"Reuse" of Biblical Quotes in Swedish 19th Century Fiction

Dimitrios Kokkinakis and Mats Malm

Introduction

Multifaceted relations between texts can be complex, abstract, diverse or subtle. Digital humanists are interested in identifying pairs of text passages likely to contain substantial overlap and empirically supporting (hopefully) new interpretations of historical texts. For instance, Cordell [3] discusses how digital interpretive tools can help make better sense of enlarged bibliographies, and the continuous digging into digital archives promises to effect exciting revisions to our literary history. Intertextual similarities [12] between *historical* texts embrace a larger set of morphological, linguistic, syntactic, semantic and copying variations, thus adding a complication to text-reuse detection. Recycled text chunks are frequently only small portions of a document and may be significantly modified [4,5]. Older language variants and dialects are less standardized; their evolution spanning centuries [1]; unlike today, e.g. verbatim quotes in older texts were not visually enclosed in quotation marks, making it hard for us to discern reuse from 'original' text; some authors quote other authors we know nothing about or whose works do not survive. Moreover, spelling and orthographic variations as well as OCR-errors can be problematic for the identification of (historical) text reuse [1,2]. Therefore, the task of detecting text re-use is challenging with NLP having a major role to play in this process. NLP techniques to discover intertextual similarities between historical texts is a major topic of considerable interest among scholars from both a theoretical and practical point of view. Biological *sequence alignment* is one available method also used for detection of similar passages in text collections [2,5,7,9,10]. We use the *Pairwise Alignment for Intertextual Relations* (PAIR[11]), a simple implementation of sequence alignment for text analysis which supports one-against-many comparisons.

Material, Experiments and Parameter Tuning

The Charles XII Bible translation into Swedish (a translation completed in 1703 and remained the official Swedish Bible translation until 1917) and the content of the Swedish prose fiction (Spf) are used in this study. *PAIR* can efficiently find similar text between a query document and a pre-indexed corpus. For each document in a collection a number of *n-grams* or *shingles* is generated and a number of parameters (such as *minimum span match*) can be tuned. The approach uses (overlapping) shingle sequences as the basic text unit and the evaluation of document pairs is performed by the intersection of n-gram features in these shingles which can be substrings or blocks of *n* words [8,9,10]. We tried different tuning parameter settings, such as the *shingle size* (i.e. the number of words to include in each shingle). An example of a preprocessed shingle tri-gram generation for the fragment: *Och han grep drakan, den gamla ormen, som är djefvulen och Satan...* (i.e. 'And he seized the dragon, that ancient serpent, who is the devil and Satan…') will be: *grep_drakan_gamla*; *drakan_gamla_ormen*; *gamla_ormen_djefvulen*; *ormen_djefvulen_satan*; here function words *Och*, *han*, *den*, *som*, *är* and *och* have been removed.

Results and Evaluation

We tried slightly different configurations A-C. A: shingle size 3; max gap 5; min pair of shingles 3; B: shingle size 3; max gap 5; min pair of shingles 2; and C: shingle size 3; max gap 12; min pair of shingles 3; and no stop words. Configuration A showed the most conservative results with the highest precision in a qualitative style evaluation (i.e., a close reading of the results). Out of the 300 volumes in Spf, configuration A could detect Bible reuses in 53 volumes, configuration B 139 and configuration C 183. Configurations B and C could caption more re-used segments but with the cost of also capturing well formulated but trivial associations such as: (i) *Hon <reste sig upp, och gick in> till grefvinnan* 'She stood up, and went in to the countess' (Spf-file lb3040528) and …*<reste han sig upp, och gick in> i staden* '…he stood up, and went into the city' (in the Bible); and (ii) *<Tre dagar och tre nätter> satt Lilia…* 'Three days and three nights sat Lilia…' (Spf-file lb99904107) and …*vara i <tre dagar och i tre nätter> i jordene* '…be three days and three nights in the earth' (in the Bible). The results from PAIR suggest that thorough (manual) evaluation is an important step since no scoring mechanism is built into the software that could rank the results according to some strength association measure.

Conclusions and Future Work

We have outlined a text reuse experiment by comparing the Charles XII Bible with the content of Spf. The results were evaluated by hand, and by adjusting some of the implemented parameters, until the results were "pleasing" which runs the risk of optimizing precision at the cost of recall. More advanced methodologies might be necessary to cope with language phenomena beyond local alignment techniques [9], taking into consideration more linguistically informed preprocessing (lemmatization, spelling variation, synonym identification, document representation and robustness to noisy texts). Software, such as TRACER [1] and its visualization component TRAViz (Text Re-use Alignment Visualization) [6] can be applied as a distant reading method in order to obtain an overview of the distribution of text reuses between each pair of texts and all texts in a corpus. Moreover, for future work we would also like to compare the content of Spf and also other Swedish digital collections with each other and also with classical, influential Swedish authors such as August Strindberg and Selma Lagerlöf.

References

[1] Büchler M., Burns P. R., Müller M., Franzini E. and Franzini G. (2014). Towards a Historical Text Re-use Detection. In *Text Mining, Theory and Applications of NLP*. Biemann, C. and Mehler, A. (eds). Pp. 221-238. Springer.

[2] Colavizza G., Infelise M. and Kaplan F. (2015). Mapping the Early Modern News Flow: An Enquiry by Robust Text Reuse Detection. In Social Informatics. *Lecture Notes in Computer Science.* Vol. 8852. pp 244-253. Springer.

[3] Cordell R. (2013). "Taken Possession of": The Reprinting and Reauthorship of Hawthorne's "Celestial Railroad" in the Antebellum Religious Press. *Digital Humanities Quarterly 7:1*. Alliance of Digital Humanities Organizations (ADHO).

[4] Ganascia J-G., Glaudes P. and del Lungo A. (2014). Automatic Detection of Reuses and Citations in Literary Texts. *Lit Linguist Computing* 29 (3): 412-421. doi: 10.1093/llc/fqu020.

[5] Horton R., Olsen M and Roe G. (2010). Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections. *Digital Studies / Le champ numérique*. Vol 2:1.

[6] Jänicke S., Geßner A., Franzini G., Terras M., Mahony S. and Scheuermann G. (2015). TRAViz: A Visualization for Variant Graphs. *Digital Scholarship in the Humanities – Digital Humanities 2014 Special Issue*.

[7] Roe G.H. (2012). Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research. *Digital Hum.* 2012. Hamburg, Germany.

[8] Seo J and Croft W.B. (2008). Local Text Reuse Detection. *ACM SIGIR*, Singapore.

[9] Smith D.A., Cordell R. and Maddock Dillon E. (2013). Infectious Texts: Modelling Text Reuse in Nineteenth-Century Newspapers. *IEEE Conf on Big Data*. Santa Clara, CA, USA.

[10] Smith D.A., Cordell R., Maddock Dillon E., Stramp N. and Wilkerson J. (2014). Detecting and Modelling Local text Reuse. *14th ACM/IEEE-CS Joint Conference on Digital Libraries Pages (JCDL)*. Pp. 183-192. ACM.

[11] text-pair. (2015). *PAIR: Pairwise Alignment for Intertextual Relations*. Online; accessed 31-August 2015: <https://code.google.com/p/text-pair/>

[12] Wikipedia. (2015). *Intertextuality*. Online; accessed 25-August-2015: <https://en.wikipedia.org/wiki/Intertextuality>[14] Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences* 13:555-556.