

Word Sense Disambiguation using Parallel Corpora for Cross Language Information Retrieval

Thitiporn Tubtimate and Chuleerat Jaruskulchai

Department of Computer Science
Faculty of Science
Kasetsart University, Bangkok, Thailand.
e-mail : g4464002@ku.ac.th , fscichj@ku.ac.th

ABSTRACT

Cross Language Information Retrieval (CLIR) focuses on the retrieval of documents based on queries formulated by users in their mother language for retrieval of documents in another language. However, the problem of basic query translation using dictionary mapping is language translation ambiguity and missing terms in the dictionary.

A parallel corpora has been used successfully in solving translation ambiguity. Thus, the Thai bitext is compiled and lexically-based techniques used to align text. Our alignment text based on paragraph and parallel corpora can increase new words about 10 %.

Keywords: Word Ambiguity, Parallel Corpora

1. Introduction

Cross Language Information Retrieval (CLIR) is a system for retrieving documents across language boundaries. A query written in one language should be translated into a representation for finding documents in another language that is consistent with the information need [3].

Approaches to CLIR can be categorized into three categories : query translation, document translation, or by using both query and document translation. Full document translation for large collections is impractical, thus query translation is the main alternative approach. Methods for translation have focused on three areas. The first is machine translation (MT) techniques. This is an automatic process that translates from one human language to another language by using context information. The second translation method is word-by-word translation using a machine-readable dictionary, which is widely available, and can be easy to develop. The third approach is parallel corpora, which is a collection of pairs of documents in two languages that are known to be translations of one another [2, 4].

Regardless of the cross-language approach taken, translation ambiguity is a problem which must be addressed. There are two types of ambiguous translation : within-language ambiguity and cross-language ambiguity [23]. Firstly, within-language ambiguity is the problem of where a source language has many possible translations – for example in the sentence, ‘เขาให้เงินได้กับเจ้าหน้าที่’, ‘ได้’

might mean ‘the corruption’ or express ‘the point of action’. Next, cross-language ambiguity is a problem where a source language has more than one translation alternative in the target language. For example, for the direct mapping word ‘ด้วง’, we found two meanings in the dictionary. In the sentence, ‘มะพร้าวคันนี้มีด้วงอยู่เต็ม’, ‘ด้วง’ means the specific name of insects. But in the sentence, ‘ฉันชอบเล่นขอด้วง’, ‘ด้วง’ means the specific name of a musical instrument.

Thus, word-by-word translation will face a challenge in word ambiguity, as the Thai lexicon can be mapped by difficult meanings in English. In this paper, we focus specifically on the problem of word translation ambiguity, using co-occurrence statistics over target language reference parallel corpora.

This paper is organized as follows: Section 2 briefly reports the previous works in parallel corpora and Thai CLIR. Section 3 describes our procedure to form parallel corpora, and the experimental results are presented in Section 4. The conclusion is then given in Section 5.

2. Previous Works

2.1 Parallel Corpora

A parallel corpora is a bitext written of the same story in two different languages, and the translation from one language to another is done by human experts. These parallel corpora serve various purposes in language technology, not only for cross language retrieval but also for machine translation. To utilize this resource, it is necessary to align this bitext. Thus, the alignment bitext is a process to identify matching sentences between languages.

Recently, automatic alignment parallel bilingual corpora have been proposed. Approaches to alignment fall into two main classes: lexical and statistical. The lexically-based techniques use an electronic bilingual dictionary to match sentences. In contrast, the statistical technique relies on the lengths of sentences, sentence position, and co-occurrence frequency. The parallel corpora’s research which is related to CLIR is summarized as follows.

The English-Chinese information retrieval [14] uses parallel corpora to build a character-based statistical translation model. This model calculates the probability of a Chinese character giving an English word and the

model predicts only characters, not the order of characters.

In English-Arabic CLIR [8], most Arabic words can be derived from the same roots which causes a high level of ambiguity. Thus, this research combined statistical MT from parallel corpora and manual lexicon translation that gave a higher weight to parallel corpora.

The next approach to solving ambiguity is co-occurrence statistics. A co-occurrence method is the ratio of the frequency for each possible pair of definitions in translations. [13] presents a technique based on co-occurrence statistics collected from unlinked corpora, which has successfully solved the ambiguity in phrase translation.

In Thai parallel corpora research, there is no work reported in this area. Although the Internet is widely used in Thai society, some of the parallel corpora are not publicly accessible. A collection of Thai bitext compiled by Doug Cooper is available at <http://crcl.th.net/> but this collection is only 6 stories with no report on how the corpora is aligned.

2.2 Thai CLIR

Most researchers in Thai CLIR try to solve unseen words which may be borrowed from foreign languages and used in Thai scripts or transliterated words. Soundex technique [19], Neural Network technique [18], and Fuzzy Matching technique [10] are implemented to solve the backward transliteration. The evaluation of these techniques are based on specific domain, for example, the effectiveness of Soundex technique is evaluated on a list of Thai names. The CLIR's work done in [7], proposed query translation using simple dictionary mapping and evaluated the retrieval effectiveness with CLEF's data test. The challenge of Thai CLIR is the coverage of the dictionary and the between-language translation ambiguity. Soundex and N-gram techniques are proposed to detect the transliterated words.

Some of Thai research on Thai word ambiguity should be categorized in the within-language ambiguity, such as Thai word recognition in Natural language processing [9] and the feature-based approach in Thai word segmentation [1]. The main problem is word meanings. The within-language ambiguity is caused by lexical or structural ambiguity. For an example of lexical ambiguity, the word 'กิน' has more than one meaning. In English, the direct meaning is 'to eat' but the implicit meaning is 'to cheat'. When a phrase or sentence has more than one structure it is said to be structurally ambiguous. Therefore, the reader can get the true meaning from the context in the sentence. Another problem of translation is the coverage of vocabulary in the dictionary [22], which comes from proper names and transliterated words [17]. An example of a proper name is 'ชินวัตร' which is a person's name. Next, for example the word : 'แบคทีเรีย', transliterated as 'bacteria', is a Thai-to-English transliterated word, and 'Thaksin', transliterated as 'ทักษิณ', is an English-to-Thai transliterated word.

In the previous work about Corpus-based dictionary, we found the publication, [20] presented a task for

building a dictionary, which exhibit explicit word boundaries. The practical solution for this task by applying the C4.5 learning algorithm for building the lexicon list.

3. Approach

Our approach of building parallel texts for word sense disambiguation comprises the following tasks: pre-processing, aligning and matching the words in paragraph boundaries.

3.1 Corpus and Lexical Resources

Since Thai bitext is not available for public usage, this bitext is manually collected Future Magazines and Idea Magazines from 1999 – 2004, totalling 120 papers (Under copyright of Odien Store Press). The content of this bitext is classified into three categories : Feature, News and Science, with 40 papers for each category. The basic statistic of Thai bitext is shown in table 1.

Table 1: Comparison of characters, words and paragraphs in Thai-English documents

	<i>Feature</i>		<i>News</i>		<i>Science</i>	
	Eng	Thai	Eng	Thai	Eng	Thai
<i>Characters</i>	52,269	62,099	44,302	50,245	46,574	57,907
<i>Words</i>	9,834	13,527	7,959	10,513	8,606	12,621
<i>Paragraphs</i>	367	367	373	373	307	307

Our procedure to compile the bitext was through the following steps. Firstly, all documents in Thai were passed through a word-segmentation process using the segmenter reported in [11]. Then, the results were inspected word boundaries manually recheck again. The title and description were separated by markup tags as follows. The topic was defined as <TITLE> tag and the description defined as <DESC> tag.

In Table 1, the number of paragraphs in the Thai and English documents are equal which we considered parallel text in paragraph boundaries.

3.2 Alignment Technique

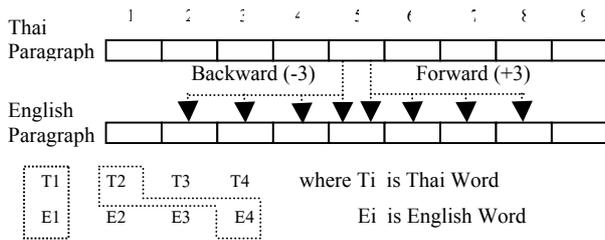
The word gathering for parallel corpora can be compiled by mapping word alignment. There are many research works in this area such as the statistical technique used in our approach that has been researched in the Slovene-English parallel corpora [21] and the French-English parallel corpora [16]. The end result of these is sentence-aligned text. The alignment information might be encoded in one of several ways, as the prototype language may be much longer when translated into the other or shorter.

LEXITRON [12] is used in the word alignment approach. Our procedure for word alignment had the following steps. Firstly, since most parallel text has a number of Thai stopwords in the sentences, we collected the list of stopwords to apply the linguistic rule. In cases that the Thai stopword has not been found in the dictionary or was found but the mapped English word was incorrect, we assumed that this word could be skipped

and no meaning translated, such as the word ‘ได้’, ‘ก็’, which is not part of the main vocabulary in the sentences.

Secondly, Thai-English mapping of words from the dictionary. If a Thai word has more than one meaning in English, we chose the best target translation by using parallel corpora. Finally, un-mapping of Thai words that cannot be translated from the dictionary. We collected co-occurrence statistics from the remaining Thai and English words considered by 3-word counting forward and 3-word counting backward.

Fig.1: Showing Thai-English Word Matching Algorithm in paragraph boundaries.

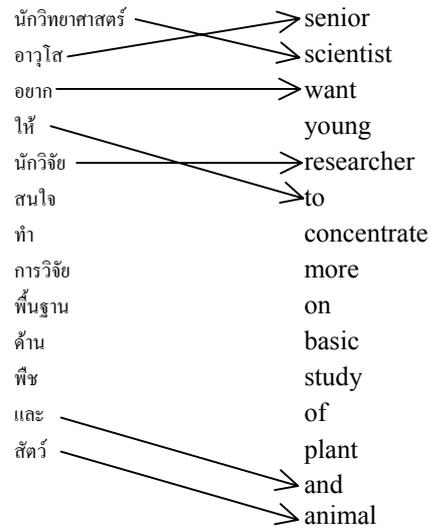


As seen in Figure 1 T_i is similarly translated to E_1 , and T_2 to E_4 by matching derived from the dictionary, so that the possibility of similarity between both languages are $P(T_3 | E_2)$, $P(T_3 | E_3)$, $P(T_3 | E_2, E_3)$, $P(T_4 | E_2)$, $P(T_4 | E_3)$, and $P(T_4 | E_2, E_3)$. The frequency of words in each meaning is used as a decision-making indicator to select the most frequently-generated words in the document to form parallel corpora [5].

4. Results

Table 2 shows detailed steps of our algorithm. The first part is an example of a Thai-English original sentence. The second part is Thai-English words after segmentation. The third part is an output of English meaning using simple mapping by LEXiTRON. The last part is the remaining un-matching parallel words will be mapped using sliding window algorithm.

Output Thai → English Translations :



Output Thai-English Word Matching :

<u>Thai</u>	<u>Eng#1</u>	<u>Eng#2</u>	<u>Eng#3</u>	<u>Frequency</u>
การวิจัย	young	#	#	1
การวิจัย	young	concentrate	#	1
การวิจัย	young	concentrate	more	1
การวิจัย	concentrate	#	#	1
การวิจัย	concentrate	more	#	1
การวิจัย	concentrate	more	on	1
การวิจัย	more	#	#	1
การวิจัย	more	on	#	1
การวิจัย	more	on	basic	1
การวิจัย	on	#	#	1
การวิจัย	on	basic	#	1
การวิจัย	on	basic	study	1

Table 2 : An example of Thai-English pair in the parallel corpora

Thai Sentence : นักวิทยาศาสตร์อาวุโส อยากให้นักวิจัยสนใจทำการวิจัยพื้นฐานด้านพืชและสัตว์

English Sentence : Senior scientists want young researchers to concentrate more on basic studies of plants and animals.

Thai Words : นักวิทยาศาสตร์, อาวุโส, อยาก, ให้, นักวิจัย, สนใจ, ทำ, การวิจัย, พื้นฐาน, ด้าน, พืช, และ, สัตว์

English Words : senior, scientist, want, young, researcher, to, concentrate, more, on, basic, study, of, plant, and, animal

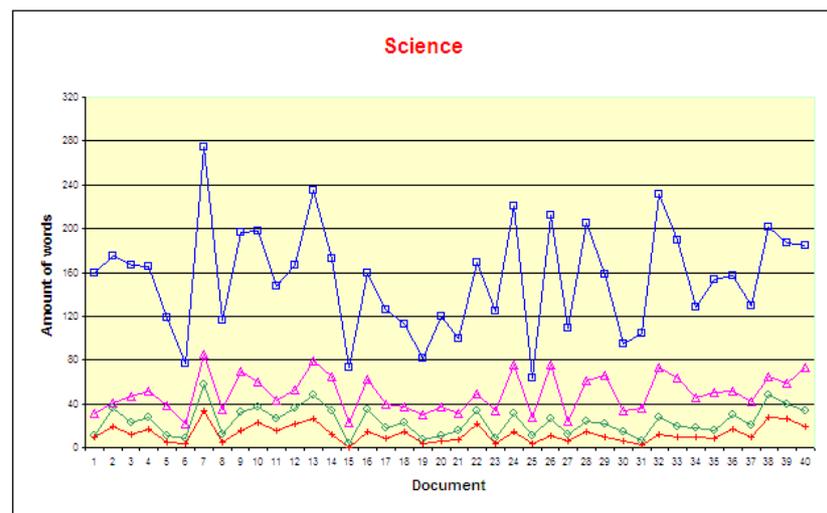
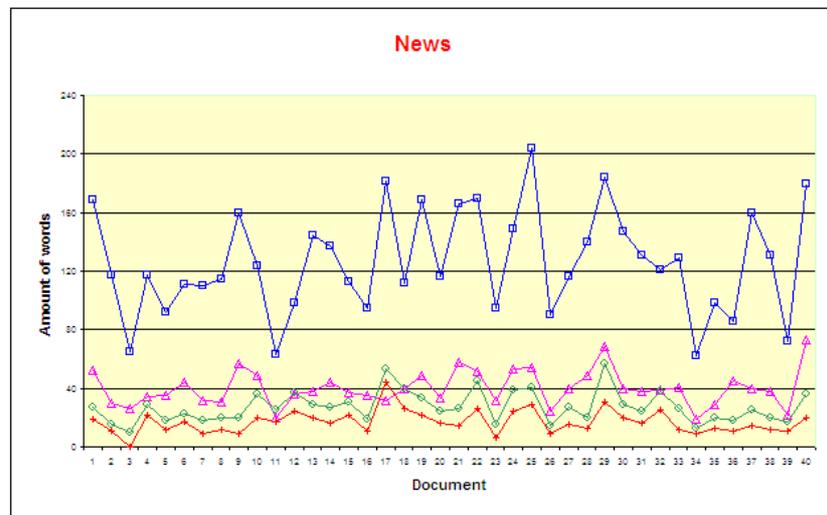
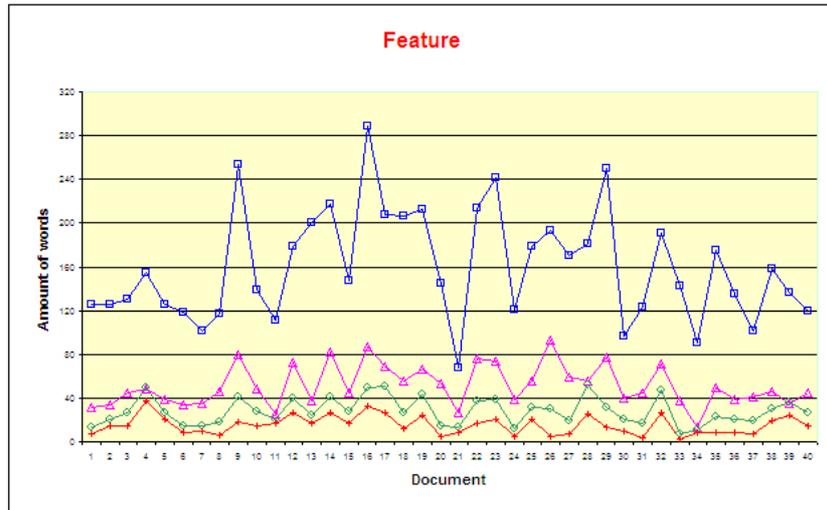


Fig.2: Indicates the results of the number of words brought for parallel corpora in 3 categories of documents : Feature, News, and Science.

The graphs in Figure 2 : A Square symbol ---□--- shows all Thai vocabulary found in the documents ; a Triangle symbol ---△--- shows the number of Thai words mapping English words found in LEXiTRON ; a Circle symbol ---○--- shows the number of Thai words mapping English words using parallel text with frequency more than one ; a Vertical symbol ---|--- shows the number of Thai words mapping English words that the meaning was never found in LEXiTRON with frequency more than one.

Figure 2, shows new words learned from parallel corpora. In particular, most new words were found in the News category more than the Science and Feature categories.

To measure the accuracy of the algorithm, we consider two statistical values : precision and recall. Test set is random 20 Thai words from 10 documents. Thai words are found various meanings in English by using dictionary mapping with frequency more than one in parallel corpus. The precision of our algorithm is 77.67% and the recall is 92.50% for the test set.

In our experiment, we faced the problem of word boundaries. Even humans still do not agree in the word boundary problem. For example, in the word : ‘ความสามารถพิเศษ’, the meaning using dictionary mapping in English is ‘talent’. If the Thai word is separated into small words, ‘ความสามารถ : ability’ and ‘พิเศษ : special’, we could not get the meaning ‘talent’. Another problem is that the data size is too small. Most of the word alignment is one-to-one alignment. This result is different from Slovene-English parallel corpora [21] and the French-English parallel corpora [16], where word alignment may be one-to-many or many-to-one.

5. Conclusion and Future work

In this paper, we compiled and studied the characteristics of Thai parallel corpora, and the basic approach for Thai-English parallel alignment was explored. The parallel text can discover new words, which can be placed into five categories : Specific name 3.69 %, Transliterated word 1.56 %, Acronym 1.28 %, Loan word 1.17 % and General word 2.30 %.

Our future work will be to evaluate this approach for solving the word ambiguity and adapting to CLIR.

6. REFERENCES

- [1] Charoenpornasawat Paisarn. 1998. *Feature-based Thai Word Segmentation*. Department of Computer Engineering, Graduate School, Chulalongkorn University : Bangkok.
- [2] D.Oard and B.Dorr. 1996. *A Survey of Multilingual Text Retrieval*. Technical Report UMIACS-TR-96-19CD-TR-3615. University of Maryland, College Park.
- [3] David A. Evans, Gregory Grefenstette, Joop van Gent, Yan Qu. 2002. *Anatomy of Commercial CLIR Application*. CLEF Workshop 2002, Rome, Italy.
- [4] Gerard Salton, and Michael J.McGill. 1983. *Introduction to Modern Information Retrieval*. New York : McGraw-Hill, Inc.
- [5] Hiroshi Masuichi, Raymond Flournoy, Stefan Kaufmann and Stanley Peters. 1999. *Query Translation Method for Cross Language Information Retrieval*. Center for the Study of Language and Information, Stanford University.
- [6] Hwee Tou Ng , Bin Wang and Yee Seng Chan. 2003. *Exploiting Parallel Texts for Word Sense Disambiguation*. Department of Computer Science , National University of Singapore.
- [7] Jaruskulchai Chuleerat. 2001. *Dictionary-based Thai CLIR : Experimental Survey of Thai CLIR*. Department of Computer Science, Faculty of Science, Kasetsart University.
- [8] Jinxi Xu, Alexander Fraser and Ralph Weischedel. 2001. *TREC 2001 Cross-Lingual Retrieval at BBN*. BBN Technologies, Cambridge, MA 02138.
- [9] Kawtrakul, A., Thumkanon, C., Poovorawan, Y., Varasrai, P. and Suktarachan, M. 1997. *Automatic Thai Unknown Word Recognition*. In Proceedings of the Natural Language Processing Pacific Rim Symposium 1997.
- [10] Khantonthong Navapat. 2001. *Automatic Backward Transliteration for Thai Text*. Department of Computer Engineering, Faculty of Engineering, Kasetsart University : Bangkok.
- [11] Kruengkrai, C. And C. Jaruskulchai. 2001. *Thai Text Document Clustering using Parallel Spherical K-Means Algorithm on PIRUN Linux Cluster*. The Fifth National Computer Science and Engineering Conference.
- [12] LEXiTRON, Thai <-> English Dictionary Software and Language Engineering Laboratory, National Electronics and Computer Technology Center, http://www.links.nectec.or.th/lexit/lex_t.html
- [13] Lisa Ballesteros, W.Bruce Croft. 1998. *Resolving Ambiguity for Cross-Language Retrieval*. Center for Intelligent Information Retrieval, Computer Science Department : University of Massachusetts, USA.
- [14] M. Franz, J.S. McCarley. 2000. *English-Chinese Information Retrieval at IBM*. IBM T.J. Watson Research Center.
- [15] Ongrungruang Siriporn. 1997. *English to Thai Word Retrieval System using Sound-Like Approach*. Department of Computer Engineering, Faculty of Engineering, Kasetsart University : Bangkok.
- [16] Peter F.Brown, Jennifer C.Lai, and Robert L.Mercer. 1991. *Aligning Sentences in Parallel Corpora*. IBM Thomas J. Watson Research Center, NY 10598.
- [17] Royal Institute. 1992. *Standard of Transliterated Word Encoding, the Royal Institute Edition* .
- [18] Soonkhlang Tassanawan. 2000. *Transliterated Word Encoding for Thai-English Cross-Language Retrieval with Neural Network Techniques*. Department of Computer Engineering, Graduate School, Chulalongkorn University : Bangkok.

- [19] Suwanvisat Prayut. 1998. *Transliterated Word Encoding for Thai-English Cross-Language Retrieval*. Department of Computer Engineering, Graduate School, Chulalongkorn University : Bangkok.
- [20] Tanapong Potipiti, Virach Sornlertlunvanich, and Thatsanee Charoenporn. 2000. *Towards Building a Corpus-based Dictionary for Non-word boundary Languages*. National Electronics and Computer Technology Center, Thailand.
- [21] Tomaz Erjavec. 2002. *Compiling and Using the IJS-ELAN Parallel Corpus*. Department of Intelligent Systems, Jozef Stefan Institutes, Slovenia.
- [22] University Lecturer. 1996. *Using Thai Language1*. Faculty of Liberal Arts, Thammasat University.
- [23] William Gregory Sakas. 2000. *Ambiguity and the Computational Feasibility of Syntax Acquisition*. Department of Computer Science, the City University of New York.