TUFS Asian Language Parallel Corpus (TALPCo)

Hiroki Nomoto† Kenji Okano† David Moeljadi‡ Hideo Sawada†
†Tokyo University of Foreign Studies ‡Nanyang Technological University, Singapore
{nomoto, okanok}@tufs.ac.jp, davidmoeljadi@gmail.com, sawadah@aa.tufs.ac.jp

Abstract

The TUFS Asian Language Parallel Corpus (TALPCo) is an open parallel corpus consisting of Japanese sentences and their translations into Burmese, Malay, Indonesian and English. This paper describes how we built it and its notable features, especially those pertaining to the choice of Japanese as the source language of translation.

1 Introduction

This paper reports the development of the TUFS Asian Language Parallel Corpus, or TALPCo, which will be available at https://github.com/matbahasa/TALPCo licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. TALPCo is a parallel corpus consisting of five languages: Japanese, Burmese (Myanmar), Malay, Indonesian and English. To the best of our knowledge, TALPCo is the second openly available parallel corpus for multiple Asian languages, with the first one being the Asian Language Treebank (ALT) Parallel Corpus (Riza et al. 2016).[1,2] As we shall describe below, TALPCo supplements ALT in several respects, thus diversifying the types of resources available for Asian language NLP tasks.

This paper is organized as follows. Section 2 gives a general overview of TALPCo and describes how the sentences in TALPCo were obtained. Section 3 provides language-specific information about the annotation of those sentences. At the time of writing this paper (December 2017), only the Burmese data have been annotated for tokens and parts of speech (POSs), though the POS tagsets is still tentative. Section 4 compares TALPCo with the ALT Parallel Corpus mentioned above and points out some merits of building a Japanese-based parallel corpus like TALPCo. Lastly, section 5 discusses our plan for improving TALPCo in the future.

[1] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

[2] Besides these two corpora, the NTU-Multilingual Corpus (Tan and Bond 2012; http://compling.hss.ntu.edu.sg/ntumc/) offers its search system to the public for Mandarin Chinese, Japanese, Indonesian and English.

2 The data

The data in TALPCo originally come from the basic vocabulary words and example sentences in 24 languages in the TUFS Open Language Resources (Kawaguchi 2007). Four languages in this database are already presented in a parallel format. These languages are Japanese, Burmese, Malay and English. The first and second authors of the present paper were involved in the creation of the Malay, English (Nomoto) and Burmese (Okano) components in the TUFS Open Language Resources database. These components are translations of the Japanese component.

The Japanese component contains 799 basic vocabulary words and example sentences for them. These basic vocabulary words are selected in accordance with the lowest level of the Japanese Language Proficiency Test, i.e. N5 (equivalent to Level 4 in the old system). The example sentences are given in a formal conversational register. An example is given in (1).[3] The number in brackets after the free translation indicates the sentence's ID in the corpus, which is inherited from the TUFS Open Language Resources data.

(1) Japanese (formal)

帰る	電車が	なかったので、
Kaeru	densha-ga	nakat-ta-node
return	train-NOM	not.exist-PST-because
友達の	家に	泊まりました。
tomodati-no	ie-ni	tomari-masi-ta.
friend-GEN	house-at	stay-POL-PST

'I stayed overnight at my friend's house because there was no train to go home.' [3156]

Compare this example with its informal equivalent in (2). The formal and informal styles differ at the level of basic grammar as well as lexical choices. In the formal style, overt case particles such as *ga* (nominative) and *ni* (dative) are used.

[3] Non-standard abbreviations not available in the Leipzig Glossing Rules; ACT: active; POL: polite; VS: verb-sentence marker.

(2) Japanese (informal)

帰る　電車　なかったから、
Kaeru　densha　nakat-ta-kara
return　train　not.exist-PST-'cause
友達ん家　泊まった。
tomodati-n-ti　tomat-ta.
friend-GEN-place　stay-PST
'I stayed overnight at my friend's place 'cause there was no train to go home.'

The formality of the original Japanese sentences is reflected in the translations as much as possible. For example, the Malay counterpart of (1) is (3), which uses formal expressions such as *oleh kerana* 'because' and *tiada* 'not exist' instead of their informal equivalents, i.e. *sebab* 'because' and *takde* 'not exist'.

(3) Malay
Oleh kerana tiada kereta api, kami
by because not.exist train we
bermalam di rumah kawan.
spend.night at house friend
'I stayed overnight at my friend's house because there was no train to go home.' [3156]

The Burmese and Malay translations were performed by a native speaker of eac respective language. The translations were then checked by Japanese native speakers who know these languages well, including the first and second authors of the present paper. The English translation was prepared by a Japanese undergraduate student who had studied at an international junior high school and then checked by a native British English speaker. British English was chosen over American English because most English-speaking countries in Asia consider the former as the norm, though the influence of the latter has been increasing in recent years.

In building TALPCo, we modified the original data from the TUFS Open Language Resources. Translation errors were corrected, and duplicates, i.e. identical sentences with different IDs, were removed. Where multiple expressions were presented, we selected one that was thought to be the most common. In addition to the four languages that were already parallel in the TUFS Open Language Resource data, we added Indonesian. The third author of the present paper translated the Japanese sentences into Indonesian. The translation was then checked by the first author.

The corpus thus built contains a total of 1,372 distinct sentences. An example of parallel sentences in Japanese, Burmese, Malay, Indonesian and English is given in (4).

(4) [J] 学校は 休みです。
[B] ကျောင်းပိတ်တယ်။
[M] Sekolah cuti.
[I] Sekolah sedang libur.
[E] There is no school. [1180]

3 Language-specific information

3.1 Japanese

The Japanese sentences in the TUFS Open Language Resources data are already tokenized. The unit of tokenization is *bunsetsu* (文節). A *bunsetsu* consists of one free morpheme and bound morphemes attaching to it, if any. For example, in (1) above, 電車が (*densha-ga*) forms a *bunsetsu*. 電車 (*densha*) 'train' is a free morpheme, whereas the nominative case particle が (*ga*) is a bound morpheme, specifically an enclitic. The only modification that we made in TALPCo was to change the boundary character from a full-width, 2-byte space to a half-width, 1-byte one.

3.2 Burmese

Burmese texts use white spaces. However, the units separated by them are not always words. We thus created a dataset in which sentences have been tokenized manually. (5) shows an example of our tokenization. Note that the second and third units separated by white spaces contain two content words.

(5) 'What's the date today?' [1335]

ဒီနေ့	ဘာလ	ဘာရက်လဲ။

↓ *tokenization*

ဒီနေ့	ဘာ	လ	ဘာ	ရက်	လဲ	။
today	what	month	what	date	Q	.

Thus far, no conclusive Burmese tokenization rules exists, except for the guidelines provided by the ALT Parallel Corpus.[4] The guidelines are based on their analysis of POSs in Burmese. However, their rules still leave some room for individual speakers' personal judgements on what is and is not a token. This may lead to indeterminable strings and inter-annotator variations. The lack of a well-organized system for spacing in Burmese orthography constitutes a major challenge for automatic processing of Burmese texts.

3.3 Malay/Indonesian

We have not tokenized Malay/Indonesian sentences.[5] However, tokenization is a relatively easy task. As seen

[4] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/Myanmar-annotation-guideline.pdf

[5] Malay (ISO693-3 zsm) and Indonesian (ISO693-3 ind) are two regional dialects of "Malay" in the broad sense (ISO693-3 msa). Malay is the national language of Malaysia, Singapore and Brunei, whereas Indonesian is the national language of Indonesia. We treat the two collectively as Malay/Indonesian here, as most linguistic features are common to both.

in examples (3) and (4) above, words are separated by a white space in Malay/Indonesian. Hence, it is possible to tokenize Malay/Indonesian sentences by using a tokenizer for English with necessary modifications. See section 5 for details.

4 Related work

As noted in section 1, the first openly available parallel corpora involving multiple Asian languages is the Asian Language Treebank (ALT) Parallel Corpus (Riza et al. 2016). The ALT Parallel Corpus consists of data in English and the following nine Asian languages: Tagalog (Filipino), Indonesian, Japanese, Khmer (Cambodian), Lao (Laotian), Malay, Burmese (Myanmar), Thai and Vietnamese. It is similar to TALPCo in mainly targeting national languages of Southeast Asian countries as well as Japanese and English. Currently, the ALT Parallel Corpus contains more languages than TALPCo, namely Tagalog, Khmer, Lao, Thai and Vietnamese. In the future, however, we would like to include these languages in TALPCo too.

Besides the number of languages covered, there are three big differences between the ALT Parallel Corpus and TALPCo concerning their data. First, the two corpora differ in genre. The ALT Parallel Corpus consists of *Wikinews* articles. It is thus a written language corpus, in particular one of the journalistic style. By contrast, the sentences in TALPCo are not journalistic but are concerned with everyday life. They are short and structurally simple. One can use them in formal conversation. TALPCo can thus be considered a (quasi-)spoken language corpus. Therefore, the two corpora complement each other.

The second difference between the two corpora is the language of the original data. The ALT Parallel Corpus is English-based; it was built by translating English *Wikinews* articles into Asian languages. TALPCo, on the other hand, is Japanese-based.

Translations are usually affected by the structure and lexical choice of the source language. It is thus expected that different translations, and hence different parallel corpora, are obtained with different source languages. For example, Japanese but not English allows the so-called *pro* drop, where arguments of a predicate are not expressed overtly, as shown in (6). Notice that none of the arguments of the predicates 'cheap' and 'to buy' are stated explicitly in Japanese and that the English translation is supplied with overt expressions for them, which are indicated by **boldface**. It is grammatical to express them overtly in Japanese, but more natural not to do so.

(6) 安ければ、 買います。
Yasu-kereba, kai-masu.
cheap-if buy-POL
'If **they** are cheap, **I** will buy **them**.' [2047]

The translations of this sentence into the other Asian languages in (7) also involve *pro* drop, reflecting the original Japanese sentence. Again, those elements which do not occur explicitly in the original Japanese sentence are indicated by **boldface**. The Burmese sentence is most similar to the Japanese sentence in that neither the subject nor the object of the main clause is expressed overtly. Malay realizes the main clause subject overtly, but the main clause object is left implicit. The translations would have been different if they had been based on the English sentence in (6).

(7) [B] ဈေး-သက်သာ-ရင် ဝယ်-မယ် ။
jhe:-sak'saa-rang' way'-may'‖[7]
price-cheap-if buy-VS.IRR

[M] ***Saya*** *akan beli jika* ***harga-nya*** *murah.*
I will buy if price-its cheap

[I] ***Saya*** *akan mem-beli-**nya*** *kalau murah.*
I will ACT-buy-it if cheap

The main clause object is not expressed overtly in all the Asian languages in TALPCo because its referent is the topic of the sentence; that is, the sentence is about the relevant entity. In this connection, the topic-comment construction, where a topic noun phrase occurs outside the subject-predicate structure (NP_{Topic} [Subj Pred]), is another linguistic structure that can be obtained more readily in a Japanese-based parallel corpus than an English-based one. In the examples in (8) below, all the Asian languages employ a special topic-comment construction, whereas English does not.

(8) [J] あの 犬は 耳が 大きいです。
ano inu-wa mimi-ga ookii-desu.
that dog-TOP ear-NOM big-POL

[B] ဟို-ခွေး-က နားရွက် ကြီး-တယ် ။
hui-khwe:-ka naa:rwak' krii:-tay'‖
that-dog-NOM earlobe big-VS.RLS

[MI] *Anjing itu telinga-nya besar.*
dog that ear-its big

[E] 'That dog has big ears.' [3225]

A Japanese-based parallel corpus of Asian languages like TALPCo has another advantage over an English-based one, i.e. it can avoid the problem of potential overspecification of linguistic features. For example, Asian languages with numeral classifiers have a number-marking system that is more complex than that of English, which lacks numeral classifiers (Nomoto 2013). Plural number marking on nouns is more restricted in the former type of language, as it competes with the number-neutral, general form, which is usually morphologically

[7] The transliteration of Burmese scripts in this paper follows the Sawada system (`http://www.aa.tufs.ac.jp/~sawadah/burroman.pdf`).

bare. No such competition exists in English. *Dogs* in English can be either *anjing* (general) or *anjing-anjing* (plural) in Malay/Indonesian, for example. Consideration of a formal parallelism between the source and target languages may lead the translator to the second, more marked option, resulting in overspecification. This problem is unlikely to happen if the source langauge is Japanese, as Japanese has the same number-marking system as Malay/Indonesian: *inu* (general) and *inu-tati* (plural).

The last point in which TALPCo differs from the ALT Parallel Corpus is that the translations have been checked by linguists whose native language is Japanese and who know the grammar and lexicon of the target languages well. Hence, the data in TALPCo are thought to contain fewer errors compared to those in the ALT Parallel Corpus. For example, TALPCo does not commit the common error in Malay/Indonesian of confusing between the passive voice prefix *di-* (spelt as part of a word) and the preposition *di* (spelt separately from the following word). This error is extremely frequent in the Malay subcorpus of the ALT Parallel Corpus, though it is not as frequent in its Indonesian subcorpus.

5 Conclusion and future work

Small though it may be, TALPCo, we believe, will become a useful and reliable resource for NLP tasks involving low-resource Asian languages. Moreover, it can also be used in linguistic research and language education.

Currently, only the Burmese subcorpus has been tokenized and POS-tagged. In the future, we will also tokenize and assign POS tags to the other subcorpora. This can be done relatively easily for Japanese and English, given the many existing tools and resources, such as MeCab (Kudo et al. 2004) and the Natural Language Toolkit (NLTK) (Bird et al. 2009). Existing tokenizers for Malay/Indonesian are generally quite simple, just adding Malay/Indonesian-specific abbreviations to an English tokenizer (e.g. Sastrawi Tokenizer[8]). More sophisticated tokenization can be done by means of POS Tag (Rashel et al. 2014), which is able to identify multi-word expressions in Indonesian. POS Tag's tokenization process also utilizes the morphological analysis provided by MorphInd (Larasati et al. 2011), which lemmatizes a word and assigns two POS tags (called "lemma tags" and "morphological tags") to each lemma. Furthermore, the first and third authors of the present paper are currently building a large morphology dictionary for Malay/Indonesian (Nomoto et al., under review). This dictionary will enable linguistically accurate morphology annotation.

Besides adding annotations to the existing subcorpora, we would also like to expand our corpus in terms of the number of languages. Our goal is to include all the Asian languages available in the ALT Parallel Corpus.

[8] `https://github.com/sastrawi/tokenizer`

Acknowledgements The research reported in this paper was conducted under the JSPS grant "Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers" offered to Tokyo University of Foreign Studies for a project entitled "A collaborative network for usage-based research on less-studied languages."

References

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media Inc.

Kawaguchi, Yuji. 2007. Foundations of Center of Usage-Based Linguistic Informatics (UBLI). In *Corpus-Based Perspectives in Linguistics*, ed. Yuji Kawaguchi, Toshihiro Takagaki, Nobuo Tomimori, and Yoichiro Tsuruga, 3–28. Amsterdam: John Benjamins.

Kudo, Taku, Koru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, 230–237.

Larasati, Septina Dian, Vladislav Kuboň, and Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In *Systems and Frameworks for Computational Morphology*, ed. Cerstin Mahlow and Michael Piotrowski, 119–129. Verlag: Springer.

Nomoto, Hiroki. 2013. Number in Classifier Languages. Doctoral Dissertation, University of Minnesota. URL `http://semanticsarchive.net/Archive/zBmYTg2Z/Nomoto2013_diss.pdf`.

Nomoto, Hiroki, Hannah Choi, David Moeljadi, and Francis Bond. (under review). MALINDO Morph: Morphology dictionary and analyser for Malay/Indonesian.

Rashel, Fam, Andry Luthfi, Arawinda Dinakaramani, and Ruli Manurung. 2014. Building an Indonesian rule-based part-of-speech tagger. In *International Conference on Asian Language Processing (IALP 2014)*. IEEE.

Riza, Hammam, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Oriental COCOSDA*.

Tan, Liling, and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-Multilingual Corpus). *International Journal of Asian Language Processing* 22:161–174. URL `http://www.colips.org/journals/volume22/22.4.2.NTU-MCTan_final.pdf`.