Intertextuality in Tibetan Texts

Daniel Labenski[1], Elad Shaked[1], Orna Almogi[2,3], Lena Dankin[1], Nachum Dershowitz[1,4] and Lior Wolf[1]

[1] School of Computer Science, Tel Aviv University, Ramat Aviv, Israel

[2] Department of Indian and Tibetan Studies, Universität Hamburg, Hamburg, Germany

[3] Khyentse Center for Tibetan Buddhist Textual Scholarship, Universität Hamburg,Hamburg, Germany

[4] Institut d'Études Avancées de Paris, Paris, France

Extended Abstract

Literary scholars are often interested in finding textual citations within a multiple-source corpus, usually the result of one source quoting or paraphrasing another one. There is a spectrum of similarity of passages, ranging from exact quotations to bags of synsets, from *ipsissima verba* (same words) to *ipsissima vox* (same meaning). The detection of texts sharing meaning naturally requires extensive semantic knowledge, but even the task of finding all exact quotes is quite complicated given a large collection of documents, such as the Tibetan-Buddhist canon we are working with, and cannot be achieved manually.

We have been developing and comparing tools for finding inexact quotations that aggregate as parallel passages between Tibetan texts and are the result of either citations (with or without attribution) or borrowing (with no indication whatsoever). In this work, our main interest is in extracting parallel passages that are meaningful for the scholar (who may be preparing a critical edition of a particular text, for example). We also address the run-time issue, since it can be very significant when the corpus is large.

The Tibetan language belongs to the Tibeto-Burman branch of the Sino-Tibetan family. It is monosyllabic (morphemes are single syllables) for which computerized language tools are largely lacking. It has a plethora of (usually) monosyllabic grammatical particles, which are often omitted. Occasionally, the same syllable can be written using one of several orthographic variations, for example, *sogs* and *stsogs*. In the case of verbs, the syllable has various inflectional forms that are often homophones, a fact that can result in variants in the reading due to scribal errors or lack of standardization.

The Tibetan-Buddhist canon consists of two parts: the *Kangyur* (108 volumes containing what is believed by tradition to be the Word of the Buddha), texts that were mostly translated directly from the Sanskrit original (with some from other languages and others indirectly via Chinese); and the *Tengyur* (about 210 volumes) consisting of canonical commentaries, treatises, and various kinds of manuals that were likewise mostly translated from Sanskrit (with some from other languages and a few originally written in Tibetan).

By identifying cited or borrowed passages within the corpora of Buddhist Indo-Tibetan (i.e. translated) and Tibetan (i.e. autochthonous) literature, other research questions can also be better addressed, including the following: determining the history of composition of individual texts; determining the relative chronology of groups of texts; determining the intellectual scholarly milieu in which the texts emerged; and determining the intellectual history behind the texts (viz. terminology and concepts). After identifying parallel passages, one can assess the frequencies of letter/syllable/word replacements in the aligned passages of selected texts or text groups. This can serve to help answer research questions like determining editorial policies and processes, such as standardization of orthography, standardization of employment of grammatical particles (i.e. according to the so-called *sandhi* rules), and identifying processes of "revisions" of translated texts.

We began by adapting the APBT method of [Barsky et al.], designed for matching DNA subsequences, to our problem [Klein et al.]. This algorithm looks for "all against all approximate matches" at the character level (within some given threshold of difference between passages) by rephrasing the problem as finding maximal paths in a matching graph. Some post-processing is required to merge adjacent parallel passages and to trim the edges of the passages to the meaningful part.

That method was modified to work with syllables as the basic building block, rather than the individual character. This change improved both runtime and the quality of results. Since on the average a syllable has 4 or 5 characters, the speedup was some twenty-fold. The quality of the results was also better because, with character-wise alignment, syllables can share many letters but have no semantic similarity. This also removes errors such as one long syllable that aligns with more than one syllable. (In the character-wise alignment a space is a regular character and counts equally.)

Next, we developed a stemmer for Tibetan [Dankin et al.] and redesigned the algorithm accordingly. Using stems enables one to detect also inexact matches, which are also common in the canon, and precision improves thereby.

A followup task is to rank the parallel passages. There are two main parameters that are relevant: the length of the shared text and an estimate of significance. This step is crucial since the APBT algorithm outputs thousands of matches, and we aim to allow scholars to work only on the most interesting and least trivial matches, moving the more "noisy" results to the bottom of the list. In our current work, we are incorporating a new ranking mechanism that counts, not only match length, but also the importance of the content of the match through a "smarter" local alignment that scores matches and mismatches of different syllables/stems differentially.

Our results are being compared to a simple baseline that uses a LCS algorithm to find overlapping candidates, and also to the eTracer algorithm [Büchler et al.], designed to detect parallel text passages in German language novels of the 16th to 19th centuries and based on a bag-of-word representation (ignoring word order).

Experiments on the entire canon were run in parallel using Spark technology, allowing us to execute all runs in a reasonable time.

References

B. E. Klein, N. Dershowitz, L. Wolf, O. Almogi, D. Wangchuk. Finding Inexact Quotations Within a Tibetan Buddhist Corpus. In: *Proceedings of Digital Humanities (DH),* 2014.

M. Barsky, U. Stege., A. Thomo, C. Upton. A Graph Approach to the Threshold All-Against-All Substring Matching Problem. *ACM Journal of Experimental Algorithmics* 12, 2008.

O. Almogi, L. Dankin, N. Dershowitz, Y. Hoffman, D. Wangchuk, L. Wolf. Stemming and Segmentation for Classical Tibetan. In: *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing),* 2016.

M. Büchl, G. Franzini, E. Franzini, M. Moritz. Scaling Historical Text Re-use. In: *Proceedings of the IEEE International Conference on Big Data,* 2014.

D. S. Hirschberg. Algorithms for the Longest Common Subsequence Problem. *J. ACM* 24(4), 1977.