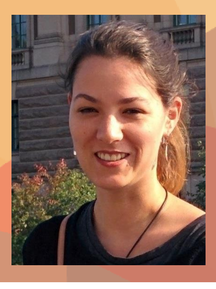
A Method for Human-Interpretable Paraphrasticality Prediction

Maria Moritz[1], Johannes Hellrich[2,3], Sven Buechel[2]



LaTeCH@COLING, August 25 2018, Santa Fe, New Mexico, USA

OVERVIEW

Reuse in historical texts is difficult to identify if it is heavily modified, and algorithmic support for its identification & analysis is limited.

Existing reuse detection techniques can tell if and how frequently a text is modified, while our technique also determines the degree and characteristics of modification (i.e., the “features”) that constitute a reuse, which is an important prerequisite for the analysis of historical text reuse giving scholars hints for deeper investigation.

Our technique is both, human-interpretable and semantically informed setting it apart from recent developments based on distributional semantics which do not allow for easy manual inspection of individual models.

DATA

plagiarism corpus[4]	Bibles[5] (1749-1859) *3 lit. translations* *5 revisions*	Bernard	
ENG	ENG	Medieval LAT	**language**
			POSITIVE
5,000	315,000	990	**training**
1,500	35,000 (15 pairings)	110	**test**
			NEGATIVE
5,000	270,000	990	**training**
1,500	30,000	110	**test**
rand. Sampled (overlap of 4)	(13 pairings)	randomly sampled	
PAN 2010 plagiarism detection challenge (Madnani et al. 2012)	www.biblestudytools.com www.mysword.info Parallel Text Project (Mayer & Cysouw 2014).	Biblindex (Mellerin 2014)	**reference**

[4] Icons made by Freepik from Flaticon is licensed by CC 3.0 BY
[5] Icons made by Smashicons from www.flaticon.com is licensed by CC 3.0 BY

OUR METHOD

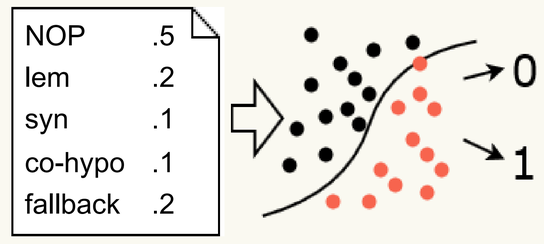
word alignment

they mount up with wings as eagles walk
lower | syn | NOP | syn | NOP | lem
They raise up the pinion as eagles go

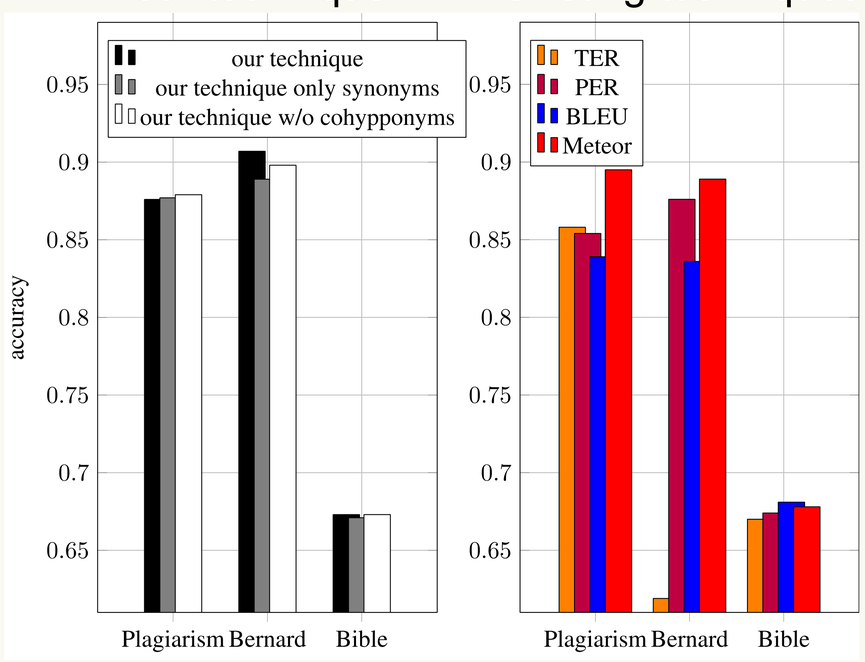
operation modeling

operation	example
No OPeration necessary	above, above
lower-casing match	LORD, Lord
normalizing match	desireth, desires
lemmatizing match	mine, my
derivation match	help, helper
short **edit dist**ance match	Phinehas, Phinees
words are **syn**onyms	went, departed
word1 is **hyper**nym of word2	coat, doublet
word1 is **hypo**nym of word2	spears, arms
words are **co-hypo**nyms	steps, feet
fallback	—

prediction of paraphrastical reuse
using operation frequencies as features



RESULTS



Co-hyponym replacement is common in historical text reuse (Moritz et al. 2016). Our method is well suited to predict this reuse, because it considers such substitutions. Meteor outperforms all other methods on the plagiarism data.

DISCUSSION

We presented a method for paraphrase detection that describes reuse based on the frequency of specific modification operations, especially semantic relations beyond synonyms.

As such, our method is especially useful for applications in the humanities as operation frequencies and their feature weights are open to manual inspection.

Modeling reuse in historical text using semantic relations as we propose, achieves results comparable to using features derived from machine translation metrics.

In future work, we plan to tune parameters and to qualitatively analyze weaknesses of our method.

1) mmoritz@etrap.eu

2) johannes.hellrich@uni-jena.de
sven-buechel@uni-jena.de

3) johannes.hellrich@uni-jena.de

SPONSORED BY THE
Federal Ministry of Education and Research
DFG Deutsche Forschungsgemeinschaft
MODELL ROMANTiK