SansTib

A Sanskrit - Tibetan Parallel Corpus and Bilingual Sentence Embedding Model

Sebastian Nehrdich

Institute for Language and Information,Heinrich-Heine-Universität Düsseldorf;
Khyentse Center for Tibetan Buddhist Textual Scholarship, Universität Hamburg

nehrdich@uni-duesseldorf.de

Abstract

This paper presents the development of SansTib, a Sanskrit - Classical Tibetan parallel corpus automatically aligned on sentence-level, and a bilingual sentence embedding model. The corpus has a size of about 317,289 sentence pairs and 14,420,771 tokens and thereby is a considerable improvement over previous resources for these two languages. The data is incorporated into the BuddhaNexus database to make it accessible to a larger audience. It also presents a gold evaluation dataset and assesses the quality of the automatic alignment.

Introduction

- Translations of Indian Buddhist texts composed in Sanskrit or Middle Indic languages have been of central importance to the formation of the Tibetan Buddhist literary tradition.
- These translations are not only relevant for the Tibetan Buddhist tradition, but also for the Indian, since a lot of Indian Buddhist texts have been lost and are nowadays only available in translation.
- Both Sanskrit and Classical Tibetan are low-resource languages for which comparatively large monolingual corpora became available in recent years.
- With the help of vecalign [5] and multilingual sentence embedding [4] a sufficiently precise automatic alignment of these digitally available texts is now possible

Main Contributions

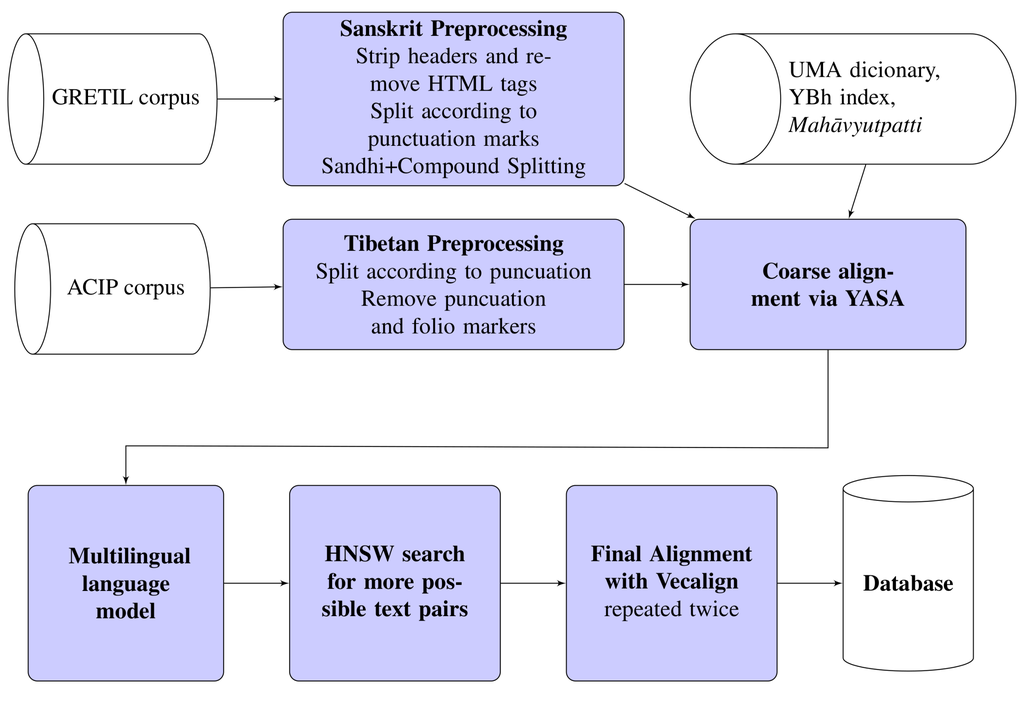
- A sentence-level aligned parallel corpus of Sanskrit Buddhist texts and their Tibetan translations with a total number of 317,289 sentence pairs which is much larger and covers a greater variety of domains than the already available bilingual resources for these two languages.
- Three manually aligned datasets with a combined size of 6,916 sentence pairs spanning different genres of Buddhist literature for the evaluation of Sanskrit Tibetan sentence alignment quality.
- A bilingual sentence embedding model that can be used for information retrieval and sentence alignment.

I make the dataset available at: `https://github.com/sebastian-nehrdich/sanstib`
The dataset is also accessible for philological research via a user-friendly interface at `https://buddhanexus.net/multi/neutral`.
The bilingual sentence embedding model is available via huggingface: `https://huggingface.co/buddhist-nlp/sanstib`

Challenges of Sanskrit-Tibetan Alignment

- Sanskrit is a classical language of the Indo-Aryan branch of the Indo-European languages while Classical Tibetan is an ergative-absolutive language belonging to the Sino-Tibetan language family.
- Sanskrit relies heavily on morphology, has a complex verbal system, rich nominal declension and employs long nominal compounds frequently; Tibetan on the other hand does not mark nouns regarding gender or number and uses particles at the end of whole noun phrases to indicate case.
- Grammar and vocabulary of Sanskrit and Tibetan are totally different and there are only few loanwords and transliterations of Sanskrit terms in Tibetan translations.
- Loss of sentences, paragraphs or whole chapters occurs, depending on the texts, on both sides due to the difficult transmission of these texts in South Asia
- Certain Indian texts have been altered and developed further after their translation into Tibetan
- The order of sub-clauses and whole sentences, especially in the case of verses, is at times inverted in Tibetan translation:
  Sanskrit:
  rājāno rājaputrāś ca amātyāḥ śreṣṭhinas tathā |
  piṇḍārthe nopadeśeta yogī yogaparāyaṇaḥ ||
  Tibetan:
  rnal 'byor gzhol ba'i rnal 'byor bas ||
  rgyal po dang ni rgyal po'i bu ||
  de bzhin blon po tshong dpon la ||
  zas kyi phyir ni bsten mi bya ||
- Punctuation conventions between Tibetan and Sanskrit are different, and within digitally available Sanskrit editions they are not consistent

Data and Processing Pipeline



- For the Sanskrit data, I use the Buddhist etexts available in the GRETIL collection[1], which number 405, have a combined size of about 60MB and 5m tokens.
- The Tibetan data consists of the Kangyur and Tengyur collections by the Asian Classics Input Project (ACIP).[2] with a combined size of 4284 files, 361MB and 85m tokens

Results

This table shows the statistics of the resulting dataset:

Collection	Category	Text pairs	Aligned sentence pairs
Scriptures (Kangyur)	Vinaya	4	9513
	Prajñāpāramitā	21	62775
	Ratnakūta	3	3160
	Avataṁsaka	1	6053
	Sūtra	26	55453
	Tantra	19	22651
	Total	74	159605
Treatises (Tengyur)	Tantra	6	481
	Prajñāpāramitā	4	10967
	Madhyamaka	26	32684
	Sūtra Commentaries	3	1529
	Yogācāra	18	38843
	Abhidharma	4	37440
	Jātaka	2	12295
	Lekha	5	1037
	Pramāṅa	19	22408
	Total	87	157684
Total		**161**	**317289**

Evaluation of the Sentence Alignment Algorithms

I use the three manually aligned datasets AKBh (2,000 sentence pairs), VkN (2,770 sentence pairs) and PSkVBh (2,146 sentence pairs) for the evaluation of the alignment quality. These datasets cover both domains of scriptures and treatises.
I report the F-measures F_A and F_S which is a combination of precision and recall ratios on alignment and sentence level respectively. F_A considers only those alignments to be positive that are identical with those in the gold data, therefore ignoring partly correctly aligned sentences. F_S on the other hand considers how many actual sentences have been aligned correctly, regardless of the size of the bisegment. A high F_S score can also be achieved when a large number of sentences is aligned in only a few bisegments. It is therefore important to consider both F_A and F_S scores in order to be able to judge the alignment quality correctly.

	hunalign		YASA		vecalign	
Dataset	F_A	F_S	F_A	F_S	F_A	F_S
AKBh	44.6	75.4	56.0	78.1	**82.8**	**94.3**
VkN	30.1	63.5	49.0	73.0	**75.1**	**90.6**
PSkVBh	63.6	83.1	66.1	80.4	**92.6**	**97.3**

Conclusions

- The evaluation on the three different datasets shows that the combination of vecalign and multilingual sentence embedding clearly outperforms length-based and dictionary based algorithms.
- While I cannot assume that an F_S score of more than 90% is met for all texts in the presented dataset, the evaluation results lead me to assume that the majority of the aligned texts have a reasonable good quality.
- The presented data can be used as a resource on its own for philological research and has already been incorporated into the BuddhaNexus database.
- Since the quality of the alignment of the presented data depends strongly on the punctuation conventions followed in both languages, I believe that standardized methods for recognizing sentence boundaries for Sanskrit could help to improve the performance of the alignment algorithms.

[1] `http://gretil.sub.uni-goettingen.de/gretil.html`
[2] `https://asianclassics.org/library/downloads/`

Selected References

[1] Oliver Hellwig and Sebastian Nehrdich. Sanskrit Word Segmentation Using Character-level Recurrent and Convolutional Neural Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[2] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

[3] Fethi Lamraoui and Philippe Langlais. Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment? In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France, September 2-6 2013.

[4] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.

[5] Brian Thompson and Philipp Koehn. Vecalign: Improved Sentence Alignment in Linear Time and Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China, November 2019. Association for Computational Linguistics.