across dimensions

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

- **Content** outperforms structural and stylistic similarity
- Best performance by combination across content and structure:
  - *longest common subsequence (content)*
  - *stopword 10-grams (content)*
  - *character 5-gram profiles (structure)*

Wikipedia Rewrite Corpus

Error analysis

- 15 out of 95 texts have been classified wrongly
- light vs. heavy revision → 67 % of all misclassification
- Annotation study: only “fair” inter-annotator agreement for this distinction

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**

$\bar{F}_1 = 0.811$

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

$\bar{F}_1 = 0.859$

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

$\bar{F}_1 = 0.967$

METER Corpus

Dataset

- Source texts:
  - News sources from the UK press Association (PA)
- Derived texts: articles from 9 newspapers that reused PA source texts.
- 2 domains: *Law & court* and *show business*
- 253 pairs of short texts
- binary classification:
  181 reused (wholly or partially) texts
  72 non-reused texts

METER Corpus

Individual measures vs. combinations

Text Similarity Feature	METER Acc.	METER $\bar{F}_1$
Majority Class Baseline	.715	.417
Ferret Baseline	.684	.535
Content Similarity		
Character 5-gram Profiles	.715	.417
ESA (Wikipedia)	.711	.484
Greedy String Tiling	.755	.645
Longest Common Substring	.719	.467
Resnik	.715	.417
Word 2-grams Containment	.727	.692
Structural Similarity		
Lemma Pair Distance	.715	.417
Lemma Pair Ordering	.715	.417
POS 3-grams Containment	.731	.701
Stopword 3-grams	.715	.417
Stopword 7-grams	.652	.482
Stylistic Similarity		
Function Word Frequencies	.715	.417
Sequential TTR	.715	.417
Sentence Ratio	.755	.625
Token Ratio	.755	.619
Type-Token Ratio	.715	.417

→ Application of individual measures often cannot exceed majority baseline

→ improvement by measure combination

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**

METER Corpus
Comparison to other approaches

- Sanchez-Vega et al. (2010):
  - Length and frequency of common word sequences
  - Relevance of individual words

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**

METER Corpus
Error analysis

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

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**

- 50 out of 253 texts were classified incorrectly
- Cause for many of the 30 errors:
  Lower similarity ⇏ no reuse
  e.g., text length (introduction of new facts, ideas etc.)

  → similarity measures could be computed per section, not per document

  → detection of text reuse for partially matching texts
- Still sufficient performance for providing authors with suggestions of potential instances

Webis Crowd Paraphrase Corpus

Dataset

- 7859 pairs of texts (original book excerpt from the *Project Gutenberg* + paraphrase acquired via crowdsourcing)

  manual assignment:

  - 52% positive samples
    good paraphrases: e.g., synonym use, changes between active and passive voice
  - 48% negative samples
    bad paraphrases: near-duplicates

INTRO

MEASURES

EXPERIMENTS

SUMMARY

Webis Crowd Paraphrase Corpus

Comparison to other approaches

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**

- Burrows et al. (2012):
  10 similarity measures on string sequences

Webis Crowd Paraphrase Corpus

Performance of individual measures

Text Similarity Feature	Webis CPC Acc.	Webis CPC $\bar{F}_1$
Majority Class Baseline	.517	.341
Ferret Baseline	.794	.789
Content Similarity		
Character 5-gram Profiles	.753	.742
ESA (Wikipedia)	.760	.753
Greedy String Tiling	.805	.800
Longest Common Substring	.743	.736
Resnik	.666	.656
Word 2-grams Containment	.801	.797
Structural Similarity		
Lemma Pair Distance	.775	.767
Lemma Pair Ordering	.785	.780
POS 3-grams Containment	.787	.783
Stopword 3-grams	.778	.776
Stopword 7-grams	.753	.750
Stylistic Similarity		
Function Word Frequencies	.727	.719
Sequential TTR	.667	.638
Sentence Ratio	.657	.653
Token Ratio	.778	.774
Type-Token Ratio	.723	.712

- Many measures achieve a very reasonable performance (> 0.7) individually

Webis Crowd Paraphrase Corpus

Performance of measure combinations

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**

- ***Content*** alone is stronger than ***Content+Structure***
- ***Content*** performs as good as Burrows et al. (2012)
- ***Content + Structure + Style***: combination of 16 features

Webis Crowd Paraphrase Corpus
Error Analysis

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

- 15% were classified incorrectly
- 759 false positives are less severe, as the users can still decide on them
- For the other 2 corpora it holds that:
  Higher similarity ⇒ higher degree of reuse
- For Webis:
  Higher similarity is annotated as bad paraphrases (including also empty samples, unrelated texts)

  → highly elaborate definition of positive and negative cases
  → difficult to learn a proper model

Summary
Hypothesis

Hypothesis:
Content alone is not a reliable indicator for text reuse because of possible modifications such as:

- split sentences
- changed order of reused parts
- stylistic variance

Investigation of three characteristic dimensions:
content, ***structure*** and ***style***

Summary
Evaluation

Evaluation based on three datasets:

Wikipedia Rewrite Corpus

METER Corpus

Webis Crowd Paraphrase Corpus

Text reuse can be best detected if measures are combined across dimensions

Summary
Conclusion

- Choice of dimensions should depend on the type of text reuse
  - Stylistic similarity performs poorly on Wikipedia Rewrite Corpus
  - Stylistic similarity performs well on the other 2 datasets
- Dimensions should be addressed explicitly in the annotation process

Summary
Future work

- Consideration of a dimensional representation should benefit in other tasks, e.g.:
  - paraphrase recognition
  - automatic essay grading (might include also measures for grammar analysis, lexical complexity or discourse measures)
- Choice of dimensions is task dependent

Thanks for your attention!

Any questions?

→ All the references used in this presentation can be found in the paper's references