Derivative Sentence Breaking for Moore Alignment

Glenn Slayden and Elias Luqman

University of Washington, Seattle

{gslayden,eluqman}@u.washington.edu

Abstract

By describing a method for efficient sentence alignment of bilingual corpora which requires no language-specific knowledge, Moore (2003) established an important new standard for this critical NLP task. However, the method still presumes sentence-broken input, posing considerable difficulties for languages with ambiguous sentence delineation. Taking Thai as our test case, we show that, when one of the languages in Moore alignment is sentence-ambiguous, deriving its sentence breaks according to the lengths of sentences in a parallel document can surpass the performance of standalone statistical sentence breaking.

1. Introduction

Sentence-aligned bilingual corpora are a staple of statistical machine translation (SMT) and methods for obtaining quality, aligned bitexts have received considerable research attention. SMT research has consolidated around the notions of the *sentence* and *word* as fundamental units of translation. This consensus has certainly fostered the development of models which are powerfully generalized—except to the extent that they take these concepts to be cross-linguistically deterministic. As the field matures, languages for which these assumptions are challenged receive increased attention as researchers aim to relieve them from the disadvantage of awkward accommodation to unnatural processing conventions.

Thai is one such language, since it uses space neither to distinguish syllables from words or affixes, nor to unambiguously signal sentence boundaries. Where these features interact with SMT output, as in the need for re-spacing Thai output according to Thai convention, solutions are relatively straightforward. More troublesome is the requirement that Thai text be presented to SMT systems in sentence-broken form, since "sentences" are neither unambiguously marked nor perhaps even linguistically imperative in the language.

In order to integrate Thai with contemporary SMT systems, researchers have investigated Thai sentence breaking (SB), which focuses on the task of classifying the spaces that appear in Thai text as either sentence-breaking or non-sentence-breaking. By viewing the problem of integrating Thai into SMT systems as a problem of Thai SB, however, we believe an important opportunity to capitalize on—rather than penalize—the language's flexibility has been missed. When considering the next step in an SMT training pipeline—for example, sentence alignment as described in Moore (2003), a method we examine in Section 4—it seems wise to accomplish Thai SB in a way that will best achieve the larger goal of better training. This is especially motivated when one considers the relatively high error rates of statistical SB: when error is inevitable, it is sensible to deploy it in a way that reduces error in the overall task.

We compare three methods for providing sentence-broken Thai input to Moore's alignment algorithm. First, as a control case, we establish that unprocessed Thai fragments are ineffective for SMT training. Then, we describe a simple length-based method that derives Thai sentence breaks from its English pair document. After another control experiment establishes that these results still require alignment, we show that our method, in combination with Moore's, outperforms a standalone maximum entropy model in BLEU evaluation, alignment running time, and out-of-vocabulary error.

This result is especially relevant when we reflect on the likely reason for the poor performance of the statistical model: dissonance between the training and application domains. Our new heuristic has no statistical model or training, so it shares the advantage of domain insensitivity with Moore's work. We also preserve the language agnosticism of Moore's work, with the caveat that one of the two languages be able to supply deterministic sentence breaks.

2. Thai Sentence Breaking

Thai is written without sentence-end punctuation, but a space character is always present between sentences. There is generally no space between words, but a space character may appear within a sentence according to linguistic or prescriptive orthographic motivation (Wathabunditkul 2003). These aspects disqualify SB methods used for many languages. Thai SB has therefore been regarded as the task of classifying each space that appears in a Thai source text as either sentence-breaking (**sb**) or non-sentence-breaking (**nsb**).

Aroonmanakun (2007) discusses Thai word- and sentence-breaking in more detail, characterizing the latter as "fuzzy." With regard to sentence breaking, he enlists Thai natives for a study on manual sentence breaking and concludes that, even for native speakers, there is little consensus on how to designate sentences in Thai.

Mittrapiyanuruk and Sornlertlamvanich (2000), define part-of-speech (POS) tags for **sb** and **nsb** and train a trigram model over a POS-annotated corpus. At runtime, they use the Viterbi algorithm to select the POS sequence with the highest probability, from which the corresponding space type is read back. Charoenpornsawat and Sornlertlamvanich (2001) apply Winnow, a multiplicative trigger threshold classifier, to the problem. Slayden et al. (2010) use a four-token window of Thai lemmas, plus additional categorical features, to train a maximum entropy (maxent) classifier. In Table 3, we compare quantitative evaluation of these three systems with evaluation of a maxent SB system we developed as a baseline for this research.

All of these systems aim to assign Thai sentence breaks *in vacuo*, making these approaches appropriate for use in monolingual applications, such as runtime SB for Thai-English SMT. While this strength of standalone models is wasted on SMT training—where parallel text is at hand—its cost in error rate is still borne. To the extent that SB error rates result in source-target misalignment, SMT for sentence-ambiguous languages is subject to a specific extra performance penalty.

These observations suggest that better SMT training might be achieved by capitalizing on the availability of parallel text during SB. In particular, sentence breaks for Thai might be derived by reference to the sentence breaks of the pair language document, which can be taken as authoritative. This is the focus of this research. We describe a method for deriving sentence breaks from parallel pair documents in bitext corpora.

3. Methodology

Our study corpus consists of 151 high-quality literary Thai bitexts downloaded from *wanakam.com*. We decoded the TIS-620 (8-bit Thai) character encoding, and cleaned and tokenized the documents, arriving at the statistics shown in Table 1. Although the derivative side of our method (Thai, in this case) is not language-specific, we do not use the French documents in this study in order to control for pair-language (non-Thai) SB consistency.

	Thai	English	French
Documents	188	151	37
Sentences	*subject work*	48,736	11,519
Words	1,225,761	895,429	164,490
HTML MB	8.17	8.23	1.71

Table 1. Study Corpus Size

In addition to this corpus, we prepared a set of reference alignments for use in BLEU evaluation. This is a literary document of 197 Thai sentences (2,357 words), where reference word breaks and reference English alignments were manually assigned according to criteria set forth in Slayden (2010).

We identified the English sentences in our corpora. For this study, these are taken as given. Next, we identified all Thai *fragments*: sequences of Thai text which occur separated by space in the source. With the exception that we did not designate fragments by breaking text inside of directional double-quotation marks, these fragments are internally space-free. The corpus contains 110,119 fragments. The minimal statement of the alignment task is to assign zero or more of these fragments to each English sentence. For this study, we add the further constraint of monotone sequence across both source and target.

We evaluate three configurations in this paper. In each configuration, we use Moore's alignment method to align a varying number of Thai "sentences" to the 48,736 English sentences in our study corpus.

The first configuration simply submits all 110,119 Thai fragments as sentences. The second configuration assigns Thai sentence breaks according to a monolingual maxent SB model that we developed. This model is described in Section 5. Next, we evaluate our new method, which is described in Section 6. Results from these experiments are presented in Section 7. First, however, since it is so central to our work, we review Moore's important method for efficiently aligning sentences in bilingual corpora.

4. Moore Sentence Alignment

Moore (2003) describes a method for aligning sentences in bilingual corpora which does not require an external lexicon, corpus-dependent anchor points, or prior paragraph alignments. A hybrid of earlier sentence-length-based methods and word-correspondence-based methods, this domain-independent method requires no knowledge of the paired languages beyond their division into words and sentences.

The core of the method is a generative probabilistic model for predicting the lengths of sentences composing sequences of minimal alignment segments, or "beads," in which sentences align *1-to-1*, *1-to-2*, *2-to-1*, *1-to-0*, or *0-to-1*. Each bead in the sequence is generated according to a fixed probability distribution over bead types, and for each type of bead there is a submodel that generates the lengths of the sentences composing the bead. Moore's model assumes that the length l_t of the sentences of the target language varies according to a Poisson distribution with mean $l_s r$, where

$$P(l_t|l_s) = \frac{exp(-l_s r)\,(l_s r)^{l_t}}{l_t!}$$

and

$$r = \frac{c_s \sum l_t}{c_t \sum l_s}.$$

Moore finds that the Poisson distribution fits empirically tested data better than the best-fitting Gaussian distribution, which is used in many other aligners. Moreover, the Poisson distribution is easier and faster to estimate because it has no hidden parameters: it does not require any iterative parameter re-estimation methods.

Moore's aligner finds the best alignment points by using a dynamic programming (DP) search. A forward-backward probability computation is used to find the highest-probability *1-to-1* beads to use in training a word-translation model. Moore employs a novel method of pruning to limit the DP search space in the forward pass. This method assumes that valid alignment points lie near the main diagonal of a spatial matrix, and confines its search to a band around the main diagonal. The band is iteratively expanded until the discovered alignment does not approach the edges of this band.

Next, the highest probability *1-to-1* beads from the initial alignment are used to train a slightly modified version of IBM Model 1 (Brown et al. 1991). Moore omits translation probabilities for rare words, and discards translation pairs which are less probable than random choice; these modifications reduce the size of the translation model by over 90%. A final re-alignment pass integrates the initial sentence-length-based model with the lexical correspondences from IBM Model 1 according to a generative model.

Moore cites precision and recall error figures of 0.030% and 0.081% respectively (compared to hand-aligned figures of 0.010% and 0.061%) for *1-to-1* alignments when a probability threshold of 0.9 was used. On a second dataset, Moore's system identified errors in hand alignments, posting precision and recall error rates of 0.006% and 0.340% (versus hand-aligned 0.029% and 0.570%). He also demonstrates how use of the translation model helps with difficult input data by deleting blocks of sentences from one side of the corpus.

Overall, Moore's use of an efficient length-based model to guide the activity of a more accurate but costlier word-correspondence-based model results in a system that is fast, accurate and domain-independent. Since this method requires sentence-broken input, we next turn to a discussion of the configurations we evaluated for sentence-breaking Thai text.

5. Maxent Model

As an evaluation baseline for current practice in Thai SMT, we used Mallet (McCallum 2002) to build a maximum entropy SB model with the features described in Slayden et al. (2010). The model in *ibid.* included commercially available corpora and was trained on 361,802 Thai sentences. Our model was trained on 64,005 Thai sentences gathered from the public sources shown in Table 2.

Because the performance of statistical models is sensitive to the training domain, we wished to include literary texts in the SB model. Our study corpus itself is not annotated with sentence breaks, so it cannot be included in the training set. ORCHID is a POS-tagged corpus of academic engineering papers from a technology conference. These are out-of-domain for our study corpus. The LEXiTRON (NECTEC 2003) and *thai-language.com* (Slayden 2009) corpora consist of sample sentences from bilingual dictionaries and many of these sentences are literary. Closest to our application domain are the 3,613 sentences which comprise six short stories downloaded from SEAlang (CRCL 2006).

Source	Sentences	%
LEXiTRON	33,378	52.2
ORCHID	23,176	36.2
SEAlang literary bitexts	3,613	5.6
thai-language.com	3,838	6.0
total	64,005	100.0

Table 2. Monolingual Thai training corpora for our maximum-entropy sentence breaker

As noted, our maxent features are based on Slayden et al. (2010). This approach uses a combination of categorical and synthetic features to characterize the proximal context of each space that occurs in Thai text as **sb** or **nsb**. Our intention was for this maxent SB to serve as a proxy for current practice in Thai SMT. Accordingly, we evaluated it against published results for standalone Thai SB. Table 3 shows results for our new system alongside the results for the systems mentioned in Section 2: POS Trigram (Mittrapiyanuruk and Sornlertlamvanich 2000), Winnow (Charoenpornsawat and Sornlertlamvanich 2001), and MSR-MT (Slayden et al. 2010).

Method	POS trigram	Winnow	MSR-MT	This Work
nsb-precision	90.27	91.48	93.18	89.92
nsb-recall	87.18	97.56	94.41	89.13
sb-precision	74.35	92.69	86.21	84.64
sb-recall	79.82	77.27	83.50	85.71
space-correct	85.26	89.13	91.19	87.72
false-break	8.75	1.74	3.94	6.4
F	82.90	89.75	89.32	87.35

Table 3. Thai Sentence Breaking Results

Sentence-breaking our study corpus with this SB, we obtained 80,250 Thai sentences. This figure represents over-breaking of 64.7% relative to the number of English sentences. We believe that the discrepancy between this figure and the characterized performance of the model may involve domain sensitivity. With 36.2% of its training sentences being jargon-laden engineering papers, it is not surprising that this model performs well when testing the same material. The texts in our study corpus, on the other hand, generally lack the categorical features described in *ibid.* and do not fare as well. Since our methodology warranted a system that was representative of current general-domain systems, we decided to proceed with these results, while noting that our results are conditioned upon the considerable domain sensitivity of statistical modeling techniques. That is, because the high empirical error of our SB model is justified by our study methodology, and because this error appears to overwhelm the error rates reported in Table 3, we deemed this SB to be sufficiently representative of current standalone Thai SB for the purposes of this study.

6. LA-1

Gale and Church (1993) establish the efficacy and reasonable accuracy of length-based sentence alignment. We extend the idea to sentence breaking with our length-based derivative sentence breaking heuristic, LA-1.

In this method, we calculate a scaling factor between the length, in characters, of the Thai and English source documents, and this factor is applied to all subsequent position comparisons. Enumerating the Thai fragments in order, they are assigned to English sentences monotonically. When adding a Thai fragment (source) would exceed the length of the current English sentence (target), the target sentence is advanced, if possible, and the fragment is then assigned without considering the size of the new target. The unused portion of the previous target, if any, is added to the available space of the new target in order to keep the approximate positions synchronized between the source and target.

The algorithm continues for all source fragments. Although it is possible for the algorithm to assign no Thai fragments to a set of trailing English sentences, in practice this did not occur, because Thai fragments greatly outnumber English sentences. Therefore, in this study, LA-1 always produced exactly one Thai “sentence” for each English sentence, and our algorithm presents 48,736 sentences to the alignment step. In fact, this characteristic of LA-1 allows us to skip the alignment step altogether and submit LA-1 “alignments” directly to GIZA++. We evaluate this, and our other experimental configurations in the next section.

7. Evaluation and Results

For our three configurations, we studied BLEU (Papinei et al 2002), alignment run time, and out-of-vocabulary (OOV) error. In each experiment, sentence alignments produced by Moore's algorithm (described in Section 4) are used to build a Thai-to-English phrasal SMT system with GIZA++ (Och and Ney 2003). Model parameters are shown in Table 4. For each experiment, the trigram target LM is built over the entire corpus, prior to pruning sentences of 40 or more tokens.

Model 1 Iterations	5
Model 2 Iterations	0
HMM Iterations	5
Model 3 Iterations	5
Model 4 Iterations	5
Phrase limit	4
Alignment	grow-diag-final-and
Reordering	msd-bidirectional-fe
Target LM	trigram

Table 4 Phrasal SMT experiment parameters

Because Thai does not use space between words, tokenization of Thai for SMT processing has certain complexities. For example, in the maxent and LA-1 experiments, where multiple Thai fragments must be adjoined to form putative Thai sentences, a special "spc" token is introduced. This is required because the SMT tools usurp the orthographic space character for the function of delimiting word tokens. The gold standard document is prepared in the same way and decoded with Moses (Hoang et al. 2007). BLEU is evaluated with the `multi-bleu.perl` script.

Results are shown in Table 5. By making use of length information in a parallel text, our simple LA-1 algorithm outperforms, for all of our study categories, our standalone monolingual statistical SB model which approximates current practice in Thai SMT training. A control experiment with raw Thai fragments disqualifies Moore alignment as the source of this improvement.

Moore's alignment utility allows alignments to be discarded based on a cut-off threshold for their first-pass forward/backward probabilities. We ran each experiment with threshold values of 0.85 and 0.90 and confirmed a consistent reduction in alignment count and increase in BLEU with the latter setting. As expected, OOV rates also increase as the number of training instances is reduced. At threshold value 0.90, our method obtains a BLEU of 5.01 versus 4.63 for maxent and 2.19 for unaligned fragments. Respective scores at threshold value 0.85 are 4.77, 4.13, and 1.65.

As noted in Section 6, by generating one Thai "sentence" per English sentence, LA-1 permits Moore alignment to be skipped. In this experiment, we obtained a BLEU score of 1.29, the lowest in our study. This suggests that, although LA-1 produces the same number of source and target sentences, these monotonic pairs are not necessarily always correctly aligned, and our algorithm does benefit from alignment according to Moore's hybrid technique. By construction, Moore's algorithm completes very quickly when correcting short-distance alignment errors such as these, however, so a key advantage of LA-1 is the quick alignment it permits.

A caveat that is apparent from these results is that our method's BLEU scores are generated on a larger training set than the maxent method was able to elicit from Moore's aligner, despite both configurations starting from the same corpus. This difference likely contributes to our minimal OOV rates as well. Although additional research is required to determine why Moore's generative model so strongly prefers our inputs, we speculate that our pre-matched sentence lengths are alluring to its length-based factor. As for maxent, of the 80,250 Thai sentences it assigns, 89.1% and 87.6% (at thresholds .90 and .85, respectively) cannot be aligned at all. Because max-

Method	Thai "Sentences"	Moore alignment			Train sentences	BLEU	OOV%
		time	threshold	sentences			
Raw Fragments	110,707	2:02:45	0.90	7,900	7,426	2.19	17.05
			0.85	10,240	9,592	1.65	14.08
Maxent	80,250	1:26:15	0.90	8,714	8,314	4.63	15.43
			0.85	9,960	9,489	4.13	14.15
LA-1	48,736	0:27:18	0.90	21,309	16,414	5.01	6.31
			0.85	21,991	17,031	4.77	5.12
		→	→	→	37,442	1.29	4.71

Table 5. Comparison of methods of assigning Thai sentence breaks for subsequent alignment with 48,736 English sentences and evaluation against a held-out gold standard of 197 sentences.

ent achieves impressive performance despite this handicap, it is likely that it would achieve significantly better results on a larger corpus, and we leave this point for later examination.

8. Conclusion and Future Work

We presented a method for assigning sentence breaks in sentence-flexible languages when parallel texts are at hand, such as in SMT preprocessing. Our method extends Moore's alignment technique by relaxing the requirement that both language pairs support deterministic sentence breaking, while preserving its appealing language and domain agnosticism.

To evaluate our approach, we built a monolingual statistical SB; provided quantitative evidence for its validity as a proxy for current practice in Thai SMT systems; obtained an empirical result that suggests that such models can be domain sensitive; and compared the proxy to our method, showing that our method is superior in end-to-end BLEU, alignment time, and OOV.

This preliminary work presents much opportunity for further development. Our SB model is small and may not accurately represent the sophistication of larger standalone sentence breakers. Derivative SB techniques should be evaluated against full models used by production SMT systems. We also report BLEU scores on phrasal SMT systems built over our tiny test corpus only. Our held-out test set is also miniscule. These scores may not reflect the complexities and parameter selections of large-scale systems, so further evaluation is warranted.

As suggested by the name of our algorithm, we originally hoped to evaluate multiple derivative sentence breaking heuristics. In fact, the LA-1 algorithm was created as a temporary placeholder for what we anticipated would be a number of candidate techniques. As our work progressed, however, we saw more value in focusing on evaluation against current practice, and we left refinement of the derivative approach for future work. Inadvertently, the strong performance of such a simplistic algorithm serves to particularly highlight the great potential of derivative approaches to sentence breaking.

References

W. Aroonmanakun. 2007. Thoughts on Word and Sentence Segmentation in Thai. *Proceedings of the Seventh International Symposium on Natural Language Processing,* Pattaya, Thailand, 85-90.

P. F Brown, J.C. Lai, and R. L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association For Computational Linguistics* (Berkeley, California, June 18 - 21, 1991). Annual Meeting of the ACL. 169-176. Morristown NJ: Association for Computational Linguistics, Morristown.

Thatsanee Charoenporn, Virach Sornlertlamvanich, and Hitoshi Isahara. 1997. *Building A Thai Part-Of-Speech Tagged Corpus (ORCHID).*

Paisarn Charoenpornsawat and Virach Sornlertlamvanich. 2001. Automatic sentence break disambiguation for Thai. In *International Conference on Computer Processing of Oriental Languages (ICCPOL)*, 231-235.

CRCL, Center for Research in Computational Linguistics. 2006. *SEAlang: Thai On-Line Library*. *http://sealang.net/thai/bitext.htm*

William A. Gale and Kenneth W. Church. 1993. *A Program for Aligning Sentences in Bilingual Corpora.* Computational Linguistics 19 (1): 75–102.

Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran and Ondřej Bojar. 2007. *Moses: Open source toolkit for statistical machine translation*.

Andrew Kachites McCallum. 2002. *MALLET: A Machine Learning for Language Toolkit*. *http://mallet.cs.umass.edu*

P. Mittrapiyanuruk and V. Sornlertlamvanich. 2000. The Automatic Thai Sentence Extraction. In *Proceedings of the Fourth Symposium on Natural Language Processing*, 23-28.

Robert C. Moore. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Machine Translation: From Research to Real Users (Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California)*, 135-244. Heidelberg: Springer-Verlag.

NECTEC. 2003. *LEXiTRON Thai-English Bilingual Dictionary*. Bangkok: National Electronics and Computer Technology Center, National Science and Technology

Development Agency (NSTDA), Ministry of Science and Technology. *http://nectec.or.th/*

Franz Joseph Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51

Kishor Papinei, Sailm Roukos, Todd Ward, and Wei-jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, July 2002*. 311-318.

Glenn Slayden, Mei-Yuh Hwang, and Lee Schwartz. 2010. Thai Sentence-Breaking for Large-Scale SMT. Submitted to *WSSANLP at COLING-2010*.

Glenn Slayden. 2009. *Thai-English Dictionary*. Seattle: thai-language.com. *http://www.thai-langauge.com*.

Glenn Slayden. 2010. An Information Structure Annotation of Thai Narrative Fiction. *University of Washington Working Papers in Linguistics* 28. Seattle: University of Washington. *http://www.thai-language.com/resources/slayden-information-structure.pdf*

Suphawut Wathabunditkul. 2003. Spacing in the Thai Language. *http://www.thai-language.com/ref/spacing*