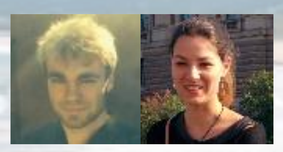
LREC 2018 MIYAZAKI

# Lexical and Semantic Features for Cross-lingual Text Reuse Classification – An Experiment with English and Latin Paraphrases



MARIA MORITZ maria.moritz@stud.uni-goettingen.de
DAVID STEDING dsteding@uni-bremen.de

7-12 May 2018 Miyazaki (Japan)

## OVERVIEW

In this work, we focus on the detection and classification of paraphrastic text reuse in historical texts. We present an experiment of cross-applying models trained for paraphrase recognition on modern English text corpora to historical texts. We analyze the impact of four different features on the resulting reuse-detection accuracy. We find that word embeddings can help to improve accuracy if lexical features (such as the overlap of similar words) fail.

### MOTIVATION

Paraphrastic text reuse classification requires ML-relying techniques. Yet, historical languages often lack enough primary material for certain time periods to adequately train ML classifiers. Consequently, basic NLP techniques (e.g., similarity thresholds over n-gram shingles), which are independent from an advanced training experience, are applicable. To improve this we need to systematically study the performance of such techniques on historical texts and we must understand in what way ancient languages behave differently than contemporary languages when they are transferred and reused paraphrastically.

### RESEARCH QUESTIONS

**RQ1:** *What features support a cross-lingual reuse classification?*

**RQ2:** *What characteristics must a source training text have to enable classification of the target language?*

## FEATURES

**F1** = the relative no. of words that two text excerpts have in common:

$$f1(text1, text2) = \frac{|text1 \cap text2|}{min(|text1|, |text2|)}$$

**F2** = the relative no. of similar words that two texts have in common:

$$f2(text1, text2) = \frac{|text2_{sim2\_text1}|}{|text2|}$$

$|text2_{sim2\_text1}|$ is the no. of words from text2 that fuzzily match at least one word of text1 with a 0.2 ratio of character edits.

**F3 =** the relative amount of words from text2 that are collocations of the words from text1:

$$f3(text1, text2) = \frac{|text2_{sim3\_text1}|}{|text2|}$$

$|text2_{sim3\_text1}|$ is the no. of words from text2 that are collocations of any word from text1. Collocations are calculated within text1 or text2 for each word of the corresponding text wit a max. distance of 2.

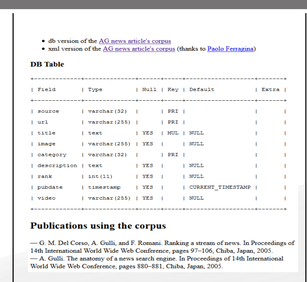
### Extra Feature

F4 = is the cosine of the angle between the averaged word vectors of text1 and text2.

$$vec_{text1} = \frac{\sum_{i=0}^{|text1|} v_{w_i}}{|text1|}, \; vec_{text2} = \frac{\sum_{j=0}^{|text2|} v_{w_j}}{|text2|}$$

$v_{w_i}$ is the word embedding of a word in text1 and text2.

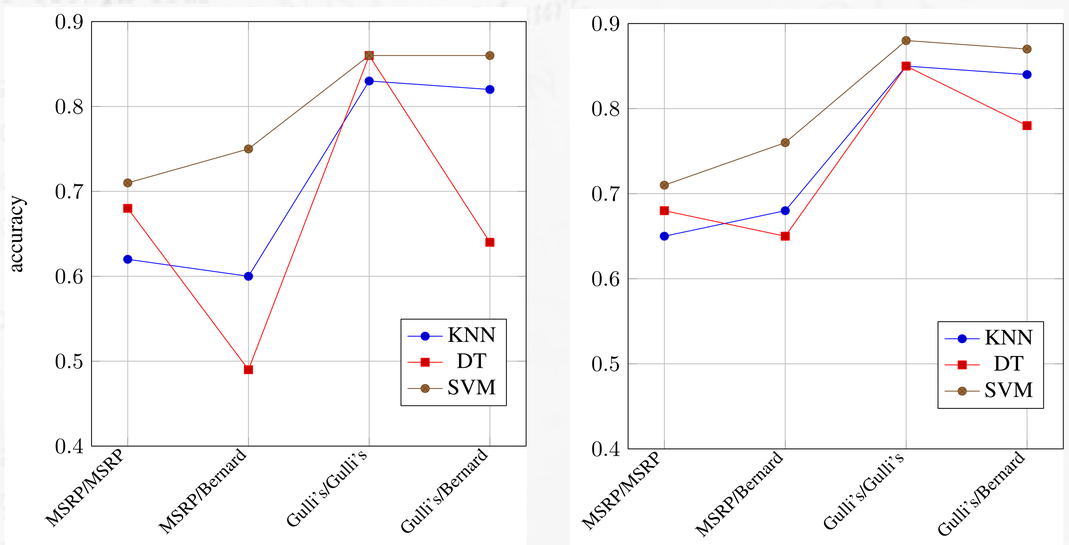
## DATA


| | MSRP | Gulli's | Bernard |
|---|---|---|---|
| **positive** | 2753 | 2656 | 1127 |
| **negative** | 1323 | 2661 | 1126 |
| **text type** | news articles | news articles (headlines & banner) | sermons & treatises |
| **reference** | B. **Dolan** & C. **Brockett**. 2005. Automatically constructing a corpus of sentential paraphrases. In: *Third International Workshop on Paraphrasing*. Asia Federation of Natural Language Processing. | A. **Gulli**. 2005. The anatomy of a news search engine. In: *14th International World Wide Web Conference*, pp. 880–881. Chiba, Japan. | L. **Mellerin**. 2014. New ways of searching with Biblindex, the online index of biblical quotations in early Christian literature. In Clivaz, Gregory, and Hamidovic, editors, DH in Biblical, Early Jewish and Early Christian Studies, pp. 175–192. Brill, Leiden. |

## RESULTS

| train | test | precision | recall | fscore | precision | recall | fscore |
|---|---|---|---|---|---|---|---|
| **KNN** | | **positive** | | | **negative** | | |
| MSRP | MSRP | .74 | .68 | .71 | .42 | .50 | .46 |
| MSRP | Bernard | .62 | .45 | .53 | .58 | .73 | .65 |
| Gulli's | Gulli's | .83 | .81 | .82 | .83 | .85 | .84 |
| Gulli's | Bernard | .82 | .82 | .82 | .83 | .83 | .83 |
| **DT** | | **positive** | | | **negative** | | |
| MSRP | MSRP | .72 | .86 | .78 | .50 | .29 | .37 |
| MSRP | Bernard | .49 | 1.0 | .66 | - | 0.0 | - |
| Gulli's | Gulli's | .88 | .82 | .85 | .84 | .90 | .87 |
| Gulli's | Bernard | .86 | .34 | .48 | .59 | .94 | .73 |
| **SVM** | | **positive** | | | **negative** | | |
| MSRP | MSRP | .72 | .94 | .81 | .62 | .21 | .31 |
| MSRP | Bernard | .96 | .51 | .67 | .67 | .98 | .80 |
| Gulli's | Gulli's | .87 | .84 | .86 | .86 | .88 | .87 |
| Gulli's | Bernard | .87 | .83 | .86 | .84 | .90 | .87 |




(a) Results of initial feature set (b) Results including the new feature

RQ1: We learn that lexical features are useful for cross-lingual classification, and that semantic characteristics support the identification of paraphrastic reuse.

RQ2: We can achieve a well-working classification when the training data is similar to the test data.

*Maria Moritz & David Steding. Lexical and Semantic Features for Cross-lingual Text Reuse Classification: an Experiment in English and Latin Paraphrases (LREC 2018). Miyazaki, Japan, May 7-12, 2018.*

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
eTRAP Electronic Text Reuse Acquisition Project
Universität Bremen

SPONSORED BY THE
Federal Ministry of Education and Research
Grant no. 01UG1509