A Framework of 2-step Bilingual Alignment for SMT: in Case Study of Thai-English Translation

Prasert Luekhong[1], Taneth Ruangrajitpakorn[2],
Rattasit Sukhahuta[3], Thepchai Supnithi[2]

[1]College of Integrated Science and Technology, Rajamangala University of Technology Lanna, Chiang Mai, Thailand
prasert@rmutl.ac.th
[2]Language and Semantic Technology Laboratory, National Electronics and Computer Technology Center, Thailand
{taneth.rua, thepchai.supnithi}@nectec.or.th
[3]Computer Science Department, Faculty of Science, Chiang Mai University, Chiang Mai, Thai-land
rattasit.s@cmu.ac.th

Abstract

This paper presents a framework of a new word alignment process that can be used in an SMT development. The method was designed to include the quality of using dictionary as prior knowledge and the ability of co-occurrence to fill unknown words. The alignment method is split into two separated steps: firstly, the dictionary-based step to guarantee the accurate word-aligning and secondly, co-occurrence-based step to handle the unknown word. In the dictionary-based step, similarity score is exploited to effectively handle partial unknown words. By testing the proposed framework against the renowned GIZA, we applied an alignment model from both systems to MOSES for proving its usefulness in a practical machine translation usage and exploited a BLEU score as a measurement. The case study in this work focused on Thai to English translation. The testing results showed that the best setting of the proposed method can overrun the result of GIZA IBM model-4 by 2.09 BLEU points.

Keywords: Word Alignment, Machine Translation, Bilingual Dictionary, Prior Knowledge, Hybrid Approach

I. Introduction

A word alignment is a basic tool to identify the same concept of a word from two texts. It is used widely in several NLP applications, such as machine translation, question-answering, information extraction and summarization. For an application such as statistical machine translation (SMT), an alignment is the core process to signify the translation capability and the key to correctness of translation result. The renowned approach of word alignment in SMT,

GIZA - a statistical machine translation toolkit [1], exploits a co-occurrence based technique by using found evidence in a parallel corpus. This approach currently has been favoured by the simplicity, coverage of words based on given corpus, and acceptable accuracy for decades. Many MT applications and researches mainly applied this approach of word alignment and they can publicly be used in practice, for example Google translation [2] [3] and Kantan machine translation [4].

However, the incorrect translation results of SMT from several systems have been reported from [5] [6] [7] and the wrong aligned word pairs in training process caused most of them. Since the co-occurrence focuses mainly be the frequency of co-existing word from parallel sentences in the given corpora, there are some miss-leading situations to align inappropriate words, which are not a translation to each other but frequently occurred to each other. This issue becomes worse for the explicitly disputative pattern of language pairs such as English-Thai translation, which belongs to a different language family [8]. For example, there is no article, no inflection to express plurality and tense, and no word-order conversion to express questioning pattern in Thai while English sentences are rich with these words and patterns are rich in. Moreover, the polysemous words and words shorten in forms which are often used in Thai natural language often lead to wrong co-relate statistic. Thus, the translation result of such language pairs from SMT is always low in accuracy as reported in several previous researches [9] [10] [11] [12].

To evidence the aforementioned reports, GIZA was tested with 130,000 pairs of sentence corpus and aligned result of random sentence is exemplified in Fig. 1 to demonstrate the inaccurately aligned word by co-occurrence. From the Fig. 1, several assigned words are obviously incorrect.

For instance, Thai word "ได้ (literal: capable to)" is often co-existed with English word "in" since the Thai word is a word frequently used in many situations referring to 1) be capable to, 2) action of receiving, 3) indicate past tense and 4) use as the positive answer while the word "in" also has many usages in English language such as a part of the proverb. This causes the ambiguity among the co-existing words in parallel corpus and leads to incorrect alignment.

To improve the alignment of words, some works such as [13] [14] [15] apply the prior knowledge, such as dictionary data and WordNet, as a given rule set to narrow the possibility of alignable words. They may improve the quality of aligned words, but it traded off to the less quantity of words that can be aligned since it is impossible to include all used lexicons, such as named-entity, slang word, jargons, and emerging words, etc., in the language pair into a dictionary. Therefore, the coverage of this approach cannot practically be used in general domain and with a natural language input since the unknown words are the major issue to the system, and increasing the dictionary entry cannot be done without the burden and time-spent to

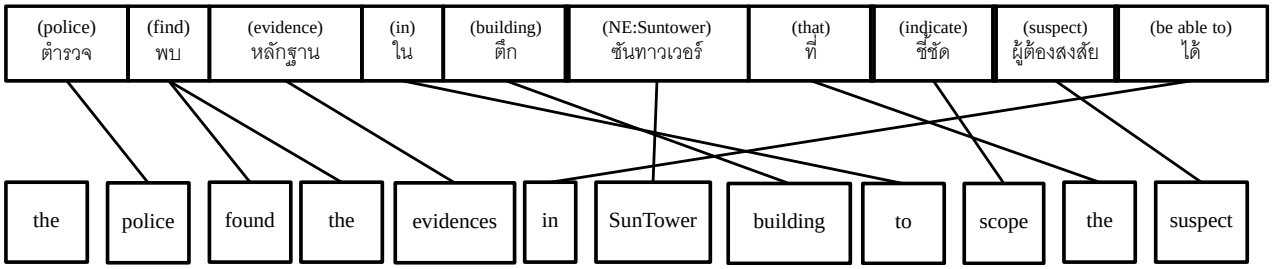

Fig. 1 An example alignment sentence from GIZA

linguists and dictionary administrator.

In this work, our research question is to find the method to obtain the quality of word alignment using bi-lingual dictionary as prior knowledge to word alignment process for machine translation application while the process remains the quantitative benefit from co-occurrence approach. Thus, we propose the two-step alignment framework that obtains the quality of dictionary based and the quantity from co-occurrence based approach.

The rest of the paper is organised as follows. Section 2 explains the design of the two-step alignment framework for MT including the example of each process. Section 3 gives the experiment setting and result. Section 4 discusses over the result and details of the framework in practical. Section 5 gives a conclusion of the paper and lists the related future work to improve the framework.

II. Related work

Since this work focuses on improving a word alignment for SMT by using a bilingual dictionary as prior knowledge, we review the existing researches related to the enhancing word alignment with a parallel dictionary. The main idea of a statistical word alignment is to find the parallel semantic of source and target word(s) in the bilingual corpus and annotate them together as a pair of translated word(s). The common method of a word alignment given in widely used tool, GIZA, is to find the alignment based on co-occurrence, but occasionally the alignment results are incorrect. A dictionary hence is used to give the process as a constraint for a more accurate result. The approaches using a bilingual dictionary in SMT are exemplified below.

Zhu & Chang [14] proposed a method to apply a bilingual dictionary as a part of training bilingual corpus. By adding a translated pair of words into a training set, they pointed out that the method would raise the probability score of the pair and result in better alignment. The idea is to adjust the existing weight of co-occurrence words to identify what should be defined as word pair and also make the words to be known within a training process. This method achieved an improvement of alignment result and was reported to outperform the baseline. In a summary, the proposed technique is easy to apply and has an obvious result, but it has a reverse effect on aligning low frequent words since it will reduce the weight of those rarely used words when they are needed. Moreover, the threshold of a number of adding the word pairs is difficult to be determined since a word in one language can have various translations and numbers of those pairs that should be added is hard to adjust properly and automatically, In addition, though this method can be flexible to cover of adding the words with their inflected form such as a plural form, this circumstance will also effect on adding more variety to the weight of possible aligned word and cause more confusion to weight of word pairs.

Y. Liu, Q. Liu, and S. Lin [16] presented an alternative algorithm for applying dictionary as a parameter to their log-linear models word alignment framework in 2005.

Their algorithm was modified from a baseline alignment of that time, IBM model 3, by adding parameters in an alignment process. These additional parameters were linguistic information including syntactic data of word and word translation. The dictionary was used to create an initial parameter and it was adjusted by frequency of the words together in manual aligned parallel corpus. By joining the new parameters with the existing parameter from the baseline, a quality of its alignment result significantly increased since it obtains more reliable constraints to decide the parallel words. This approach shows high potential, but it requires a richful resources such as a POS tagger for both languages, a sufficient amount of examples in manual aligned corpus, and a good bilingual dictionary. This approach will become unavailable for those low resource languages which lacks of aforementioned tools and corpora. Moreover, using linguistic data as parameter cannot assure a correct aligning result since a score of another parameter can make a confusion on deciding and ignore the correct alignment to wanted candidate.

Some works such as Ker and Chang [17] applied dictionary as a reference for aligning words without using parallel corpus as a training data. Ker and Chang's work only used a bilingual dictionary to align the words in a bilingual text. With dice coefficient measure, the score for each word align is calculated and the highest score path is chosen as an alignment. For those words which are contained in a dictionary, this approach works effectively, but this work suffered an issue of an unknown word since data in a dictionary cannot represent all the using words in natural language. In details, it is impossible to include all used lexicons in parallel corpus, such as named-entity, slang word, jargons, emerging words, etc., into a dictionary. Moreover, an ambiguity and a variation of a translation are another limitation of this approach because a natural language is likely to be used in a comparable meaning instead of a direct word-to-word translation.

In a summary, a bilingual dictionary was used in several parallel word alignments to improve their accuracy. It helps to enhance the aligned output by adding a reference of possible translation in word paring consideration. However, the limitation of previous works can be grouped as a limit number of lexical entries of a dictionary and the usage of it within a statistical model. The former causes an incompetence in quantity of word alignment because of an unknown word issue.

The latter cannot guarantee the wanted qualitative result since it mainly relies on probability score.

III. Two-step alignment framework for MT

In this work, the new alignment process for MT is proposed. The system employs the benefit of both co-occurrence and prior knowledge from a dictionary. For ease of observation, the system is designed to work separately as two steps. Fig. 2 illustrates the overview of the proposed framework. The first step is to align the word in a parallel sentence based on translations of a word given in a dictionary entry to guarantee the quality of alignment, and the second step is for handling the unaligned words by using co-occurrence to complete the alignment of the entire sentence. The second step can handle the numeral expression, proper-name, etc. The format of aligned output is designed to resemble the output of GIZA which can be used in other machine translation processes by Moses. The required inputs of the framework are bilingual dictionary and parallel corpus.

A. Data Preparation

As a resource for the system, data used in the framework should be pre-processed to meet the requirements.

a. Bilingual Dictionary

The dictionary for comparing the words in two different languages. The reference dictionary must contain the word base entry in the root form and its possible word translation, not the word description. To possible future usage, each entry should contain the headwords and translated words as a synonym set with the part of speech (POS). The idiom and proverb should be prevented to be included since the word matching processes can handle only word level in the current state.

b. Bilingual Corpus

The corpus is a collection of parallel sentences which will be used as the learning source for automatic word alignment. The word must be segmented in order to ascertain the boundary of lexicon for word based alignment. To prevent an unnecessary ambiguity, symbolic function markers such as full-stop and question mark are removed. Moreover, the articles also should be removed for the language pair of different language family since they are frequently shown in one language while another language does not have such function word.

B. Two-step Alignment

a. Step 1 # Dictionary-based Alignment

This process aims to map the words according to lexical entry given in the reference dictionary. The required inputs for this process are the reference dictionary and pre-processed parallel corpus. It consists of 3 modules as shown in Fig. 3.

i. Word to Metric Generation Module

To efficiently compare the words in both source and target sentence, the metric is generated. The source words are designed to position vertically while target words are placed horizontally.

ii. Dictionary Mapping Module

In order to find the appropriate translation of word from source sentence, the source word is looked up in the reference dictionary. It is possible for each source word that has several translatable words in target language. The variation of the translations is all listed to the word in order to compare them

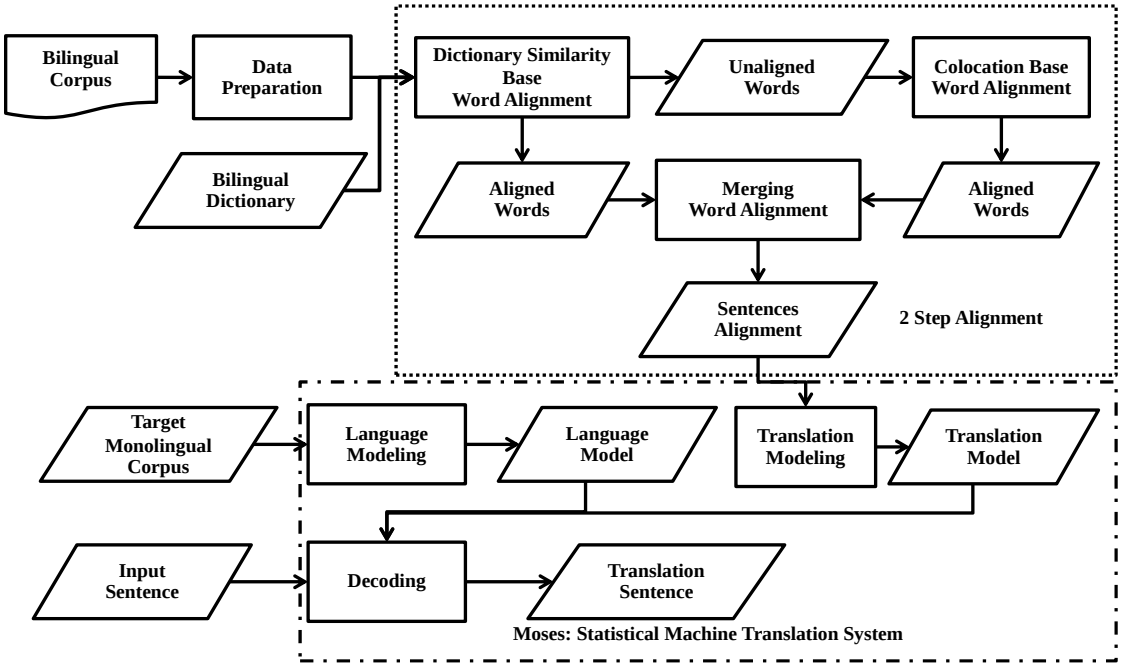

Fig. 2 An over view of the framework of 2-step alignment for SMT

with each of target word by the apparent likeliness.

iii. Similarity Score Calculation Module

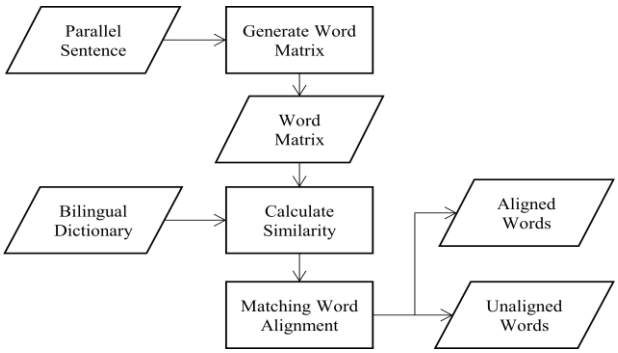

Fig.3 An overview of dictionary similarity-based word alignment

Upon using dictionary as reference for word translation, it is possible for the word in sentence to be in a different form to the dictionary form because of an inflection, a derivation, a different localising usage (British form and American form for example), and a typo. Similarity score (*SimScore*) is exploited to increase the chance to map the translation word in target language. The *SimScore* algorithm was published by Islam and Inkpen 2008 [8] and it is redefined to fit the framework as followings.

We proposed an algorithm combine dictionary and string similarity to matching lexical of source sentence with a words in target sentence [7] . The redefine of string similarity algorithm exposed by Islam and Inkpen 2008 [8] . The formal of our algorithm described as following

$Sim(s_i,D_i)$ [8] represents the string similarity between words. We exploit a merge of normalized longest common subsequence ($NLCS$), maximal consecutive longest common subsequence starting at character 1($NMCLCS_1$) and maximal consecutive longest common subsequence starting at any character n ($NMCLCS_n$). When $s_i$is a source word we use D to denote dictionary, which has a set of source entries $ds_1, \dots, ds_n$ and a set of target entries$dt_1 \dots, dt_n$. In the dictionary, there is $D = \{ds_i, dt_{i-to-j}\}$. The formulae are as follows:

$$v_1 = NLCS(s_i,D_i) = \frac{length(LCS(s_i,D_i))^2}{length(s_i)\times length(D_i)} \quad (1)$$

$$v_2 = NMCLCS_1(s_i,D_i) = \frac{length(MCLCS_1(s_i,D_i))^2}{length(s_i)\times length(D_i)} (2)$$

$$v_3 = NMCLCS_n(s_i,D_i) = \frac{length(MCLCS_n(s_i,D_i))^2}{length(s_i)\times length(D_i)} (3)$$

We use the weighted sum of these individual valuesv_1,v_2 and v_3 to determine string similarity score, where w_1, w_2, w_3 are weights and $w_1+w_2+w_3 = 1$. We set w value by use a linguistic knowledge as a prior knowledge then use EM algorithm to find a significance of each parameter by v . Therefore, the similarity of the two strings is:

$$Sim(s_i,D_i) = w_1v_1 + w_2v_2 + w_3v_3 \quad (4)$$

When we get a maximum similarity scores for each words in source and target sentence. We put the results into matrix. Then we used Hungarian algorithm propose by Kuhn 1955 [9]. Mills-tettey et al developed to dynamic Hungarian algorithm in 2007 [10]. We apply that to assign the best path in the matrix.

All of the SimScores of each possible translation are assigned in each metric slot, and the highest score will be selected as the translation of the source word.

iv. Best-in-Slot Selection Module

Once all the metric slots are filled with SimScore, it is possible that the values are the same in the horizontal row. To find the most suitable translation in each slot, Hungarian algorithm [9][10] are employed to find the best total score of all slots.

To ascertain the translation selection in terms of accuracy, the system is designed to be assignable of the prefer SimScore to limit the unwanted aligning. For example, if only the words exactly matched to the dictionary entry are wanted, the limit of SimScore must be set to 1 which will allow the system to align only the words with 1 SimScore. Furthermore, in case the limit of SimScore is set to 0.8, all the words with 0.8 or higher SimScore will be aligned. The rest of words which are not aligned in this process will later be used in co-occurrence based alignment in the second step.

b. Step 2 # Co-occurrence-based Alignment

From the step 1, if all of the words are not handled, there will be the words left undone from either source or target sentence or both. To fill the left over words, co-occurrence based approach, i.e. GIZA, will be exploited to align the rest by using the frequency and the statistic of co-existing. The chosen version of GIZA in this work is the GIZA with IBM model 4 [21] since this is the base-line version for phrase base MT [22] and hierarchical phrase base MT [23] suggested in workshop on statistical machine translation from 2006-2013 [24] for its best proficiency. GIZA is a freeware alignment toolkit utilizing co-occurrence given in a parallel corpus. However, the data sparseness from a parallel corpus often causes the word alignment model to not indicate associations between source and target words correctly [16] [25]. Hence, the already aligned words from Step#1 greatly helps GIZA to reduce the sparseness and makes it to work more effectively. For example from Fig. 5-B, the left over words are the word in the unbold box. From preliminary experiment, those words were aligned in more accurate than the alignment of the full text by GIZA as shown in Fig. 1 as for instance the word “in” (the forth English word) was aligned and will not cause the confusion in the co-occurrence.

With GIZA, all the leftover words will be aligned even though the word exists once or it is a named-entity. Moreover, it allows the one to many word alignments so the rest will be efficiently handled. Therefore, this step will increase the quantity of an alignment to fulfil the uncertain words left from the alignment based on a dictionary which represents for qualitative alignment. Once GIZA aligns the rest of the words, they will be merged with the output of Step#1.

i. Merging Word alignment

To complete the word alignment process, the output from both step 1 and step 2 must be integrated into one completed alignment. The output format of this process is designed to compatible with MOSES [21]. An algorithm to merge an output from step 1 and step 2 are described in fig.5.

```
# Merging Words Alignment Algorithm
DO
String C= READ LINE Colocation Word Alignment
String D = READ LINE Dictionary Similarity Word Alignment
Hash Table CW = Split String C to words
Hash Table DW = Split String D to words
FOR i =1 to DW.size
FOR j =1 to CW.size
IF DW.key[i]=CW.key[j]
SET Value in DW[i] = Value in CW[j]
END FOR
END FOR
WRITE DW to File
UNTIL end of file
```

Fig. 5 An algorithm to merging 2 word alignments.

IV. Experiment

To prove the potential of the proposed framework, an experiment was set up and the result was compared to GIZA, the widely used word alignment in SMT development. Furthermore, the output aligned sentence was used to train machine translation implement by using MOSES to see the impact of the proposed method in the machine translation practical environment.

A. Experiment Setting

The bilingual corpus used in this experiment contains Thai-English parallel sentences gathered from two sources: BTEC (Basic Travel Expression Corpus) [26] and HIT London Olympic Corpus [27]. The total number of the sentence pairs is 149,000 sentences while it randomly separated into three sets. First is a training set containing 134,000 sentences. Second contains 1,000 sentences preserving as a development set that is for tuning the weights in the translation process. The rest of the sentences is for testing the SMT as testing sentence set that acquires 14,000 sentences. The reference dictionary developed by integrating of Lexitron [28] [29], opened Thai-English bilingual dictionary containing 40,850 lexical entries and dictionary for RBMT [30]. Pre-process, as mentioned in data preparation section, was applied to both sources.

The translation result training by GIZA alignment using IBM model-4 was assigned as a baseline to compare with the proposed method. Since the proposed system designed for using in Thai to English SMT application, we choose MOSES [21] for applying both alignments to get translation results for comparison their capability. However, in MOSES, a common method for creating word alignment models [31] requires both of a Thai to English alignment and English to Thai alignment. Therefore, we need to prepare these two set of alignments by both systems. In the proposed method, a SimScore has to be set for the Step#1, and we set a SimScore using in this experiment as 0.9 and 0.7 for Thai to English and English to Thai respectively as these scores produced the best result of a preliminary experiment. The BLEU [32] [33] point from the translation result used as a translation accuracy measurement.

B. Experiment Result

From the setting, the BLEU point results from each set up and base-line are shown in Table 1.

Table 1 The comparation results of our framework and baseline Thai to English machine translation

Test set		**Average BLEU score**		
Sentence length	**Amount**	**Baseline**	**Baseline +Dictionary**	**2Step**
1-5	4941	44.17	44.27*(+0.07)*	47.81*(+3.64)*
6-10	5777	28.47	28.54*(+0.05)*	29.44*(+0.97)*
11-15	2004	22.53	22.58*(+0.05)*	24.23*(+1.70)*
16-20	733	19.08	19.13*(+0.04)*	20.10*(+1.02)*
21-30	472	15.93	15.97*(+0.04)*	17.16*(+1.23)*
31-50	73	16.27	16.31*(+0.04)*	16.36*(+0.08)*
total	14000	32.18	32.26*(+0.08)*	**34.27*(+2.09)***

From the result, the proposed system against a renowned existing parallel word alignment, GIZA, we applied an alignment model from both systems to MOSES for proving its usefulness in a practical machine translation usage and exploited a BLEU score as a measurement. The case study in this work focused on Thai to English translation. Our framework shows potential over the baseline and baseline with dictionary for having better BLEU point for 2.09 and 2.01 points respectively in the overall translation result.

V. Conclusion

This paper presents a new method to develop a parallel word alignment for using in statistical-based machine translation. The main idea relies on the separation of the process into two steps. The first step is to consult the mapping with a dictionary entry to assure the quality of word alignment. To enhance a matching the minor typo and word inflections, the similarity score calculation to possibly match a dictionary entry to a practical word which perchance slightly varies in an appearance. The second step is designed to handle the unknown words left from the first step according to co-existing words in a parallel text to complete the alignment in terms of a quantity based on existing evidence. By testing the proposed system against a renowned existing parallel word alignment, GIZA, we applied an alignment model from both systems to MOSES for proving its usefulness in a practical machine translation usage and exploited a BLEU score as a measurement. The case study in this work focused on Thai to English translation. The result showed that the proposed system works better against the baseline as obtaining the higher BLEU point for 2.09 points from the translation outputs.

To improve the proposed system, we plan to include a phrasal bilingual dictionary to cover a proverb usage in first step matching for preventing confusion from a comparable sentence. Moreover, we plan to implement a method to handle language diversity from across language typology by using predefined rules to capture grammatical expressions in a sentence and transform them into a universal tag for reducing ambiguity. Last, an extended method to capture similar words

will be researched for reducing the unknown word issue in the quality based alignment step.

References

[1] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[2] D. Wu and D. He, "A Study of Query Translation Using Google Machine Translation System," in *2010 International Conference on Computational Intelligence and Software Engineering*, 2010, pp. 1–4.

[3] "Google translation." [Online]. Available: https://translate.google.com/. [Accessed: 29-Jul-2014].

[4] "Kantan Machine Translation Services." [Online]. Available: http://www.kantanmt.com. [Accessed: 29-Jul-2014].

[5] F. J. Och and H. Ney, "The Alignment Template Approach to Statistical Machine Translation," *Comput. Linguist.*, vol. 30, no. 4, pp. 417–449, Dec. 2004.

[6] S. Kumar, F. F. Och, and W. Macherey, "Improving Word Alignment with Bridge Languages," *EMNLP-CoNLL*, no. June, pp. 42–50, 2007.

[7] W. Hua and W. Haifeng, "Improving statistical word alignment with a rule-based machine translation system," in *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, 2004, no. 1, p. 29–es.

[8] J.-P. Koenig and N. Muansuwan, "The Syntax of Aspect in Thai," *Nat. Lang. Linguist. Theory*, vol. 23, no. 2, pp. 335–380, May 2005.

[9] C. Nusai, Y. Suzuki, and H. Yamazaki, "Estimating Word Translation Probabilities for Thai–English Machine Translation using EM Algorithm," *Int. J. Comput. Intell.*, vol. 4, no. 3, 2008.

[10] N. Labutsri, R. Chamchong, R. Booth, and A. Rodtook, "English Syntactic Reordering for English-Thai Phrase-Based Statistical Machine Translation," in *6th International Joint Conference on Computer Science and Sofrware Engineering (JCSSE 2009)*, 2009.

[11] P. Luekhong, R. Sukhahuta, P. Porkaew, T. Ruangrajitpakorn, T. Supnithi, and R. Sukhauta, "A Comparative Study on Applying Hierarchical Phrase-based and Phrase-based on Thai-Chinese Translation," in *International Conference on Knowledge, Information and Creativity Support Systems*, 2012, no. Ldc, pp. 126–133.

[12] P. Nakov and T. N. Hwee, "Improved statistical machine translation for resource-poor languages using related resource-rich languages," in *Methods in Natural Language Processing: Volume 3*, 2009, no. August, pp. 1358–1367.

[13] D. Tufis, R. Ion, and N. Ide, "Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets," *Proc. 20th Int. Conf. Comput. Linguist.*, p. 7, 2005.

[14] D. Zhu and B. Chang, "Bootstrapping word alignment by automatically generated bilingual dictionary," in *2008 International Conference on Natural Language Processing and Knowledge Engineering*, 2008, pp. 1–7.

[15] P. Luekhong, T. Ruangrajitpakorn, T. Supnithi, and R. Sukhahuta, "Pooja : Similarity-based Bilingual Word Alignment Framework for SMT," in *Proceedings of the 10th International Symposium on Natural Language Processing, Phuket, Thailand*, 2013.

[16] Y. Liu, Q. Liu, and S. Lin, "Log-linear models for word alignment," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 459–466.

[17] S. J. Ker and J. S. Chang, "A Class-based Approach to Word Alignment," *Comput. Linguist.*, vol. 23, no. 2, pp. 313–343, 1997.

[18] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans. Knowl. Discov. Data*, vol. 2, no. 2, pp. 1–25, Jul. 2008.

[19] H. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logist. Q.*, vol. 2, no. 1–2, pp. 83–97, 1955.

[20] G. A. Mills-tettey, A. Stentz, and M. B. Dias, "The Dynamic Hungarian Algorithm for the Assignment Problem with Changing Costs," 2007.

[21] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, N. Bertoldi, B. Cowan, C. Moran, C. Dyer, A. Constantin, E. Herbst, H. Hoang, and A. Birch, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 2007, no. June, pp. 177–180.

[22] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2010.

[23] D. Chiang, "Hierarchical Phrase-Based Translation," *Comput. Linguist.*, vol. 33, no. 2, pp. 201–228, Jun. 2007.

[24] "ACL 2013 Eighth Workshop on Statistical Machine Translation." [Online]. Available: http://statmt.org/wmt13/. [Accessed: 29-Jul-2014].

[25] C. Mi, Y. Yang, X. Zhou, X. Li, and T. Osman, "Co-occurrence Degree Based Word Alignment: A Case Study on Uyghur-Chinese," *Chinese Comput. Linguist. Nat. Lang. Process. Based Nat. Annot. Big Data*, pp. 259–268, 2014.

[26] "BTEC Task | International Workshop on Spoken Language Translation." [Online]. Available: http://iwslt2010.fbk.eu/node/32. [Accessed: 27-May-2012].

[27] M. Yang, H. Jiang, and T. Zhao, "Construct trilingual parallel corpus on demand," *Chinese Spok. Lang. Process.*, pp. 760–767, 2006.

[28] K. Trakultaweekoon, P. Porkaew, and T. Supnithi, "LEXiTRON Vocabulary Suggestion System with Recommendation and Vote Mechanism," in *SNLP2007 The Seventh International Symposium on Natural Language Processing*, 2007, pp. 43–48.

[29] "LEXiTRON:Thai-English Electronic Dictionary." [Online]. Available: http://lexitron.nectec.or.th/2009_1/. [Accessed: 10-May-2013].

[30] T. Ruangrajitpakorn and and T. S. Wasan. na Chai, Prachya Boonkwan, Montika Boriboon, "The Design of Lexical Information for Thai to English MT," in *Proceeding of SNLP 2007*, 2007.

[31] F. Och and H. Ney, "A comparison of alignment models for statistical machine translation," in *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 2000, pp. 1086–1090.

[32] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU : a Method for Automatic Evaluation of Machine Translation," *Comput. Linguist.*, no. July, pp. 311–318, 2002.

[33] N. Madnani, "iBLEU: Interactively debugging and scoring statistical machine translation systems," in *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, 2011.