Phrase-based alignment of classical Chinese and English

Donald Sturgeon and John S. Y. Lee*

Introduction

Aligned parallel corpora are useful for a variety of purposes including machine translation and statistical studies, as well as making possible new and innovative digital tools for use in pedagogy and research. Alignments can be made at various levels of granularity, a common type being alignment of sentences. In the case of classical Chinese in particular, databases containing such alignments are also of direct utility to scholars and linguists due to the complex semantics of individual terms of the language, the limited size of the extant body of writing, and a lack of sufficiently comprehensive bilingual dictionaries.[1] Aligned corpora make possible automated extraction of relevant linguistic data for arbitrary terms, while avoiding the prohibitively high cost involved in manual construction of an adequate bilingual dictionary. [2]

While in many modern languages sentences are delimited in the written form by the presence of certain punctuation marks, classical Chinese was for many centuries written without any punctuation marks whatsoever, and later with punctuation that delimited only boundaries between phrases.[3] Modern editions of classical Chinese texts include punctuation marks corresponding closely to (and greatly influenced by) modern English punctuation, but often disagree on the precise details of such punctuation, highlighting the degree of freedom present in adding such marks. Due to the grammar of classical Chinese, this freedom often extends to choices determining apparent sentence boundaries. Similarly and partly as a result of this, English translations of these texts often differ in the precise delimiting of sentences in the source text.

As a result of these linguistic and historical factors, sentence-based alignment of classical Chinese texts and their modern translations is problematic, as sentences of the source and target languages often fail to correspond exactly due to different choices made in punctuating the text, even where these do not correspond to significant differences in interpretation. By contrast however due to the much lower degree of freedom involved, different modern editions of early texts exhibit much less disagreement regarding the delimiting of phrases.[4]

---

* The work described in this paper was supported by a Teaching Development Grant from City University of Hong Kong (Project No. 6000489).

[1] Two publicly available databases incorporating such alignments between classical Chinese and English are the Chinese Text Project (http://ctext.org), and Thesaurus Linguae Sericae (http://tls.uni-hd.de).

[2] The manual alignments used for comparison in this study are used for these purposes on the Chinese Text Project website – see http://ctext.org/introduction#dictionary

[3] Additionally, Chinese was, and still is, written without spaces or other delimiters indicating word boundaries.

[4] Also of note is that in our chosen corpus while there were 25%-34% more English sentences than Chinese sentences (depending upon whether or not

Motivated by these factors, this study investigates automated phrase-wise alignment of a corpus of classical Chinese texts and their English translations, comparing unsupervised machine-generated phrase-wise alignments versus sentence-wise alignments by means of human annotated results.

Data

The corpus used in this study consists of four texts from the Chinese Text Project database from the pre-Qin and Han period.[5] This corpus contains classical Chinese texts punctuated according to modern punctuation standards, and English translations that have been manually aligned both at the paragraph level and at phrase level.[6] We begin by splitting each paragraph of text in the corpus into phrases based upon the (modern) Chinese punctuation of the text, and similarly for the English translation, splitting into likely phrases based upon English punctuation. For instance, the Chinese sentence:

庖有肥肉，廐有肥馬，民有飢色，野有餓莩，此率獸而食人也。

is split by punctuation into the five phrases:

庖有肥肉，
廐有肥馬，
民有飢色，
野有餓莩，
此率獸而食人也。

and the corresponding English (in this case written as three sentences),

> In your kitchen there is fat meat; in your stables there are fat horses. But your people have the look of hunger, and on the wilds there are those who have died of famine. This is leading on beasts to devour men.

is similarly split by its punctuation into the five phrases:

In your kitchen there is fat meat;
in your stables there are fat horses.
But your people have the look of hunger,
and on the wilds there are those who have died of famine.
This is leading on beasts to devour men.[7]

---

semicolons are counted as sentence delimiters), there were only 14% more English phrases than Chinese phrases.

[5] http://ctext.org. The texts used in this study are the *Analects*, *Mengzi*, *Yangzi Fayan*, and *Zhuangzi*.

[6] In some cases phrases (as defined by punctuation boundaries in the Chinese text and their English equivalent) of the source and target text do not match in sequence (e.g. if the translation reorders the phrases, or a construction extends over several phrases). In these cases, the shortest sequential matching segments are aligned.

[7] In this example, the five phrases match one another exactly. There is considerable freedom in punctuating both the Chinese and the English translation in this case, and both could easily be rewritten to consist of varying numbers of sentences. However, it would be difficult in this example to correctly

Methodology

To perform the phrase-wise alignment, we use the "Bilingual Sentence Aligner" tool published by Microsoft Research,[8] which implements the algorithm described in Moore (2002). This method uses a combination of sentence length correspondence and extracted word correspondence to align source and target sentences fully automatically. Although this algorithm and tool were both designed for sentence alignment, we apply them instead to the phrases extracted from our corpus. The phrases given as input to the tool are grouped by the manually aligned paragraphs specified in the corpus, which for the texts selected contain an average of 5 sentences and 13 phrases.[9]

The tool produces as output those pairs of phrases in the source and target languages that, according to its model, match 1:1 with a probability greater than a specified threshold value. We compare these results with the expected 1:1 matches according to the manually annotated Chinese Text Project data. Finally, we perform the same analysis again on the same corpus of texts, this time splitting the corpus into sentences rather than phrases, and compare the accuracy of these sentence-wise alignments with the phrase-wise ones.

Results

The selected corpus contained 12743 manually annotated 1:1 phrase alignments.[10] For varying probability thresholds, the automated method correctly identified up to 52% of these (recall), and the pairs it identified were accurate in 82-89% of cases (precision). Performing the evaluation using the same four texts, but including additional texts in the automated word association model-building phase, further improved the recall to up to 58% at some cost to precision.[11]

Repeating the assessment using sentences instead of phrases, the corpus contained a total of 6993 manually annotated 1:1 sentence alignments. For varying probability thresholds, the automated method correctly identified up to

---

punctuate the Chinese in any way that would not result in precisely these five phrases once split using punctuation as described here.

[8] http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

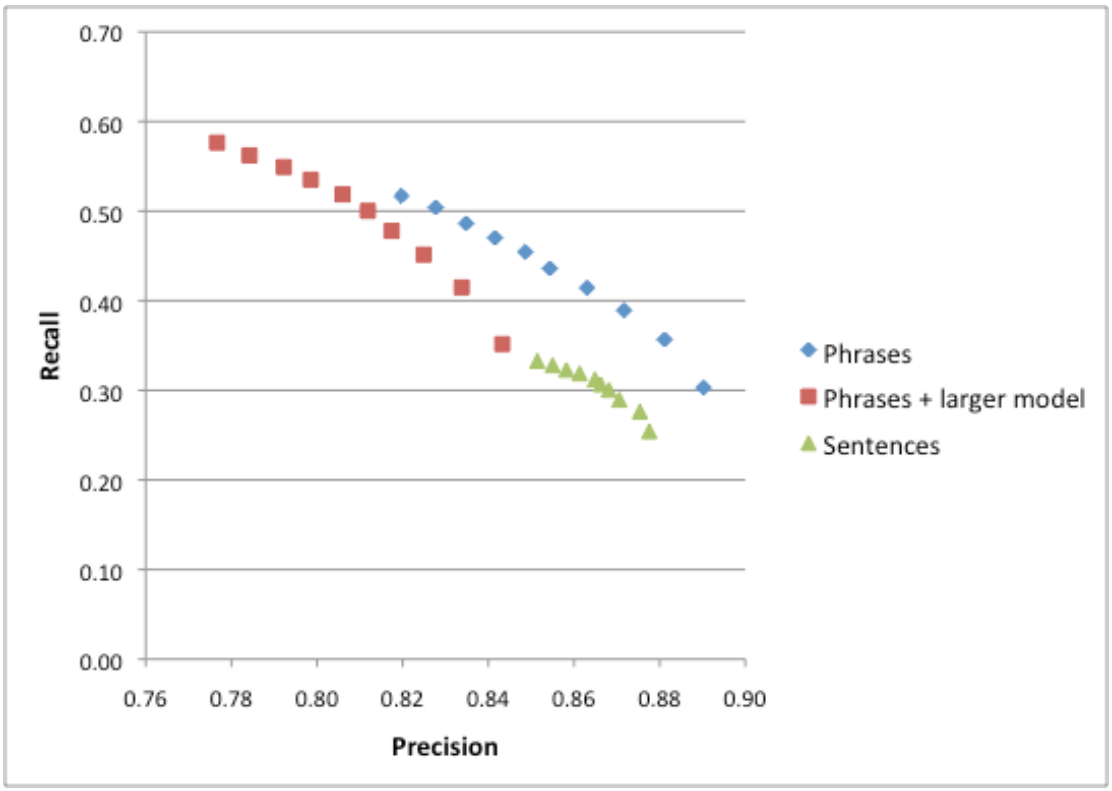
[9] The Microsoft tool has two modes of operation. In the first, no sentence groupings are specified – for our corpus, this mode of operation did not produce satisfactory results. In the second, sentences are specified in groups, such that sentences of any particular source group (in our case, the phrases of one particular paragraph) should match only a sentence in the corresponding target group.

[10] Not all alignments in the corpus are 1:1 – in the chosen corpus, 44% of total characters occur within 1:1 matches.

[11] Additional texts: *Mozi, Art of War, Liji, Shang Jun Shu, Shang Shu, Book of Poetry, Book of Changes.*

33% of these (recall), and the pairs it identified were accurate in 85-87% of cases (precision).

Phrase-wise alignment produced a greater volume of data corresponding to significantly more of the corpus (up to 69% more by character count), while simultaneously achieving significantly higher accuracy than sentence-wise alignment.



Conclusion

The results obtained show that phrase-based alignment of classical Chinese is feasible using algorithms and tools developed for sentence-based alignment. A high degree of precision was shown to be possible using a relatively small corpus and an algorithm that made few assumptions about the languages involved. Furthermore, it was demonstrated that for the chosen corpus, phrase-wise alignment achieved higher recall and precision than sentence-wise alignment with the same algorithm.

Works cited

Moore, R.C. Fast and Accurate Sentence Alignment of Bilingual Corpora. In: Richardson. S.D. (ed.) ATMA 2002. LNCS (LNAI), vol 2499, pp. 135-144. Springer, Heidelberg (2002)