Introducing the Asian Language Treebank (ALT)

Ye Kyaw Thu†, Win Pa Pa‡, Masao Utiyama†, Andrew Finch†, Eiichiro Sumita†

†Advanced Speech Translation Research and Development Promotion Center, NICT, Kyoto, Japan
‡Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar,
{yekyawthu, multiyama, andrew.finch, eiichiro.sumita}@nict.go.jp, winpapa@ucsy.edu.mm

Abstract

This paper introduces the ALT project initiated by the Advanced Speech Translation Research and Development Promotion Center (ASTREC), NICT, Kyoto, Japan. The aim of this project is to accelerate NLP research for Asian languages such as Indonesian, Japanese, Khmer, Laos, Malay, Myanmar, Philippine, Thai and Vietnamese. The original resource for this project was English articles that were randomly selected from Wikinews. The project has so far created a corpus for Myanmar and will extend in scope to include other languages in the near future. A 20000-sentence corpus of Myanmar that has been manually translated from an English corpus has been word segmented, word aligned, part-of-speech tagged and constituency parsed by human annotators. In this paper, we present the implementation steps for creating the treebank in detail, including a description of the ALT web-based treebanking tool. Moreover, we report statistics on the annotation quality of the Myanmar treebank created so far.

Keywords: Asia Language Tree Bank (ALT), Myanmar language, word segmentation, alignment, part-of-speech tagging, tree

1. Introduction

We introduce and describe ongoing work in the creation of the Asian Language Treebank (ALT) corpus. Although the ALT corpus is mainly designed for statistical machine translation (SMT) of Asian languages, it will be useful in general for natural language processing (NLP) research. The corpus contains essential information for NLP tasks such as word segmentation, word alignment to parallel English, part-of-speech (POS) tagging and constituency parse trees. Additionally, this will be the first open Asian language treebank corpus. The ALT project is one of the language resource development projects of ASTREC and aims to accelerate research of NLP for Asian languages. Currently there is no publicly available POS-tagged and constituency tree corpus for most of the Asian languages and thus the APT project was created.

In this paper, we present the ALT corpus builing for Myanmar (the official language of the Republic of the Union of Myanmar) in detail. We will also introduce the ALT web-based software tool that was developed for annotating manual translations with word segmentation, word alignment, POS tagging and parse trees. This paper is organized as follows: Section 2. gives an overview of the design of the ALT corpus. In Section 3. the rules and development steps for the Myanmar language are described in detail. In Section 4., we give a brief introduction to the Asian Language Treebanking Tool (ALT Tool). In Section 5. an evaluation of the quality of the work in progress is given. Finally, Section 6. concludes and describes how we expect the ALT project to progress in the near future.

2. Overview of the Asian Language Treebank (ALT)

ASTREC, plans to coordinate the development of the ALT Corpus between 2014 to 2018. As a first step, the corpus is scheduled to cover: Indonesian, Japanese, Khmer, Laos, Malay, Myanmar, Philippine, Thai and Vietnamese languages by the end of this time span. In 2014, the project commenced development for the Japanese and Myanmar langauges. The domain is news and 1888 articles were randomly selected from English Wikinews (Wikinews, 2014). 20,000 sentences for building the corpus. Although preparing a parallel corpus may be sufficient for building a standard statistical phrase-based SMT system (Koehn et al., 2003), we also added manual alignment, POS tagging and constituency trees to facilitate further study on SMT and also for other NLP fundamental research. In order to create the corpus, we implemented a web-based tool. This tool will be used in collaboration with research institutions of several Asian countries. The data was represented in XML format for all development steps. The following is an example of the XML data for English sentence "Visitors at the hotel were evacuated to the exhibition hall at street level.":

```
<source>
  <text><![CDATA[Work began in 1900.]]></text>
      <words>
         <word><![CDATA[Work]]></word>
         <word><![CDATA[began]]></word>
         <word><![CDATA[in]]></word>
         <word><![CDATA[1900]]></word>
         <word><![CDATA[.]]></word>
      </words>
</source>
```

3. Developing for Myanmar Language

In this section, we will explain all the development steps in the construction of the Myanmar language ALT corpus. The Myanmar ALT corpus was developed in a collaboration between ASTREC and the University of Computer Studies, Yangon (UCSY). All the steps from translation to tree building were done manually with the help of UCSY within 1 year. In detail, 200 people worked on English-Myanmar translation and word segmentation, and 10 members of the Natural Language Processing Lab., UCSY worked on word alignment, POS tagging and tree building.

3.1. Translation

There are many ways to translate the same sentence, especially for professional translators and it is difficult to decide which is the best translation. (Secară, 2005) discussed different frameworks used in the process of translation evaluation with special focus on error classification schemes used both in the translation industry and in translation teaching institutions. Generally, there are groups of human translation errors and they are: content, lexis, grammar and text (Chiho et al., 2015). We carried out translation from English to Myanmar with non-professional but English-fluent UCSY staff (most of whom were teaching staff). We prepared a general instruction set for translators as follows:

1. Don't miss necessary information
   (e.g. Wrong Translation: A, C and Correct Translation: A, B, C)
2. Don't add unnecessary information
   (e.g. Wrong Translation: A, B, C and Correct Translation: A, B)
3. Take care to minimize spelling mistakes
   (e.g. Mistakes based on phonetic similarity are common in the Myanmar language, such as ကျား (tiger) and ကြား (hear))
4. Use the written style of the Myanmar language
   (e.g. using သည်, မည် instead of တယ်, မယ်)
5. If possible generate Myanmar that can be directly aligned to the English
   (i.e. avoid using idiomatic translation)

3.2. Word Segmentation

In Myanmar text, words composed of single or multiple syllables are usually not separated by white space, but in rare cases can be.

For the ALT corpus, we defined Myanmar 'words' to be meaningful units that correspond to a defined set of POS tags (in Section 3.4.). All suffixes and prefixes of nouns, adjectives, adverbs and verbs were defined were included (with exceptions) within words. For example, the particles "များ" and "တို့" corresponding to the English suffixes "s", "es" make nouns plural. These were defined to be included as part of the noun. Post positional markers of verbs "သည်", "၏", "ပြီ" and particles of adverb "စွာ" are also used within words. Loan words were segmented based on how they are segmented in the English corpus such as "ဂရိတ်တား မန်ချက်စတာ" for "Greater Manchester", "အင်တာနက် ဖိုရမ်" for "internet forum", "ဝီကီနယူးစ်" for "Wikinews". However, segmentation of some Myanmar particles and post positional markers were heuristically segmented depending on the context. In the example below we consider the particles "သာ" (meaning: only) and "ခဲ့" (past tense) in the sentence သူသာပထမဖြစ်ခဲ့သည် (meaning: Only he was the first.):

Segmented text: သူ သာ ပထမ ဖြစ်ခဲ့သည်
with POS: သူ\PRO သာ\PART ပထမ\N ဖြစ်ခဲ့သည်\V

Here, the word "သာ" has the meaning of "only" and thus segmented as a word the word "ဖြစ်ခဲ့သည်" is combination of root verb "ဖြစ်", past tense suffix "ခဲ့" and post positional marker "သည်" and was segmented as a word.

Although we already have a supervised CRF-based in-house Myanmar word segmenter for the basic travel expression domain (Win Pa et al., 2015), the domain of the ALT corpus is harder to segment, and to get higher word segmentation accuracy, we decided to do manual word segmentation.

3.3. Alignment

Bitext word alignment is an important step for SMT (Koehn et al., 2003). For this reason, the ALT corpus was designed to contain manually annotated word alignments between the original English and translated Asian language sentence. Some words like determiners and prepositions in English are omitted in translated Myanmar sentences. Similarly, some words like post positional markers in Myanmar don't have particular English words to align to. For example, the alignment of "the last" in "the last cars to finish" would be aligned to နောက်ဆုံး and မှ in နောက်ဆုံး မှ ပန်းဝင်သော ကားများ and likewise, the alignment of "to finish" would align to ပန်းဝင်သော. All words on both sides were required to be aligned to words in the other language. Unaligned words (null alignments) were not allowed (see Figure 1).

3.4. POS Tagging

For the ALT Project, we did not use existing POS tagsets for Myanmar such as (Phyu Hninn et al., 2011). Our POS tag set was intended to be simple and universal similar to proposal of (Slav et al., 2011) (Petrov et al., 2012). The main difference with the Universal POS Tagset is we added necessary language specific POS tags to a core tagset that will be shared with other languages. Myanmar parts of speech are different from English since the grammatical structure is subject, object, verb. Some English parts of speech like determiners, prepositions and auxiliary verbs are not used in Myanmar and some Myanmar parts of speech like post positional markers are not used in English. Although the Myanmar Thadda (a book on Myanmar grammar)

(Commission, 2005) defined 10 parts of speech (adjectives, adverbs, conjunctions, interjections, nouns, particles, post positional markers, pronouns, punctuation and verbs), we defined 14 POS tags to be used in the ALT corpus in order to get more detailed syntactic information of both source and target languages. They were as follows:

1. Abbreviation (ABB)
   (e.g. abbreviation of high school အထက, CNN စီအန်အန်, BBC ဘီဘီစီ)
2. Adjective (ADJ)
   Myanmar adjectives can be before or after a noun (e.g. beautiful girl can be written [လှသော]မိန်းကလေး or မိန်း[လှ])
3. Adverb (ADV)
   Myanmar adverbs are always before the verb and there can be more than one adverb for a verb. Most Myanmar adverbs are adjectives combined with the suffix "စွာ" (e.g. run quickly [မြန်မြန်]ပြေးသည်, [အလွန်] [မြန်ယှက်စွာ] စား)
4. Conjunction (CONJ)
   (e.g. and နှင့်, but သော်လည်း)
5. Foreign words (FOR)
   generally used for words of transliterated foreign origin
   (e.g. video ဗီဒီယို, Taekwondo တိုက်ကွမ်ဒို)
6. Interjection (INT)
   (e.g. follow follow လိုက်ဟ လိုက်ဟ, Oh God! ဘုရား ဘုရား)
7. Noun (N)
   (e.g. the largest city in Myanmar: Yangon ရန်ကုန်, happiness ပျော်ရွှင်ခြင်း)
8. Number (NUM)
   (e.g. 1 ၁, 2 ၂, 3 ၃)
9. Particle (PART)
   (e.g. friends သူငယ်ချင်း[များ])
10. Post positional marker (PPM)
    (e.g. ကျွန်တော်တို့ [သည်], Here the word [သည်] (I) followed by a Noun ကျွန်တော်တို့ (we) to show that the noun is a subject)
11. Pronoun (PRON)
    (e.g. I don't like it. ကျွန်တော် အဲဒါ ကို မကြိုက်ဘူး, Here ကျွန်တော်(I) and အဲဒါ (that) are pronouns)
12. Punctuation (PUNC)
    In Myanmar language, there is no question mark, exclamation mark, colon and semicolon, usually only two punctuation symbols are used, and the little section symbol (as comma in English) "၊" and section symbol (as full stop in English) "။". For the ALT project, we wished to represent the English language side punctuation for the alignment process and thus ‘, ’, (,), [,], {, }, “, ” and ⋯ were used on the Myanmar side.
13. Symbol (SB)
    There is no special symbols for Myanmar other than punctuation marks but there can be many symbols in translated Myanmar sentences like @, <, >, %, #, &, +, − and ∗.
14. Verb (V)
    Myanmar verbs formed by concatenating a stem word and a post positional marker. Tense information is encoded in a particle suffix to the stem word and is terminated by a post positional marker.
    (e.g. for the word "go", present tense: သွား[သည်, past tense: သွား[ခဲ့]သည် and future tense: သွား[လိမ့်]မည်)

3.5. Treebanking

The last step of ALT corpus development was the treebanking. Initially all words are child nodes of the root node. Then, nodes were repeatedly merged to form meaningful phrases until the root node had only two or three child nodes. Every new node was POS-annotated with the POS of a head chosen from the children. For example joining an adjective node and a noun node would result in either an adjective or noun parent node. The tree is generally a binary tree but some nodes can have 3 children. For example, the three nodes "(/PUNC'', "1/NUM" and ")/PUNC" form the phrase "(1)/NUM''.

4. Asian Language Treebanking Tool (ALT Tool)

The ALT Tool is a web-based application which allows the user to interact with the original English Wikinews data in order to make translations to the target language, and then perform word segmentation, word alignment, POS tagging and tree building on the target side in a user-friendly working environment. The application was developed with PHP, Javascript, HTML and CSS and is served using the Apache web server (Apache Software Foundation, 1997). jQuery JavaScript library (The jQuery Foundation, 2005) and Apache Bootstrap CSS Library were used. For the database, MySQL (Oracle Corporation and/or its affiliates, 2005) was used. This application also supports users, groups and task management and has a multilingual help system. The administrator can assign users to tasks based on their background knowledge (such as assigning only translation tasks etc.). If the user is assigned all tasks he/she can continuously and efficiently work through all processes for given sentence. For word alignment, POS tagging and tree building user input can be performed using only the mouse. Figure 1 shows the user interface for word alignment annotation between an English sentence and the corresponding translated Myanmar sentence, and Figure 2 shows the user interface for constituency tree building. Alignment between source and target words is easily done by clicking them (see Figure 1). It also supports two types of line colour that represent different levels of confidence of aligment. POS tagging is also easily done by selecting a POS tag from the dropdown list

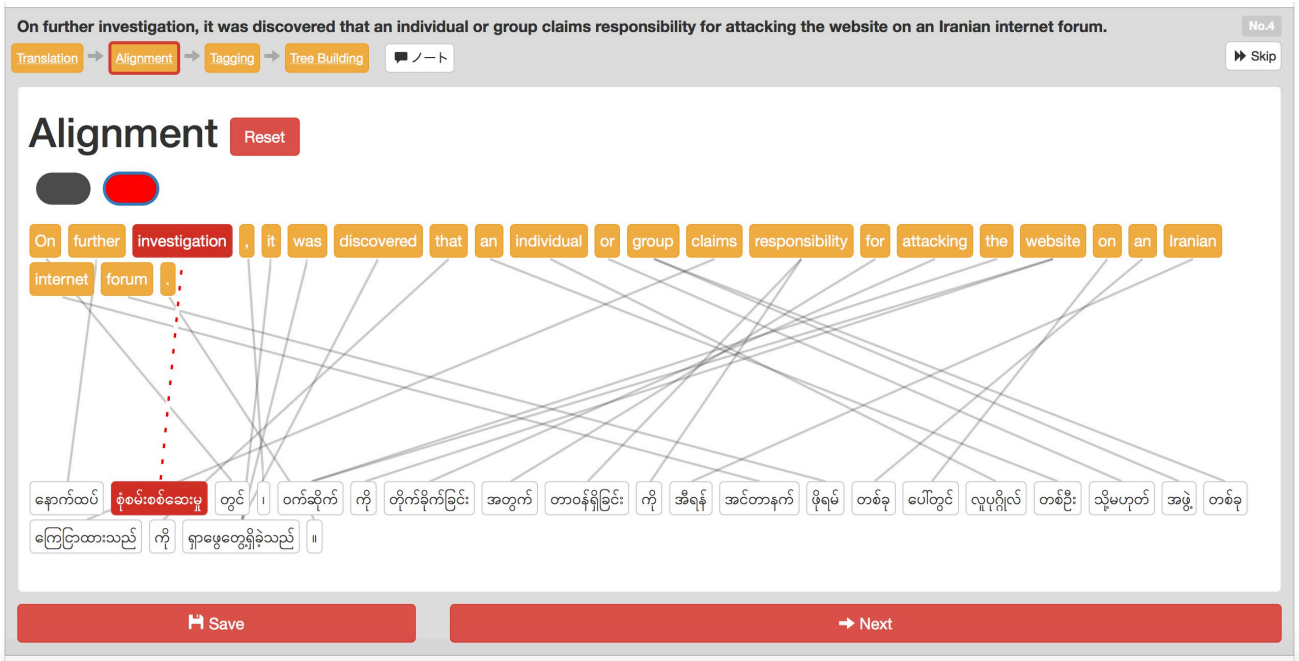

Figure 1: ALT Tool: Alignment interface.

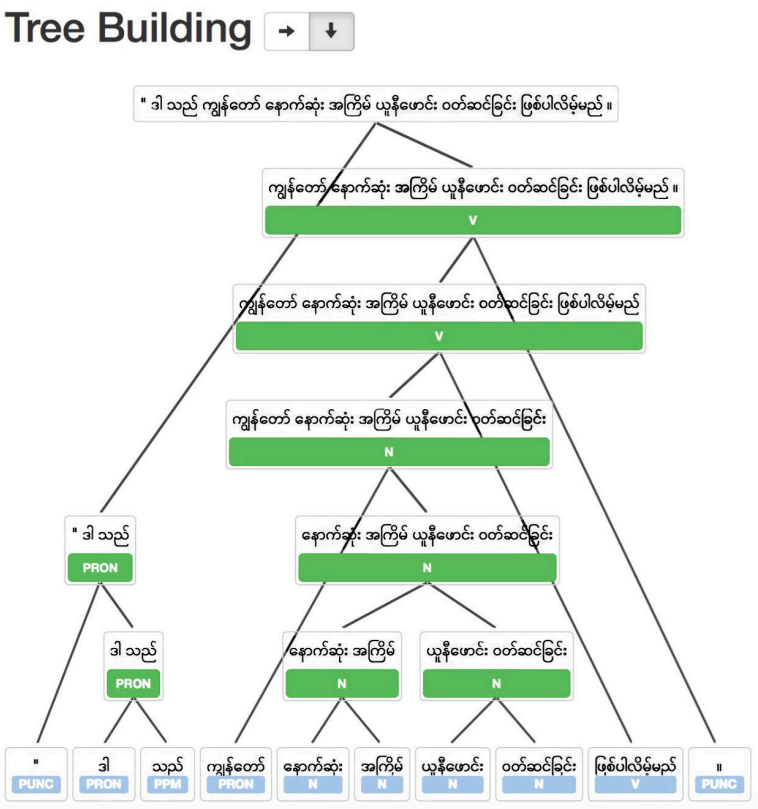

Figure 2: ALT Tool: Tree building interface.

of defined POS tags. Tree building is done bottom-up by clicking two nodes that will form a constituent (see Figure 2). The rule labels are selected from a drop-down menu.

5. Evaluation

There is no golden standard for evaluation of word segmentation, POS tagging, alignment and trees of Myanmar language. Error analysis evaluation was done manually by randomly sampling 10% (2,000 sentences) of the whole ALT Myanmar corpus and check-

ing for consistency with our annotation rules (Section 3.). The sentence-level error rate for translations was 2.88% (only semantically equivalent sentences were counted as correct, spelling and grammatical errors were counted as errors), the sentence-level segmentation error rate was 8.53%, and the sentence-level POS tagging error (excluding sentences with segmentation errors) was 12.28%. Statistics on the alignment quality and parsing error are currently being collected and will be available in the near future.

6. Conclusion and Future Work

This paper describes the process of building the Myanmar language component of the ALT corpus. This corpus is intended to accelerate NLP development in low resource Asian languages. The corpus consists of 20000-sentences from the news domain that will consist of Asian language translations from a shared English source text together with accompanying manual word segmentation, word alignment, POS tagging, and constituent parses. The project will expand in scope to include Indonesian, Japanese, Khmer, Laos, Malay, Philippine, Thai and Vietnamese in the short term, and extend to other languages in the long term through collaboration with international research organizations.

7. Bibliographical References

Apache Software Foundation. (1997). Apache web server.

Chiho, T., Kikuko, T., Anthony, H., and Kageura, K. (2015). Error categories in english to japanese translations. In *The 21st annual meeting of the natural language processing Japan (NLP2015)*, Kyoto, 3.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Oracle Corporation and/or its affiliates. (2005). Mysql.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Phyu Hninn, M., Tin Myat, H., and Ni Lar, T. (2011). Bigram part-of-speech for myanmar language. In *Proceedings of the 2011 International Conference on Information Communication and Management (ICICM 2011)*.

Secară, A. (2005). Translation evaluation- a state of the art survey.

Slav, P., Dipanjan, D., and Ryan, M. (2011). A universal part-of-speech tagset. In *IN ARXIV:1104.2086*.

The jQuery Foundation. (2005). jquery javascript library.

Win Pa, P., Ye Kyaw, T., Finch, A., and Sumita, E. (2015). Word boundary identification for myanmar text using conditional random fields. In *Genetic and Evolutionary Computing*. Springer International Publishing Switzerland.

8. Language Resource References

Myanmar Language Commission. (2005). *Myanmar Thadda (in Myanmar language)*. Ministry of Education, Combination of 3 Volumes, 1st.

Wikinews. (2014). *https://en.wikinews.org/wiki/Main_Page, Accessed: 2014-07-15*.