A Purely Monotonic Approach to Machine Translation for Similar Languages

Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita
Multilingual Translation Lab.,
Universal Communication Research Institute,
National Institute of Information and Communications Technology,
Kyoto, Japan
{*yekyawthu, andrew.finch, eiichiro.sumita*}*@nict.go.jp*

Yoshinori Sagisaka
GITI/Dept. of Applied Mathematics Language &
Speech Science Research Lab.,
Waseda University,
Tokyo, Japan
ysagisaka@gmail.com

***Abstract*—This paper investigates the effect of taking a strictly monotonic approach to machine translation for a restricted set of suitable language pairs. We studied the effect of decoding monotonically for a set of language pairs which has similar word order characteristics and found that for some language pairs - namely language pairs where both languages are in SOV order - there was almost no difference in machine translation quality. The results of this experiment motivated the extension of the monotonic approach into the alignment stage of the training. We used a Bayesian non-parametric aligner that has been shown to outperform GIZA++ in combination with the $grow$-$diag$-$final$-and heuristic on transliteration data. Our results show that the monotonic aligner was able to match the performance of the GIZA++ baseline, and gains in translation performance were obtained by integrating both aligners into the systems.**

***Keywords*-monotonic decoding; machine translation; bilingual alignment;**

I. Introduction

Reordering is used in phrase-based machine translation (PBPT) to handle the long distance movements of translations of words and phrases during translation. The process of reordering, possibly above all others in the machine translation process adds considerable complexity to the task, and it affects both the training process (the word alignment must be performed non-monotonically) and the decoding process (bilingual phrase pairs that generate the target from left to right can come from positions in any order on the source side). The resulting increase in complexity of has led to pre- and post-ordering approaches which attempt to mitigate the issues associated with the complexity of reordering by handling it separately, leaving a predominantly monotonic process of translation.

In the framework of PBMT, local reordering can be modeled by reordering that occurs within the bilingual phrase pairs themselves (for brevity we will refer to these as Translation Units, or TUs from now on). For some language pairs, typically those that are close in origin, long distance reordering is relatively rare. It is our hypothesis that it may be possible to perform translation between these languages more efficiently and more effectively by adopting a strictly monotonic phrase-based translation approach that performs only local reordering.

This paper contributes a study of the effect of constraining the decoding process to be strictly monotonic for languages with similar word ordering characteristics. We also provide a study on the effect of leveraging some of the techniques developed for transliteration generation to in an attempt to improve machine translation performance for language pairs that may benefit from a strictly monotonic translation strategy. The structure of the paper is as follows: in the next section we describe some existing work that has used a monotonic approach to translation. Then in Section III we describe a set of experiments that aim to assess the effect of monotonic decoding on machine translation performance. In Section IV we introduce monotonicity into the training by using a monotonic bilingual aligner. Finally, in Section V we draw our conclusions.

Throughout this paper we will use the following abbreviations for the languages used in our experiments (en=English, id=Indonesian Malay, ja=Japanese, ko=Korean, ms=Malaysian Malay, my=Myanmar, th=Thai, zh=Chinese).

II. Related Work

The task of monotonic sequence transduction we are undertaking is very similar to the task of transliteration generation and therefore we draw on the existing work in this field. Most related to our work is the work on using PBSMT techniques to perform transliteration (for example [1], [2], [3]) since here the overlap of technology is the greatest.

There is much interest in the field of machine translation in both pre- [4] and post-ordering, [5], [6]. The aim of these approaches is to handle the complex task of reordering in a separate process from the decoding in order to simply the decoding and facilitate the use of richer models of reordering.

III. Monotonic Decoding

A. Motivation

In this section we study the effect of constraining the decoding process to be monotonic for a set of language pairs with similar word ordering. We expect many (most) language pairs to be unsuitable for this type of approach since they require long range reordering, and have deliberately excluded them from our study. Our focus is on those pairs for which our approach might be expected to be effective.

978-0-7695-5063-3/13 $26.00 © 2013 IEEE
DOI 10.1109/IALP.2013.31

B. Methodology

In these experiments standard PBSMT systems were trained with lexical reordering models using the MOSES [7] toolkit. The decoding was performed using the same models both with and without reordering. The systems were evaluated using the BLEU score [8]. The weights for the log-linear models were tuned using the standard MERT [9] procedure, however to avoid variance on the results we chose to re-use the MERT weights trained with reordering for the models that decoded monotonically. Therefore the results for the monotonic decoding processes represent a lower bound for their performance; that is, it may have been possible to obtain higher BLEU scores for these systems had we tuned the model weights specifically for them.

C. Experiments

1) Corpora: We used three Asian SOV languages (ja, ko and my), and three SVO languages (en, th and zh) from the multilingual Basic Travel Expression Corpus (BTEC), which is a collection of travel-related expressions [10]. We used word segmentation for all languages except Myanmar for which there is no available word segmenter; in this case we used syllable segmentation. We created bilingual corpora by pairing languages with the same basic word order characteristics together: that is we created SOV-SOV and SVO-SVO pairs. The corpus statistics of the languages are summarized in Table I. The number of sentence pairs used in the experiments was the same for each language: 65k training sentences, 10k development sentences and 1k test sentences.

LANGUAGE	TRAIN	DEVELOPMENT	TEST
en	527,268	86,934	7,901
id	474,542	77,446	7,801
ja	594,127	95,727	9,266
ko	559,243	89,517	8,777
ms	479,054	77,777	7,611
my	835,030	123,961	12,654
th	512,054	86,401	8,811
zh	485,151	77,101	7,711

Table I
CORPUS SIZE IN WORDS FOR EACH LANGUAGE

2) Results: The results of this experiment are shown in Figure 1. It can be seen from the graph that the language pairs we selected have clustered into two groups: the SOV-SOV language pairs in one cluster and the SVO-SVO language pairs in the other. The monotonic decoding had a negligible impact on BLEU scores of the systems based on the SOV-SOV languages. Most of the systems were less than 0.05 BLEU points different, the largest differences of around 0.23 BLEU points were observed for the systems that translated from Myanmar. This motivated us to extend our approach from monotonic decoding to monotonic alignment and decoding.

IV. MONOTONIC ALIGNMENT

A. Motivation

The motivation for using a monotonic alignment strategy is two-fold. First, from a machine-learning point of view, the task of bilingual alignment is considerably simpler without the reordering component. Our hypothesis was that it should be possible to obtain more accurate alignments that lead to higher machine translation quality as a result. Second, the construction of a phrase table via a different methodology provides an opportunity to combine both alignment approaches to produce a phrase table containing potentially more useful entries and reliable statistics for the entries that these approaches mutually discover.

B. Methodology

To perform the bilingual alignment we chose to use the Bayesian non-parametric method of [11]. The advantage of this approach is that it is capable of many-to-many alignment without the tendency to overfit the training data. In addition we adopted the TU extraction heuristic used in their paper which proceeds as follows: within a single bilingual word-pair, agglomerate all contiguous bilingual sequence-pairs in all possible ways, but limit the size of the resulting source and target phrases to match the *maximum phrase length* parameter used to train the PBSMT system (this was set to 7 in our experiments). This is not strictly necessary, but we performed this step to keep the phrase-table generated from our Bayesian alignment comparable to that generated by the baseline system. The TU composition process for a single bilingual word-pair is illustrated in Figure 2.

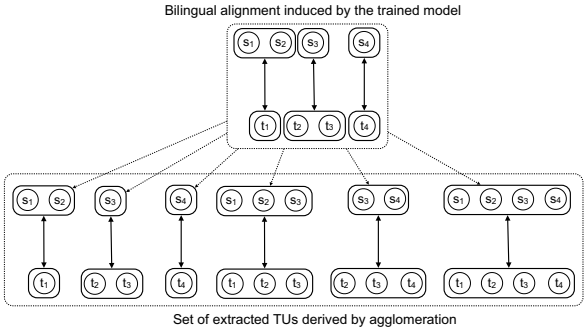

Figure 2. The sequence-pair extraction process for a single bilingual word pair, using TU agglomeration.

C. Experiments

1) Corpora: The corpora used for these experiments were the same set described in the previous section (summarized in Table I). We also added another pair of related Asian languages to the experiments (id and ms).

2) Experimental Conditions: We ran two sets of experiments, in the first set we linearly interpreted the phrase tables from both alignment methods together. We also added an indicator feature into the log-linear model of the PBSMT system to indicate those TUs that were in the intersection of both phrase tables. This feature was unity for TUs that were in only one table, and a constant value (e) for those in the intersection. This feature had a log-linear weight that was tuned together with the other weights during MERT. In the second set of experiments,

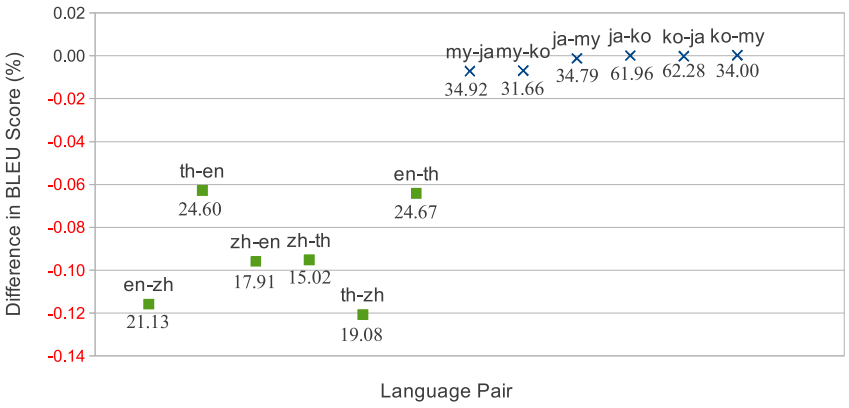

Figure 1. The effect of decoding monotonically on translation performance (cross points on the graph are SOV-SOV language pairs; squares are points for the SVO-SVO language pairs; absolute BLEU scores are given below the points).

we removed this indicator feature from the model and tuned the model that excluded this feature with MERT.

In addition we attempted to build a system based purely on transliteration technology. We trained the DirecTL+ string transduction model [12], (a transliteration model capable of state-of-the-art performance in transliteration generation [13]) on the alignment induced by our Bayesian alignment, and used the model to decode the test data. We were unable to obtain results that were comparable with those from the PBSMT approaches, the BLEU scores being several percentage points lower. We therefore performed our experiments solely with the PBSMT systems, but we only investigated a limited number of configurations of the DirecTL+ models, and it may be possible to achieve better performance with different settings. We believe the difference in the characteristics of translation data may be the cause. The type set size for transliteration is typically small, whereas for translation it is far larger and could potentially lead to issues due to data sparseness.

3) Results: The results for the experiments with the phrase-tables arising from both alignment methods and including the indicator feature are shown in Table III. The cells highlighted in gray are the optimal values for the interpolation parameter selected by the BLEU score attained on the development data. An interpolation parameter of 0 is equivalent to the case where only the phrase table generated by using GIZA++ was used. Conversely a weight of 1 is equivalent to a system that only used the phrase table derived from the Bayesian alignment. The two alignment methods seem to give approximately comparable performance in isolation. The phrase table integration lead to a respectable improvement in performance, with an average gain of around 0.7 BLEU points.

The size of the phrase tables resulting from both approachs along with the size of the integrated tables is shown in Table II. There is quite a large overlap between the phrase tables, and the phrase tables from each of the alignment approaches are similar in size.

The results for the experiments in which the indicator feature was removed from the model are shown in Table IV. These scores are slightly lower than in the previous experiment, but the differences are small . This suggests the indicator feature is not contributing much to the system, and that the real gains are coming from the introduction of new TUs and/or from improvements in the quality of their associated features.

LANG. PAIR	GIZA++ TUs	Bayesian TUs	Integrated TUs
my-ko	734,936	1,746,263	118,482
ko-my	738,895	1,744,945	120,525
ja-ko	859,035	634,081	405,931
ko-ja	858,955	630,315	407,163
id-ms	1,050,174	1,179,212	930,854
ms-id	1,050,223	852,935	763,788

Table II
PHRASE TABLE STATISTICS

V. CONCLUSION

This paper explored the idea of using a fully monotonic training and decoding process for the translation of a limited subset of all language pairs. Our experimental results show that for a reasonably broad set of SOV Asian languages, restricting the decoding process to monotonic has very little effect on the machine translation output quality. The decoding complexity is dramatically reduced by decoding monotonically and this motivated our study on extending the monotonic approach to the training phase. We aligned our corpora using both GIZA++ and a non-parametric Bayesian monotonic sequence aligner not previously used on translation data. Unlike the results in the transliteration field, we were unable to improve MT performance using the phrase table from the Bayesian aligner alone, but the monotonic alignment was sufficiently different from the GIZA++ alignment that were able to obtain a substantial improvement (averaging approximately 0.7 BLEU points) in MT performance on all of the language pairs in our experiments.

REFERENCES

[1] A. Finch and E. Sumita, "Phrase-based machine transliteration," in *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, vol. 1, Hyderabad, India, 2008.

LANGUAGE PAIR	INTERPOLATION WEIGHT										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
my-ja	33.96	**34.08**	33.98	33.32	33.10	32.79	32.46	31.42	30.55	29.72	26.70
ja-my	**32.96**	32.04	32.09	31.43	31.40	31.22	30.95	29.83	28.63	28.07	26.84
my-ko	**25.30**	25.01	24.80	24.34	23.66	22.44	21.43	20.23	19.34	18.69	17.78
ko-my	**26.08**	25.69	24.80	24.41	24.22	23.66	22.63	21.31	20.18	19.54	19.51
ja-ko	58.53	**59.33**	59.03	58.82	58.38	56.25	52.82	51.13	49.43	48.57	46.43
ko-ja	58.41	**59.07**	58.94	58.71	58.53	55.72	53.63	52.50	50.66	50.91	49.01
id-ms	**65.62**	65.21	65.01	64.96	65.00	64.77	64.14	63.96	63.22	63.03	61.77
ms-id	68.07	68.63	68.74	68.67	**68.89**	68.59	68.07	68.19	67.89	66.70	66.18

Table III
THE EFFECT ON THE BLEU SCORE OF INTERPOLATING THE PHRASE TABLES FROM DERIVED FROM TWO DIFFERENT ALIGNMENTS (BOLD NUMBERS INDICATE THE HIGHEST BLEU SCORE WITH TEST DATA AND HIGHLIGHTED CELLS INDICATE THE HIGHEST BLEU SCORE WITH DEVELOPMENT DATA)

LANGUAGE PAIR	INTERPOLATION WEIGHT										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
my-ja	33.84	**34.07**	33.87	33.35	33.13	32.98	32.24	31.39	30.56	29.56	26.55
ja-my	**32.76**	32.26	32.03	31.88	31.11	31.33	31.07	30.36	29.09	27.86	26.75
my-ko	**25.24**	25.11	24.80	24.35	23.61	22.44	21.32	20.20	19.47	18.76	17.75
ko-my	**26.12**	25.58	25.14	24.21	24.09	23.23	22.65	20.82	20.33	19.89	19.44
ja-ko	58.65	**59.31**	59.12	58.74	58.18	56.23	52.80	51.12	49.46	48.57	46.81
ko-ja	58.22	58.53	**58.92**	58.72	58.04	55.91	53.63	52.51	50.66	50.93	49.03
id-ms	64.99	**65.12**	65.01	65.04	65.00	64.89	64.11	63.99	63.56	63.19	61.63
ms-id	68.11	68.58	68.74	68.67	**68.89**	68.58	68.01	68.01	67.68	66.62	66.17

Table IV
THE EFFECT ON THE BLEU SCORE OF REMOVING THE INDICATOR FEATURE (BOLD NUMBERS INDICATE THE HIGHEST BLEU SCORE WITH TEST DATA AND HIGHLIGHTED CELLS INDICATE THE HIGHEST BLEU SCORE WITH DEVELOPMENT DATA)

[2] T. Rama and K. Gali, "Modeling machine transliteration as a phrase based statistical machine translation problem," in *NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 124–127.

[3] S. Noeman, "Language independent transliteration system using phrase based smt approach on substrings," in *NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 112–115.

[4] F. Xia and M. McCord, "Improving a statistical mt system with automatically learned rewrite patterns," in *Proceedings of the 20th international conference on Computational Linguistics*, ser. COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. [Online]. Available: http://dx.doi.org/10.3115/1220355.1220428

[5] K. Sudoh, X. Wu, K. Duh, H. Tsukada, and M. Nagata, "Post-ordering in statistical machine translation," in *Proc. MT Summit*, 2011.

[6] I. Goto, M. Utiyama, and E. Sumita, "Post-ordering by parsing for japanese-english statistical machine translation," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ser. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 311–316. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390665.2390737

[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowa, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," in *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007): demo and poster sessions*, Prague, Czeck Republic, June 2007, pp. 177–180.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2001, pp. 311–318.

[9] F. J. Och, "Minimum error rate training for statistical machine translation," in *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, 2003.

[10] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proceedings of EUROSPEECH-03*, 2003, pp. 381–384.

[11] A. Finch and E. Sumita, "A Bayesian Model of Bilingual Segmentation for Transliteration," in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, M. Federico, I. Lane, M. Paul, and F. Yvon, Eds., 2010, pp. 259–266.

[12] S. Jiampojamarn, C. Cherry, and G. Kondrak, "Joint processing and discriminative training for letter-to-phoneme conversion," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 905–913. [Online]. Available: http://www.aclweb.org/anthology/P/P08/P08-1103

[13] S. Jiampojamarn, K. Dwyer, S. Bergsma, A. Bhargava, Q. Dou, M.-Y. Kim, and G. Kondrak, "Transliteration generation and mining with limited training resources," in *Proceedings of the 2010 Named Entities Workshop*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 39–47. [Online]. Available: http://www.aclweb.org/anthology/W10-2405