

Alignment Tool for Treatment ParSit Result

Tipraporn Thanakulwarapas, Sittha Paholpinyo and Thepchai Supnithi

Text Processing Section, Division of Research and Development on Information,
Nation Electronics and Computer Technology Center

112 Paholyothi Road, Klong 1, Klongluang, Pathumthani, 12120, Thailand

{tipraporn.thanakulwarapas,sittha.paholpinyo,
thepchai.supnithi}@nectec.or.th

Abstract

In this paper we propose an alignment tool in ParSit_TREAT, which is an extension of ParSit, an English to Thai machine translation. We aim the system to assist ParSit to improve the accuracy of MT.. The alignment tool is functioned to support linguists to modify result from ParSit to handle the some errorcases.

1 Introduction

The rapidly growth of Internet society causes a lot of information in digital resources to construct. Exchanging information is required to serve a globalized level of communication. Based on language diversity, translation becomes an important topic as a way to assist people realize and exchange information to others. In addition, it is a way to reduce the digital divide based on language barrier.

There are many English to Thai machine translation softwares. Most of them are developed based on ruled based approach [3][6]. There are very few systems that apply memory-based approach [4][9]. The quality of translation is around 30-70% based on the complexity of input sentence.

In our project, we developed ParSit, an English to Thai machine translation, with the collaboration with NEC Corporation, Japan. The accuracy is around 60%. To improve the accuracy of MT, there are two main approaches; one is to improve the analysis rule and generation rule in ParSit, another is to apply a post-processing module to correct the incorrect translation result. We conducted an experiment on applying the first method to ParSit and found that there are some problems due to conflict among rules. Some correct translation patterns give the incorrect translation result. Currently, we aim at the latter technique by adding a

Post-Edit module in ParSit system as a post-processing process. [7] analyzed problems in translation error types and introduced a framework to add post-edit module in ParSit.

To realize the post processing process, it is necessary to collect the Post-Edit corpus, which is input data for generating a set of rule for Post-Edit module. This subsystem is called ParSit_TREAT.

In this paper, we describe visualization tool for aligning data. Alignment tool is a free, open-source, portable, user-friendly, GUI-driven program written in Java that provides a visual representation of word correspondences between bilingual pair of sentence. Our interface includes both, alignment¹ visualization and post-editing tool for word alignment.

The structure of this paper is shown as follows. Section 2 gives the overview of ParSit system architecture. Section 3 explains the implementation, important feature and example usage. Section 4 illustrates the data model in our application respectively. Finally, conclusion and future work is stated.

2 System Architecture

As illustrated in Figure 1, the system can be divided into two subsystems. The first subsystem is ParSit with Post-Edit Module. The second subsystem is ParSit_TREAT.

2.1 ParSit with Post-Edit module

As shown in the right side in Figure 1, ParSit is an English to Thai machine translation system. It applied rule-based technique with interlingua approach. When an end user submits an English sentence into ParSit, its translation result will be consequently constructed. The translation results from ParSit are sometimes not satisfactory

¹ We will use the term “alignment” to refer to the set of all word correspondences between a sentence pair.

because of linguist problems, such as, ungrammatical sentence construction,

incorrect word orderings, word omissions, etc. Post-Edit

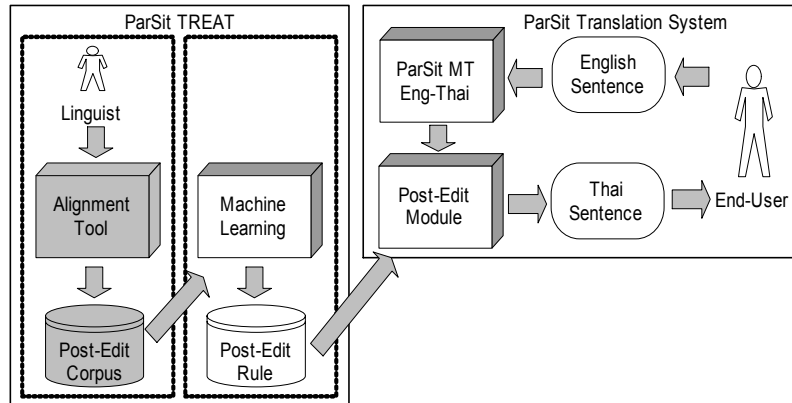


Figure 1 Parsit with Alignment Tool

Module plays an important role as a module to refine the translation results of ParSit before sending back to the end user. It applies our prepared post-edit rule set as a decision criterion.

2.2 ParSit_TREAT

As shown in the left side in Figure 1, ParSit_TREAT is a subsystem that enables us to retrieve linguists' criteria and generate rules for Post-Edit module. TREAT stands for translation result editing alignment tool. When a linguist accepts a list of ParSit translation results, he/she will compare the original source and target source, and then edit the ParSit translation results. In this step, linguists have to edit them by considering the linking between original, ParSit result, and edited result simultaneously. The corrected translations, which are corrected by linguists, are accompanied with the original ones, and piled up into the Post-Edit corpus. Afterwards the post-edit corpus is learned by Machine Learning algorithms to construct Post-Edit rules.

2.3 Alignment Tool in ParSit_TREAT

Constructing the Post-Edit corpus is a major task in ParSit_TREAT. Since it is difficult to map situations among original, ParSit translation and edited translation simultaneously, linguists need to analyze a translation visually and manually correct it. Alignment Tool is developed to facilitate their works. Resembling the graph-based interface of Cairo [8][5], the graphical user interface is comprehensive and is easy to correct the translations. The work presented in this paper is emphasized in this tool.

3 Implementation on Alignment Tool

Alignment Tool was designed in graph based interface to assist linguists easily edit the incorrect results. It was implemented in JAVA (with JDK 1.5.0) [2].

3.1 Alignment Tool GUI

Alignment Tool displays the given three sentences with lines drawn between aligned words. The GUI of Alignment Tool is shown in Figure 2.

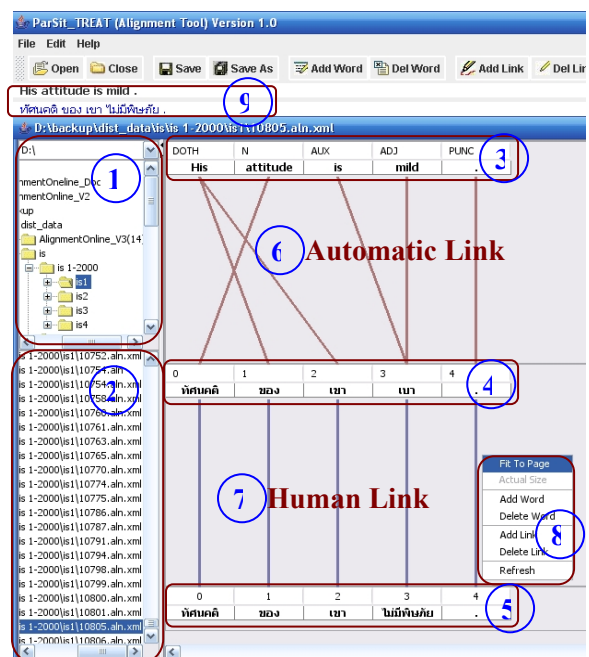


Figure 2 Alignment Tool program

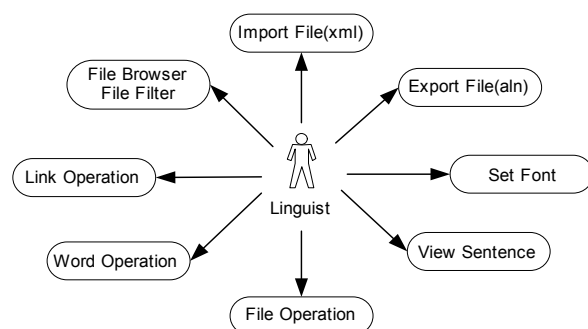
When the user firstly loads an alignment file, the user selects the drive and directory to locate the alignment result as shown in (1). All alignment files in the selected directory, which restrictedly

have the ALN² and XML extensions, will be listed in (2). When he/she selects an alignment file, the translation alignment will be displayed in the right window. It composes of the English sentence, ParSit translation result, as displayed in (3) and (4) respectively. (5) represents an edited sentence which is corrected by linguists. If a file is opened first time, (5) displays the same content as shown in (4) as the default content. Word alignment is automatically represented in (6) and (7) respectively. (6) is an output generated from ParSit. The correspondences among (3) to (7) help a linguist to correct the translation. The linguist is able to correct the translation only in (5) and (7). He/she can insert a new word, remove words, and edit incorrect words in (5). In parallel, he/she can correct the linking mistakes in (7) by adding new link, removing the inappropriate links. The linguist is guided to provide the correction based on the correspondence between the original translation and the human-edited translation. These correspondences are the most vital, since they will be learned by the Machine Learning algorithms for automatic translation correction. Linguists may use popup menu in each correction process as shown in (8). If the English sentence or the translation sentence cannot be fitted in the screen, the sentence previewer is also provided in (9).

In the figure 2, an English sentence “His attitude is mild” is an original sentence. It is translated to “ที่ ศนคติของ เขาเบา”. Since a word “mild” which is translated to “เบา” is incorrect, linguist correct from “เบา” to “ไม่ มีพิษภัย”.

3.2 Features

Alignment Tool is applied to correct translation result from ParSit. Features on Alignment Tool are shown in Figure 3.



² ALN file is an output file from ParSit

Figure 3 Features on Alignment Tool

Import File: to convert from ALN file format to XML file format.

Export File: to convert from XML file format to ALN file format.

Set Font: to set font type and set font size of words.

View Sentence: to preview sentence after edit.

File Operation

- Save As: to save XML, ALN file into new file name or new file format.

- Save: to save XML file into the same directory.

- Open: to open XML, ALN file in a new window.

- Close: to close file.

Word Operation

- Delete Word: to remove an unnecessary word.

- Add Word: to insert a new word.

Link Operation

- Delete Link: to remove unnecessary link.

- Add Link: to insert link.

File Browser and File Filter: to select file and display only XML and ALN

4 Data Model for Alignment Tool

Based on the GUI structure, there are two major objects represented our data. A node represents a word in a sentence. A link represents a relation between two nodes.

4.1 Data Structure Format

We represent information in node and link based on XML format. Table 1 presents tags and description of tags in each object.

4.2 Example XML File

Figure 4 illustrates an alignment file produced from Alignment Tool. An English sentence, its Thai translation from ParSit, and the human-edited translation are embraced in the tags <source> – the source language – in (a), <target> – the target language – in (b), and <reference> in (c), respectively. Each word is separated by a blank. English words are also annotated with their parts of speech separated by the vertical bar ‘|’; *attitude|N* for the word ‘attitude’ with the part of speech ‘N’ (noun), for example. Word alignments between the English sentence and the ParSit’s translation are stored in the tag <alignment> in (d), while word

alignments between the ParSit's translation and the human-edited translation are stored in the tag `<reference_alignment>` in (e). Within these two tags, word IDs are aligned and stored

as links $W_{\text{begin}}-W_{\text{end}}$, where W_{begin} is a beginning word ID, and W_{end} is the terminal word ID.

Table 1: XML tags of the alignment file and their descriptions

Object	Tag Name	Tag Description
NODE	<source>	Source sentence
	- <sentence>	Original sentence in "word part of speech" format
	- <words>	Word count in Original sentence
	- <word id>	Word ID
	- <content>	Word content
	- <pos>	Part of speech
	<target>	Target sentence
	- <sentence>	Automatic translation sentence from ParSit
	- <words>	Word count in automatic translation sentence
	- <word id>	Word ID
	<reference>	Reference sentence
	- <sentence>	Human edited translation sentence
- <words>	Word count in human edited translation sentence	
- <word id>	Word ID	
LINK	<alignment>	Alignment link between source and target sentence
	- <link>	Link ID
	- <begin>	Begin point of a link from source sentence
	- <end>	End point of a link from target sentence
	<reference_alignment>	Alignment link between target sentence and reference sentence
	- <link>	link ID
- <begin>	Begin point of a link from target sentence	
- <end>	End point of a link from reference sentence	

The screenshot shows an XML file named '10805.xml' in Microsoft Internet Explorer. The XML content is as follows:

```

<?xml version="1.0" encoding="UTF-8" ?>
- <application minor="b001" name="alignmentonline2" structure-version="1.0" version="1.0">
  - <source>
    - <sentence>His|DOTH attitude|N is|AUX mild|ADJ ./</sentence>
    - <words count="5">
      - <word id="0">
        - <content>His</content>
        - <pos>DOTH</pos>
      - </word>
      ...
    - </words>
  - </source>
  - <target>
    - <sentence>ทัศนคติ ของ เขา เขา ./</sentence>
    - <words count="5">
      - <word id="0">ทัศนคติ</word>
      ...
    - </words>
  - </target>
  - <reference>
    - <sentence>ทัศนคติ ของ เขา 'ไม่มีทัศนคติ' ./</sentence>
    - <words count="5">
      - <word id="0">ทัศนคติ</word>
      ...
    - </words>
  - </reference>
  - <alignment>
    - <content>0-1, 1-0, 2-0, 3-2, 3-3, 4-4</content>
    - <links count="6">
      - <link id="0">
        - <begin>0</begin>
        - <end>1</end>
      - </link>
      ...
    - </links>
  - </alignment>
  - <reference_alignment>
    - <content>0-0, 1-1, 2-2, 3-3, 4-4</content>
    - <links count="5">
      - <link id="0">
        - <begin>0</begin>
        - <end>0</end>
      - </link>
      ...
    - </links>
  - </reference_alignment>
</application>

```

Annotations in the image:

- Original Sentence:** Points to the source sentence tag.
- Word and POS:** Points to the word 'His' and its part of speech 'DOTH'.
- Automatic Sentence:** Points to the target sentence tag.
- Word:** Points to the word 'ทัศนคติ' in the target sentence.
- Human Sentence:** Points to the reference sentence tag.
- Word:** Points to the word 'ทัศนคติ' in the reference sentence.
- Original Link:** Points to the alignment content '0-1, 1-0, 2-0, 3-2, 3-3, 4-4'.
- Start and End Link:** Points to the begin and end tags of the first alignment link.
- Human Link:** Points to the reference alignment content '0-0, 1-1, 2-2, 3-3, 4-4'.
- Start and End Link:** Points to the begin and end tags of the first reference alignment link.

Figure 4. An example of alignment file generated from Alignment Tool

5 Conclusion and Future work

In this paper, we presented Alignment Tool, an important part of ParSit_TREAT, with a user interface for interactive word alignment. This tool provides flexible interface for revising words and connection links. This tool makes convenience for linguists to edit translation results because of their new user-friendly interface, and also reduces the time consumption to handle with various forms of editing. It was applied to collect a post-edit corpus. Currently, there are ... sentences that are collected.

In the future, we would like to add several features, such as displaying alternate word's meaning (help linguists edit e.g. the program highlights words, then shows the list of meanings for selection) to the word alignment interface. Optional features for complex sentence are possibly added. To reduce confusing from unsighted some part of sentence in a screen, enabling the system shows only each clause of sentence is also taken into consideration .

Acknowledgment

Special thanks to Dr. Krit Kosawat, Ms. Monthika Boriboon in collaboration with the development on ParSit English-Thai MT system. Mr. Sawawut Kongyang , Mr. Nattapol Kritsuthikul, and Mr. Prachya Boonkwan to give a valuable advice in programming technique and alignment concept.

References

- [1] Al-Onaizan, Yeser, Jan Curin, Michael E. Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith and David Yarowsky. 1999. *Statistical Machine Translation: Final Report*, Johns Hopkins University 1999 Summer Workshop (WS 99) on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, USA.
- [2] Cay S. Horstmann and Gary Cornell, 2005. *Core Java 2 fundamentals*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.
- [3] John W. Hutchins and Harold L. Somers. 1992. *An introduction to machine translation*, Academic Press, London, UK.
- [4] Satoshi Sato, Makoto Nagao. 1990. *Towards memory-based translation*. Proceedings of COLING'90, Helsinki, 247-252.
- [5] Sentence Alignment and Word Alignment Projects:
<http://www.cs.unt.edu/~rada/wa/#projects>

- [6] Sergei Nireburg, Jaime Carbonell, Masaru Tomita and Kenneth Goodman. *Machine Translation: A Knowledge-Based Approach*.
- [7] Sitthaa Phaholphinyo, Teerapong Modhiran, Nattapol Kritsuthikul and Thepchai Supnithi. 2005. *A Practical of Memory-based Approach for Improving Accuracy of MT*, in the proceeding of Conference Machine Translation Summit 10 th (MT SUMMIT X), Phuket, Thailand.
- [8] Smith, Noah A. and Michael E. Jahr. 2000. *Cairo: An Alignment Visualization Tool*, in the proceedings of the Second International Conference on Language Resources and Evaluation(LREC 2000), Athens, Greece.
- [9] Van den Bosch A., and Daelemans W. 1999. *Memory-based Morphological Analysis*, Proceedings of ACL'99, 37th Annual Meeting of the Association for Computational Linguistics, Maryland, USA.