A Method for Human-Interpretable Paraphrasticality Prediction

Maria Moritz[1], Johannes Hellrich[2,3], and Sven Buechel[3]
[1]Institute of Computer Science, University of Göttingen, Germany
[2]Graduate School “The Romantic Model”, Friedrich-Schiller-Universität Jena, Germany
[3]JULIE Lab, Friedrich-Schiller-Universität Jena, Germany
mmoritz@gwdg.de

Abstract

The detection of reused text is important in a wide range of disciplines. However, even as research in the field of plagiarism detection is constantly improving, heavily modified or paraphrased text is still challenging for current methodologies. For historical texts, these problems are even more severe, since text sources were often subject to stronger and more frequent modifications. Despite the need for tools to automate text criticism, e.g., tracing modifications in historical text, algorithmic support is still limited. While current techniques can tell if and how frequently a text has been modified, very little work has been done on determining the degree and kind of paraphrastic modification—despite such information being of substantial interest to scholars. We present a human-interpretable, feature-based method to measure paraphrastic modification. Evaluating our technique on three data sets, we find that our approach performs competitive to text similarity scores borrowed from machine translation evaluation, being much harder to interpret.

1 Introduction

Why is Text Reuse important? The term text reuse refers to the repetition of a text within a new context. Examples are citations, paraphrases of a text, allusions, or even cases of cross-linguistic reuse in the form of translations. In the humanities context, the detection of text reuse helps tracing down lines of transmission, which is essential to the field of textual criticism (Büchler et al., 2012). Text reuse detection can also help consolidating today’s digital libraries by assuring the consistency of content by inter-linking related documents (Schilling, 2012).

Background: To this date, a lot of effort has been put into the investigation of detecting *plagiarism*, a special kind of text reuse. However, while constantly improving (see Ferrero et al. (2017)), contemporary detection techniques are still quite unreliable when text is heavily modified. Historical text is even more challenging through incompleteness, copying errors, and evolution of language. Thus, only limited algorithmic support exists for the identification and analysis of (especially paraphrastic) repetition in such documents.

While existing reuse detection techniques are able to tell *if* and *how frequently* a text has been modified, it is important to also determine the degree and characteristics of paraphrastic modification, i.e., the “features” that constitute a given modification. As such, understanding type and degree of reuse is an important prerequisite for enhancing reuse detection techniques for historical texts as well as giving scholars hints for deeper investigation. In this work, we present a technique to measure paraphrastic modification which is both human-interpretable and semantically informed. This interpretability sets our method apart from recent approaches based on distributional semantics which do not allow for easy manual inspection of individual model decisions (Wieting et al., 2015).

We already investigated descriptive characteristics of paraphrasing in a specific humanities use case (Moritz et al., 2018). We found changes in inflection, synonym replacement and co-hyponym replacement to be the most frequent paraphrastic modifications, thus supporting the feasibility of feature-based approaches to this problem.

Method and Questions: We measure the degree of modification based on a list of *modification operations* that we count in a prioritized order based on relations between aligned, parallel sentences. These

Proceedings of Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 113–118
Santa Fe, New Mexico, USA, August 25, 2018.

relationships between two words can range from exact copy (no operation necessary) to co-hyponymy, see Table 1. Compared to scores such as Meteor that make use of synonymy, but do not model other relationships, our score also includes information on hypernymy, hyponymy, and co-hyponymy. This is especially useful in historical text, since meaning and, therefore, relationships change over time. The order in which these operations are counted is intuitive and follows the usual prepossessing steps that one would perform to reduce variance in a text corpus. Table 2 shows an example of the alignment output, thus illustrating our method. The relative frequencies of the operations then serve as input features for a binary classifier.

In this contribution we investigate, how our human-interpretable method compares against text similarity metrics borrowed from machine translation evaluation (also serving as input for a classifier). In particular, we examine the performance of those approaches for semantic equivalency in: (**RQ1**) a modern English paraphrase corpus; (**RQ2**) a parallel Bible corpus; and (**RQ3**) a medieval Latin text reuse dataset.

Operation	**Example Pair**
no operation necessary	*above, above*
lower-casing match	*LORD, Lord*
normalizing match	*desireth, desires*
lemmatizing match	*mine, my*
derivation match	*help, helper*
short edit distance match	*Phinehas, Phinees*
words are synonyms	*went, departed*
word1 is hypernym of word2	*coat, doublet*
word1 is hyponym of word2	*spears, arms*
words are co-hyponyms	*steps, feet*
other	—

Table 1: Overview of transformation operations.[2]

2 Related Work

Surface Feature Approaches: Levenshtein's (1966) edit distance, which is based on character-level removal, insertion, and replacement operations, can be considered as one of the earliest works to measure text similarity. Büchler et al. (2012) use overlapping bi-grams to maximize recall in a reuse detection task of Homeric quotations, showing a good precision of more than 70% at the same time. Those techniques rely on surface features (token and character-level) only. Thus, our proposed method differs by also incorporating semantic information (lexico-semantic relationships between aligned word pairs).

Semantic Approaches: Computing the semantic similarity between two sentences is a popular task in NLP (Xu et al., 2015). Osman et al. (2012) present a plagiarism detection technique based on semantic role labeling. They analyze text by identifying the semantic space of each term in a sentence and find semantic arguments for each sentence. They also assign weights to the arguments and find that not all of them affect plagiarism detection. Techniques from the field of paraphrase detection can be used for e.g., sentence similarity, entailment, and sentiment classification. Wieting et al. (2015) use embedding models to identify paraphrastic sentences in such a mixed NLP task employing a large corpus of short phrases associated with paraphrastic relatives. Their simplest model represents a sentence embedding by the averaged vectors of its tokens, the most complex model is a long short-term memory (LSTM) recurrent neural network. They find that the word averaging model performs best on sentence similarity and entailment, and the LSTM performs best on sentiment classification. Although these methods generally show good results, they typically allow no manual inspection of why a specific judgment is made and are thus ill-suited for applications in the humanities.

Approaches Based on Machine Translation (MT) Evaluation Metrics: Madnani et al. (2012) conduct a study on the usefulness of automated MT evaluation metrics (e.g., BLEU, NIST and Meteor) for the task of paraphrase identification. They train an ensemble of different classifiers using scores of MT metrics as features. They evaluate their model on two corpora for paraphrase and plagiarism detection, respectively, finding that it performs very satisfyingly. This approach to paraphrase and plagiarism detection based on MT metrics combines surface and semantic features since Meteor incorporates synonymy information (see below). Yet, the number of semantic features used is limited and so is also the interpretability of this approach.

[2] Note that we distinguish operations with and without changes in part-of-speech, hence in total we work with twenty one different operations.

OP	NOP	NOP	cohypo	NOP	syn	NOP	fallback	NOP	NOP	NOP	NOP	NOP	syn	fallback
token s1	It	is	unlawful	he	contends	to	co-operate	with	any	one	who	is	doing	wrong
token s2	It	is	law	he	argues	to	-	with	any	one	who	is	performing	-

Table 2: Example of operation (feature) based alignment. Features here are no operation =9/14 (NOP), cohyponym =1/14 (cohypo), synonym=1/14 (syn), and fallback =2/14.

3 External Resources

Tools: We use BabelNet (Navigli and Ponzetto, 2012) as a resource for retrieving relationships between *English* words, namely synonym, hypernym, hyponym and co-hyponyms. For the *Latin* evaluation dataset we use the Latin WordNet by Minozzi (2009).[3] To normalize, lemmatize, and part-of-speech (POS) tag the text data, we use MorphAdorner,[4] a tool for lemmatizing Early Modern English text which is also applicable to contemporary English. For the *Latin* dataset, we use the respective TreeTagger model (Schmid, 1994). To align sentences from a given parallel corpus on the token level we use the Berkeley Word Aligner (DeNero and Klein, 2007), a statistical, unsupervised word aligner originally designed for machine translation.

Contemporary Paraphrase Detection: As a gold dataset for paraphrase prediction, we use an English corpus of semantically equivalent sentences that originates from the PAN 2010 plagiarism detection challenge. Starting from text that was aligned on the *paragraph* level, Madnani et al. (2012) generated a set of aligned *sentences* using heuristics. Negative pairings were created by sampling non-aligned sentences with an overlap of four words. The training and test set comprise 10,000 and 3,000 sentence pairs, respectively. Both sets are balanced regarding positive and negative labels.

Bible Translation Class Prediction: We use a parallel corpus of eight English Bible translations that we gathered from three sources.[5] We split them in two classes: literal translations—those being directly translated from the primary languages Hebrew and Ancient Greek coming with rich linguistic diversity—and translations that mainly follow the translation tradition of the Anglican Church (standard). Table 3 lists the detailed edition names accompanied by its publishing date and its class. For the experiments we extract parallel verses from two different editions and try to predict if they come from the same or different translation classes (literal vs. standard).

Bible	**Published**	**Class**
Douay-Rheims Challoner Rev. (DRC)	1749-1752	standard
King James Version (KJV)	1769	standard
The Webster Bible (WBT)	1833	standard
Darby Bible (DBY)	1867-1890	standard
English Revised Version (ERV)	1881-1894	standard
English Septuagint (LXXE)	1851	literal
Young's Literal Translation (YLT)	1862	literal
Smith's Literal Translation (SLT)	1876	literal

Table 3: Overview of English Bible translations used.

Latin Reuse Detection: Excerpts from a total of twelve works and two work collections from the 12th century Latin writer Bernard of Clairvaux constitute our third dataset. The team behind the Biblindex project (Mellerin, 2014)[6] manually identified 1,100 instances of text reuse in these writings and bundled them into a corpus. Every instance of reuse relates to a Bible verse from the Biblia Sacra Juxta Vulgatam Versionem and is typically half as long as the original verse. Negative training data of equal size were obtained by randomly shuffling the initial dataset.

[3] http://multiwordnet.fbk.eu/english/home.php

[4] http://morphadorner.northwestern.edu/

[5] http://www.biblestudytools.com/, www.mysword.info/, Parallel Text Project (Mayer and Cysouw, 2014).

[6] http://www.biblindex.mom.fr/

4 Methods

Our method relies on the relative frequencies of modification operations (see Table 1) in an aligned sentence pair which later serve as features for a classifier:

$$x_i = \frac{\#o_i}{\sum_{j=0}^{m} \#o_j} \tag{1}$$

where x_i is the relative frequency of a modification operation i in an aligned sentence pair, m is the number of features, and o_i is the absolute frequency of operation i.[7] Our method, hence, can be understood as a collection of features that are represented as relative frequencies of edits obtained from empirical values. These features are used as input to a maximum entropy classifier to predict if two sentences are paraphrases of each other. MaxEnt was chosen due to its simplicity, relying on a linear combination of features. Thus feature weights can be roughly interpreted as importance of the respective modification operation after fitting the model. Recall the example alignment presented in Table 2 illustrating the high interpretability of our approach. Our method will be denoted "multi_f" (multiple features) for the remainder of this paper.

We evaluate our method by comparing it to several reference methods based on machine-translation evaluation metrics.[8] To adapt these to our different paraphrase detection tasks, the source Bible provides the reference sentence (ref) and the target Bible (and Bernard's reuse respectively) provides the system output (sys). From the Gold corpus, also the source text (numbered in the repository with 1, see Madnani et al. (2012)) serves as reference, and the paraphrastic reuse of it (numbered with 2), provides the system output.

Reference Methods: Often, machine translation metrics are based on simple edit distance measures. Unlike simple word error rate (WER; Su et al. (1992)), which depends on a strict word order, the position-independent error rate (**PER**; Tillmann et al. (1997)) uses a bag-of-words approach. Popović and Ney (2007) define PER based on counts of independent words that system output and reference sentence have in common. We adapt their document-wide score to the sentence level:

$$PER = \frac{1}{2 \cdot N_{ref}}(|N_{ref} - N_{sys}| + \sum_{e} |n(e, ref) - n(e, sys)|), \tag{2}$$

where N_{sys} is the length (in words) of the target reuse text—in MT a.k.a. *the system output* version of a text—and N_{ref} is the length of the source text—in MT a.k.a. *the reference sentence* for a system output—, and $n(e, \mathit{ref})$ is the frequency of a given word e in the reference sentence.

The translation edit rate (**TER**; Snover et al. (2006)) is the number of edits that a system output should undergo so that it matches a reference sentence. TER[9] is normalized by the length of the reference input. Following Papineni et al. (2002), we define a sentence-based **BLEU** score:

$$BLEU = BP \times \exp(\sum_{n=1}^{N} \frac{1}{N} \log p_n) \tag{3}$$

where N is the maximum n-gram size, which we set to 2. p_n is a precision score that is calculated based on n-grams in both, source and target texts (see Papineni et al. (2002)). We omit BLEU's brevity penalty which would otherwise dominate our sentence level analysis.

The last measure we consider is **Meteor** 1.5 (Denkowski and Lavie, 2014). Meteor especially differs from other scores by considering not only precision, but also recall. It further takes synonymy and paraphrases into account. Meteor introduces so called matchers that are represented by exact match, stem match, synonym match or paraphrase match. The hypothesis (system) and reference texts h and r are split into content words h_c and r_c, and function words h_f and r_f. Precision and recall measures are then used to

[7] $m = 18$ because we dropped three features after development experiments, i.e., no operation necessary, lemmatization match and hypernym match.

[8] We had to change some of the metrics to capture distance (instead of a similarity) by using their complement.

[9] We use the implementation from: `www.cs.umd.edu/\%7Esnover/tercom/`, acc. May '18

determine the harmonic mean F_{mean}. Together with a fragmentation penalty that measures the degree of chunks, the Meteor score is calculated by $Meteor = (1 - penalty) \times F_{mean}$.
Similar to Madnani et al. (2012) we use these MT scores separately in a classification task to predict paraphrasticality where the respective MT score is fed into a MaxEnt classifier as only feature.

5 Results

Detecting Paraphrases (RQ1): Using the relative operation count from the alignment as features in a classification task, we determine the classification accuracy of our approach on the gold corpus. We run a maximum entropy classifier on our operation features. The results in Table 4 show that Meteor performs best on that task, followed by our approach.

name	**RQ1**	**RQ2**	**RQ3**
multi_f only WN	87.6	67.2	-
multi_f synonyms only	87.7	67.1	88.9
multi_f w/o cohyponyms	87.9	67.3	89.8
multi_f (all features)	87.6	67.3	**90.7**
TER	85.8	67.0	61.9
PER	85.4	67.4	87.6
BLEU	83.9	**68.1**	83.6
Meteor	**89.5**	67.8	88.9

Table 4: Accuracy in solving our three tasks.

Predicting Translation Classes (RQ2): Here we want to determine if two aligned Bible editions are of the same translation class (labeled with 0), or of different classes (labeled with 1); we distinguish between standard vs. literal translations. We use the operation counts based on two aligned verses as features in this binary classification task. Our operations equip us with a fine-grained description of the degree of modification of two text excerpts. The Bible corpus is a suitable source for measuring the degree of modification, since it holds a broad variety of paraphrastic reuses. To estimate a human judgment of deviation, we assume that standard translations are more homogeneous to each other (based on their evolution history) than literal translations that demand for more creative language use (Moritz et al., 2018). We use 10-fold-cross validation on the shuffled dataset. The results in Table 4 show that all methods under consideration perform comparably well. We also find that our proposed method suffers from a drop of accuracy when semantic features are ablated. When only WordNet, not BabelNet, is used for identifying lexico-semantic relations, performance increases slightly, which we attribute to noise that comes with using BabelNet.
Detecting Latin Reuse (RQ3): Finally, we predict reuse in the medieval Latin dataset. Moritz et al. (2016) found out that co-hyponymy (besides synonymy) can be a common means of substitution in reuse, especially in medieval texts. Consequently, our method is well suited for this task, because it considers semantic relations beyond synonymy.[10] Again, we use 10-fold cross-validation on the shuffled dataset. Table 4 shows that dropping features such as co-hyponyms indeed worsens the accuracy. The low score of TER may be explained by the fact that this metric's normalization term is based on the length of the *reference* version of a sentence. In our setup the Bible verse is the reference and the system output is the reuse. The reuse however, is often shorter than the Bible verse (see above).

6 Discussion and Conclusion

We presented a method for paraphrase detection that describes reuse based on the frequency of specific modification operations and is thus easily interpretable for humans. We showed that modeling reuse in historical text using semantic relations beyond synonyms achieves results comparable to using features derived from machine translation metrics. Moreover, our method is especially useful for applications in the humanities as operation frequencies, their respective feature weights, and, by extensions, individual model decisions are open to manual inspection. In future work, we plan to tune parameters and to qualitatively analyze weaknesses of our method (e.g., due to the tools used for pre-processing and alignment).

Acknowledgements

This research was partially funded by the German Federal Ministry of Education and Research (grant 01UG1509) and the Deutsche Forschungsgemeinschaft (grant GRK 2041/1).

[10] Note that Meteor only contains synonym data in English, which can also influence its accuracy.

References

Marco Büchler, Gregory Crane, Maria Moritz, and Alison Babeu. 2012. Increasing recall for text re-use in historical documents to support research in the humanities. In *Theory and Practice of Digital Libraries*, pages 95–100.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *ACL 2007*, pages 17–24.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jérémy Ferrero, Laurent Besacier, Didier Schwab, and Frédéric Agnès. 2017. Using word embedding for cross-language plagiarism detection. In *EACL: Volume 2, Short Papers*, pages 415–421.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, pages 707–710.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *NAACL 2012*, pages 182–190.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *LREC 2014*.

Laurence Mellerin. 2014. New ways of searching with biblindex, the online index of biblical quotations in early christian literature. In *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*. Brill, Leiden.

Stefano Minozzi. 2009. The latin wordnet project. In Peter Anreiter and Manfred Kienpointner, editors, *Innsbrucker Beitrge zur Sprachwissenschaft*, pages 707–716.

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In *EMNLP 2016*, pages 1849–1859.

Maria Moritz, Johannes Hellrich, and Sven Buechel. 2018. Towards a metric for paraphrastic modification. In *Digital Humanities 2018*, pages 457–460.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Albaraa Abuobieda. 2012. An improved plagiarism detection scheme based on semantic role labeling. *Applied Soft Computing*, 12(5):1493–1502.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.

Maja Popović and Hermann Ney. 2007. Word error rates: Decomposition over pos classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48–55.

Virginia Schilling. 2012. Introduction and Review of Linked Data for the Library Community, 20032011. `http://www.ala.org/alcts/resources/org/cat/research/linked-data`.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *COLING 1992: Volume 2*, pages 433–439.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Proceedings of the fifth European Conference on Speech Communication and Technology*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *SemEval 2015*, pages 1–11.