An Automated Approach to Model the Transformation Process of the Reuse in Bernard de Clairvaux: How Do Lexical Resources Help?

Maria Moritz
mmoritz@etrap.eu
University of Goettingen, Germany

Marco Büchler
mbuechler@etrap.eu
University of Goettingen, Germany

Abstract

To fortify the research of automated, historical text reuse detection, it is necessary to investigate the way in which a text is reused (e.g., verbatim, paraphrased) in order to understand the broader context of a reuse. Our long-term goal is to build a formal theory behind reuse transformations. We have previously investigated two datasets of Bible reuse to analyze how reuse is modified and how linguistic resources support this. In this work, we investigate the ratio of non-literal text reuse, and we measure to which extent the Ancient Greek WordNet—which also contains Latin WordNet—and BabelNet can support identifying lexical relations in Latin reuse excerpts. In doing so, we also show the lack and need of resources for ancient data.

Introduction

The automated detection of historical text reuse is still in its early stages. To reinforce its research, it is necessary to investigate the way a text is reused in order to understand the broader context. Here is where the necessity of lexical resources supporting this task comes in, especially when a text is non-literally reused, and words are substituted with semantic equivalents, such as synonyms or other semantically similar words. Our long-term goal is to formally model reuse transformations. The analysis of the amount and type of substitutions of words with lexically related words enables insights into how text is reused. Applying these insights into future development of detection methods helps to improve them. We have previously investigated two datasets of Bible reuse, trying to understand how reuse is modified (when operations are performed on word pairs) and how linguistic resources support this task. To achieve this, we need to study more and different cases of reuse. In this short paper, we propose and report on work that extends the number of reuse excerpts we investigated in previous work (Moritz et al., 2016), and take another linguistic resource, BabelNet into account. We aim at investigating the current state of lexical resources' support for a Latin reuse dataset. We compare the support we can get from an additional lexical resource to previous results. Specifically, we investigate BabelNet (BN) (Navigli and Ponzetto, 2012), a multiple resource network pulling from different sources, and we compare the reuse detection support (how many words are covered) between BN and Ancient Greek WordNet (AGWN) (Bizzoni et al., 2014), which also contains Latin WordNet (Minozzi, 2009). Both are recently developed resources and the most common for the Latin language. BabelNet is produced from a range of different, contemporary sources, such as Wikipedia and Wikidata. We are interested in the extent to which BabelNet is able to cover words and relations from an ancient reuse dataset. We are especially curious about what words are still supported by current resources. Our ultimate goal is to simulate a transformation process that also supports non-literal reuse. This can help to model the changes that were applied to an ancient text during its reuse history.

Background

The field of automatically detecting historical text reuse is still in its early stages. To date, Büchler (2013) combines state-of-the-art NLP techniques to address reuse detection scenarios for historical texts, ranging from near copies to text excerpts with a minimal overlap, using a method, which selects n-grams from an upfront pre-segmented corpus. While the approach can discover historical and modern text reuse language-independently, it requires a minimal text similarity. Recognizing modified reuse is difficult in general. Alzahrani et al. (2012) study plagiarism detection techniques, such as n-gram-, syntax-, and semantics-based approaches. However, as soon as reused text is modified (e.g., word substitution), most systems fail. Finally, lexical resources support the identification of relationships between words, but they are not free from issues (Jing, 1998) that can appear when they are used to adapt a general lexicon to a specific domain (Miller et al., 1990).

Data

Our dataset contains excerpts from twelve works—mainly sermons and treaties (Literature)—and two work collections—sermons and letters—from the Latin writer Bernard of Clairvaux (c.f.,

Moritz et al., 2016). All those texts come from the *Sources Chrétiennes* collection (c.f., Mellerin, 2014) (changes in format and orthography may be inserted by the editor). The Biblindex project (Mellerin, 2014) extracted over thousand Bible reuse exerpts from these works, each of which points to a Bible verse. We use the Latin Bible from Biblindex, called *Biblia sacra juxta vulgatam versionem* (Weber R., 1969) to link the excerpts to the respective Bible verses. We come up with 1,128 unique reuse-to-bible-verse pairs. Table 1 shows one example.

	Bible verse	Bernard reuse
Proverbs 18 3	impius cum in profundum venerit peccatorum contemnit sed sequitur eum ignominia et obprobrium	Impius , cum venerit in profundum malorum , contemnit
English	The wicked man, when he has come into the depth of sins, despises: but ignominy and reproach follow him.	The wicked man, when he has come into the depth of evils, despises

Table 1: Example of reuse

Methodology

We use AGWN, which is automatically constructed from Greek-English digitized lexicons, which again were provided by the Perseus Project (Crane, 1985) and also aligns to Minozzis Latin WordNet (Minozzi, 2009). BabelNet (Navigli and Ponzetto, 2012) is a multilingual semantic network that integrates lexicographic and encyclopedic knowledge from WordNet (Fellbaum, 1998), Wikipedia, and others. We further use lemma lists from the Biblindex project, as well as the Latin lemma list from the Classical Language Toolkit (CLTK), which is available in the online GitHub repository of the CLTK (Johnson et al., 2014 2016), to increase the hit rate when querying both resources.

To model the transformation in-between two text excerpts, we define replacement operations (OPs) (see Table 2) that represent the transformation of a reuse to the Bible verse it refers to, as well as an algorithm that identifies those operations between word pairs of a reuse and a Bible verse in a prioritized order. Our algorithm first finds all possible operations for a reuse word, and then applies the most literal operation using the counterpart Bible verse word, which fulfills this operation. This means that if no perfectly or lemmatized matching word is found, relationships of semantic closeness (such as synonyms) for a given word are retrieved. We call the group of semantic operations non-literal operations (c.f., Table 3). We apply our algorithm (which identifies the operations) on Bernard's reuse, first using the relationships queried from AGWN and second, using BabelNet. Afterwards, we show which operations are identified, and calculate a support value for both processes.

operation	description	example
NOP(reuse_word, orig_word)	Original and reuse word are equal.	*NOP(maledictus,maledictus)*
upper(reuse_word, orig_word)	Word is lowercase in reuse and uppercase in original.	*upper(filio,Filio)*
lower(reuse_word, orig_word)	Word is uppercase in reuse and lowercase in original.	*lower(Gloriam, gloriam)*
lem(reuse_word, orig_word)	Lemmatization leads to equality of reuse and original.	*lem(penetrat, penetrabit)*
repl_syn(reuse_word, orig_word)	Reuse word replaced with a synonym to match original word.	*repl_syn(magnificavit, glorificavit)*
repl_hyper(reuse_word, orig_word)	Word in bible verse is a hyperonym of the reused word.	*hyper(cupit, habens)*
repl_hypo(reuse_word, orig_word)	Word in bible verse is a hyponym of the reused word.	*hypo(dederit, tollet)*
repl_co-hypo(reuse_word, orig_word)	Reused word and original have the same hyperonym.	*repl_co-hypo(magnificavit, fecit)*
lemma_missing(reuse_word, orig_word)	Lemma unknown for reuse or original word.	*lemma_missing(tentari, inlectus)*
no_rel_found(reuse_word, orig_word)	Relation for reuse or original word not found in word net.	*no_rel_found(gloria, arguitur)*

Table 2: List of operations and corresponding examples (cf. Moritz et al., 2016)

Results

Table 3 shows the identified operations. Using AGWN, we encounter a high ratio of synonyms (repl_syn), a lot of co-hyponyms and a significant number of hyperonyms and hyponyms. With BabelNet these figures are about a tenth as high. Table 3 shows that the values for **NOP**, **lower** and **lem** (matching words, and words with same lemma) slightly differ in-between both word nets. This is caused by a design decision of our algorithm, which pragmatically permits to reassign a word when it already was used in an association with an earlier word.

	literal				non-literal				unclassified		
	NOP	upper	lower	lem	repl_syn	repl_hyper	repl_hypo	repl_co-hypo	no_rel_found	lem_missing	total
AGWN	4521	1	396	770	397	125	124	316	2470	450	9570
BN	4524	1	397	771	25	21	36	112	3233	450	9570

Table 3: Absolute numbers of operations identified

Fig. 1 shows that AGWN outperforms BabelNet in identifying semantic relations, which represent non-literal text reuse, because these ratios are much lower for BabelNet than for AGWN. We further encounter three significant descents: between 0% and 10%, 30% and 40%, and 50% and 60%. Looking into samples deeply, we find three patterns: i) the more semantic related words are replaced in a reuse, the more likely it is an allusion or analogy, and the less paraphrased or verbatim it is; ii) short allusions are better covered by the Latin synsets than paraphrases with a high ratio of semantic related words; iii) paraphrases with a high literal ratio are covered best. We summarize that both word nets cover paraphrased reuse to a certain extent of replaced words, and AGWN better identifies allusions.

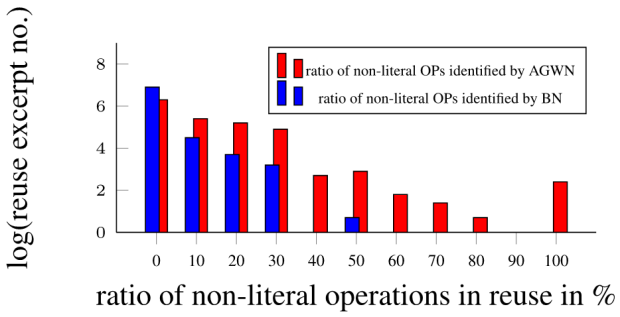

Figure 1: Ratio of non-literal (semantic) operations, aggregated in 10%-steps in relation to the whole reuse length. The reuse number is displayed logarithmically due to clarity reasons.

Lastly, we calculate a support value, which determines the ratio of non-literal operations (c.f., Table 3) compared to them including unsuccessful resource look-ups (no_rel_found) in both, AGWN and BN. For AGWN this value is about 28%, for BabelNet about 6%. Both values are to be understood as lower bounds, because often there is no reasonable relationship in-between two words.

Even if BN coverage is poor, its results tell us, which words of a dataset of medieval, Biblical Latin and Latin of the church fathers are prevailed in a current resource. Some examples are words such as **gloria** (glory) (contained in 17 synsets), **corona** (crown) (contained in 10 synsets), or **nemo** (nobody) (contained in 4 synsets).

Conclusion

We identified the ratio of non-literal reuse in a Latin dataset and showed the support of two lexical resources. Our results show that language resources for Latin reuse are limited and that only a small part of the required coverage is supported. This result raises awareness for the lack of resources for ancient data, despite the growth of language resources for modern languages. Our future work includes refining our operation set, analyzing more languages, increasing the size of our datasets, and investigating probability measures for those data in lexical hierarchies. Since lexical resources will never completely cover the vocabulary at hand, we further consider the application of a form of word embedding.

Acknowledgements

We thank Laurence Mellerin for providing the dataset we used, and for advice on its content. Our work is funded by the German Federal Ministry of Education and Research (grant 01UG1509).

Bibliography

Alzahrani, S. M., Salim, N., and Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. Trans. Sys. Man Cyber Part C, 42(2):133–149.

Bizzoni, Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., and Crane, G. (2014). The making of Ancient Greek WordNet. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).

Büchler, M. (2013). Informationstechnische Aspekte des Historical Text Re-use (English: Computational Aspects of Historical Text Re-use. Ph.D. thesis, Leipzig University, Germany).

Crane, G. (1985). Perseus digital library. http://www.perseus.tufts.edu/hopper/.

Fellbaum, C. (1998). WordNet: An electronic lexical database: Bradford book.

Jing, H. (1998). Usage of WordNet in natural language generation. In Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems (COLING-ACL'98). Columbia University Academic Commons.

Johnson, K.P., Burns P.J., Hollis, L., Pozzi, M., Shilo, A., Margheim, S., Badger, G., and Bell, E. (2014–2016). Cltk: The classical language toolkit. https://github.com/cltk/cltk.

Mellerin, L. (2014). New ways of searching with Biblindex, the online index of biblical quotations in early Christian literature. In Claire Clivaz, Andrew Gregory, and David Hamidovic, editors, Digital Humanities in Biblical, Early Jewish and Early Christian Studies, chapter 11, pages 175–192. Brill, Leiden.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.J. (1990). Introduction to WordNet: An on-line lexical database. International Journal of Lexicography (special issue), 3(4):235–312.

Minozzi, S. (2009). Innsbrucker Beitrge zur Sprachwissenschaft, volume 137, chapter The Latin WordNet Project, pages 707–716. Institut für Sprachen und Literaturen der Universität Innsbruck, Innsbruck.

Moritz, M., Wiederhold, A., Pavlek, B., Bizzoni, Y., and Büchler, M. (2016). Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In Empirical Methods in Natural Language Processing (EMNLP'16), Austin, TX, USA. Association for Computational Linguistics.

Navigli, R., and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. 193:217–250.

Weber, R., Gribomont, J., Fischer, B., Eds. (1969) 1969, 1994, 2007. Biblia Sacra Juxta Vulgata Versionen. Deutsche Bibelgesellschaft.