Exploiting Directional Asymmetry in Phrase-table Generation for Statistical Machine Translation

Andrew Finch and Eiichiro Sumita
Language Translation Group
MASTAR Project
National Institute of Information and Communications Technology
{andrew.finch, eiichiro.sumita}@nict.go.jp

Abstract

This paper presents a method that can improve the translation quality of a phrase-based statistical machine translation system without the need for additional training data. The technique exploits the asymmetry of the phrase-table generation process during training. In our experiments we use the GIZA++ toolkit for alignment, and the phrase extraction utilities that are provided with the MOSES decoder. These tools are commonly used in the field, and serve as the benchmark by which other techniques are measured. Our experiments show that if the corpus's word order (both source and target) is reversed during the word alignment/phrase extraction phase of the training, the resulting phrase table is significantly different to that generated from the un-reordered corpus. Typically only about 30-60% of the phrase-pairs are shared between the forward- and reverse-generated phrase tables. Our approach attempts to exploit this asymmetry by integrating these phrase-tables into a single larger table, and use this integrated phrase table for decoding. The phrase-table integration is done by linearly interpolation. The benefits of this approach are two-fold. Firstly, the larger number of phrases present in the integrated phrase-table allows for greater coverage of the test data. Secondly, phrases that occur in both tables receive contributions to their probability mass from both entries in the tables during the interpolation process. This effectively boosts the probability of the more reliable phrases that occur in both tables relative to less reliable phrases that occur in only one of the tables. To evaluate our approach we ran a total of 272 experiments on all language-pairings from a set of 17 languages, and evaluated using a set of seven machine translation evaluation metrics. Our training data consisted of approximately 160,000 sentence pairs from the ATR BTEC1 corpus. The test set was 5000 single-reference sentences drawn from the same sample. We show consistent gains in over 95% of our experiments, over baseline systems trained in the usual manner on unreversed training data.

1. Introduction

Phrase-based statistical machine translation systems (SMT) currently pervade the field of machine translation research. These systems are simple in operation relative to other techniques, and offer state-of-the-art performance.

During the translation process the source sentence is implicitly segmented by the decoder, and the source word sequences arising from the segmentation are translated using bilingual word sequences called *phrase-pairs*. These phrase-pairs are extracted automatically from the corpus during training and are stored in a table, called the *phrase-table*. Since these phrase-pairs are used as the building blocks of the translation system, their reliability and also their number and variety are key in determining the quality of a phrase-based statistical machine translation system. Errors in the alignment process can give rise to erroneous phrase-pairs in the phrase table, or cause some genuinely useful phrase-pairs occurring in the bilingual training data to be missed by the extraction process.

The *de facto* standard process for phrase-table construction from a bilingual corpus is to use the GIZA++ toolkit, in combination with utilities provided with the MOSES machine translation decoder. This technique is commonly used in the field, and provides the benchmark by which other competing techniques are measured. The construction of the phrase-table in this manner involves two steps. In the first step the bilingual sentence pairs are word-aligned in a 'one-to-many' fashion. This alignment is carried out using the IBM Model 4 [2]. This model includes the notion of a CEPT, the sequence of 'many' words, generated according to the model from the 'one' word. The model causes words in the CEPT to be generated in a sequence from left-to-right, and thus the alignment process is not symmetrical in the sense that if the order of the word sequences in the data are reversed then the alignment obtained will not necessarily be the same for forward and reversed data. In the second step, both alignments are combined and a set of heuristics are used to extract a set of phrase-pairs consistent with the alignment. Note that in the traditional approach, an attempt is made

to symmetrize the process with respect to the order of the languages (source and target). Our approach attempts to symmetrize with respect to the word order.

2 Experiments

2.1 Methodology

2.1.1 Phrase-table Generation

Our approach involves running the word alignment and phrase extraction components of the machine translation system training scheme twice. The first run is the same as in the normal training process. The second run uses the same procedure with one difference; the source and target word order of the sentences in the corpus is reversed. Each of these processes gives rise to a different phrase table (see Section 3.1 for details).

2.1.2 Phrase-table Interpolation

The two phrase-tables produced by the above process are combined by linear interpolation of their model probabilities, into a single integrated phrase table file which is then used in the normal way by the decoder.

2.1.3 Decoding

The decoder used is a standard phrase-based machine translation decoder that operates according to the same principles as the publicly available PHARAOH [6] and MOSES [7] SMT decoders. In these experiments 5-gram language models built with Witten-Bell smoothing were used along with a lexicalized distortion model. The system was trained in a standard manner, using a minimum error-rate training (MERT) procedure [8] with respect to the BLEU score [9] on held-out development data to optimize the log-linear model weights.

2.1.4 Experimental Data

The experiments were conducted on all possible pairings among 17 languages, giving rise to a total of 272 experiments. A key to the acronyms used for languages together with information about their respective characteristics is given in Table 1.

We used all of the first ATR Basic Travel Expression Corpus (BTEC1) [5] for these experiments. This corpus contains the kind of expressions that one might expect to find in a phrase-book for travelers. The corpus is similar in character to the IWSLT06 Evaluation Campaign on Spoken Language Translation [10] J-E open track. The sentences are relatively short (see Table 1) with a simple structure and a fairly narrow range of vocabulary due to the limited domain.

The experiments were conducted on data that contained no case information, and also no punctuation (this was an arbitrary decision that we believe had no impact on the results).

We used a 1000 sentence development corpus for all experiments, and the corpus used for evaluation consisted of 5000 sentences with a single reference for each sentence. The evaluation set was deliberately large to minimize the variance in the results.

3 Results

3.1 Asymmetry

Our method relies on the fact that the word order of the sentences being processed influences the phrases extracted from the corpus. We measured the degree to which the phrase tables differ by calculating the percentage of all phrases extracted that are shared by both of the systems. These figures are given in Table 2. The table clearly shows that the phrase-table overlap is dependent on the language pair. In our experimental set, the lowest overlap is only 25% for Arabic-Japanese. The highest overlap being 95% for Malaysian Malay and Indonesian Malay. Languages in the table appear have a high overlap with languages that have a similar word order. Japanese and Korean, for example, have a low amount of phrase-table overlap with all languages, with the exception of each other. Japanese and Korean have a similar grammatical structure, but are also relatively free with their word order. This may create differences in the relative sentence positions of corresponding source and target words with other languages, which in turn amplifies the asymmetry in the phrase generation process to yield quite different phrase-tables. Indonesian and Malaysian in contrast have almost identical word order, and give virtually identical phrase tables.

It is clear from Table 2 in most cases the phrase-tables generated are significantly different.

3.2 System Evaluation

The results presented in this paper are given in terms of the BLEU score [9]. This metric measures the geometric mean of n-gram precision of n-grams drawn from the output translation and a set of reference translations for that translation.

There are large number of proposed methods for carrying out machine translation evaluation. Methods differ in their focus of characteristics of the translation (for example fluency or adequacy), and moreover anomalous results can occur if a single metric is relied on. Therefore, we also carried out evaluations using the NIST [3], METEOR [1], WER [4], PER [12] and TER [11] machine translation evaluation techniques. However, the results were similar in character no matter which technique was chosen for evaluation. BLEU was chosen, as it is most commonly used in the field.

Abbreviation	Language	#Words	Avg. sent length	Vocabulary	Order
ar	Arabic	806853	5.16	47093	SVO
da	Danish	806853	5.16	47093	SVO
de	German	907354	5.80	23443	SVO
en	English	970252	6.21	12900	SVO
es	Spanish	881709	5.64	18128	SVO
fr	French	983402	6.29	17311	SVO
id	Indonesian (Malay)	865572	5.54	15527	SVO
it	Italian	865572	5.54	15527	SVO
ja	Japanese	1149065	7.35	15405	SOV
ko	Korean	1091874	6.98	17015	SOV
ms	Malaysian (Malay)	873959	5.59	16182	SVO
nl	Dutch	927861	5.94	19775	SVO
pt	Portuguese	881428	5.64	18217	SVO
ru	Russian	781848	5.00	32199	SVO
th	Thai	1211690	7.75	6921	SVO
vi	Vietnamese	1223341	7.83	8055	SVO
zh	Chinese	873375	5.59	14854	SVO

Table 1. Key to the languages, corpus statistics and word order. SVO denotes a language that predominantly has subject-verb-object order, and SOV denotes a language that predominantly has subject-object-verb order

4. Conclusion

The average improvement over all the experiments was 0.38 BLEU percentage points. This improvement is consistent across languages, and in 95% of our experiments we were able to show an improvement over the baseline system. Furthermore, this technique is simple to implement and adds only a small amount of time to the training process and requires no additional data or resources to be used.

Acknowledgment

This work is partly supported by the Grant-in-Aid for Scientific Research (C) Number 19500137.

References

[1] S. Banerjee and A. Lavie. Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In *ACL-2005: Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[2] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[3] G. Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the HLT Conference*, San Diego, California, 2002.

[4] M. J. Hunt. Figures of merit for assessing connected-word recognisers. In *In Proceedings of the ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases*, pages 127–131, 1989.

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

[6] P. Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Machine translation: from real users to research: 6th conference of AMTA*, pages 115–124, Washington, DC, 2004.

[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 *ACL 2007: proceedings of demo and poster sessions*, pages 177–180, Prague, Czeck Republic, June 2007.

[8] F. J. Och. Minimum error rate training for statistical machine trainslation. In *Proceedings of the ACL*, 2003.

[9] 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*, pages 311–318, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[10] M. Paul. Overview of the iwslt 2006 evaluation campaign. In *Proceedings of the IWLST*, 2006.

[11] M. Snover, B. Dorr, R. Schwartz, J. Makhoul, L. Micciula, and R. Weischedel. A study of translation error rate with targeted human annotation. Technical report, University of Maryland, College Park and BBN Technologies, July 2005.

[12] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated dp based search for statistical translation. In *In European Conf. on Speech Communication and Technology*, pages 2667–2670, 1997.

	ar	da	de	en	es	fr	id	it	ja	ko	ms	nl	pt	ru	th	vi	zh
ar	-	45.4	43.2	56.6	50.7	42.2	50.3	48.5	25.7	30.2	50.5	43.8	48.7	44.5	40.5	39.6	34.2
da	45.4	-	67.5	79.0	60.5	55.5	56.6	58.8	27.2	29.7	57.3	69.4	59.4	54.5	45.1	51.2	34.9
de	42.9	67.4	-	76.2	59.1	53.8	53.2	56.6	28.2	29.0	53.5	75.5	57.2	54.3	40.9	47.1	34.2
en	56.6	79.0	76.2	-	72.7	65.7	63.9	69.9	30.9	33.8	63.7	78.7	73.0	60.4	51.7	58.4	37.7
es	50.6	60.5	59.2	72.8	-	60.3	53.7	72.2	29.8	31.2	53.9	62.1	76.1	52.9	42.9	47.1	35.8
fr	42.1	55.5	53.7	65.8	60.3	-	51.6	62.8	27.1	29.4	50.7	57.9	59.1	48.6	40.6	44.1	34.3
id	50.2	56.5	53.1	63.9	53.7	51.6	-	51.8	33.8	38.5	95.7	54.2	52.6	55.4	57.6	59.4	45.0
it	48.2	58.7	56.5	69.9	72.1	62.7	51.9	-	28.2	29.5	52.3	58.9	72.1	50.4	40.5	44.6	34.7
ja	25.5	27.3	27.9	31.3	29.6	27.2	34.0	28.3	-	71.0	32.8	27.6	29.6	27.1	26.8	27.5	38.7
ko	29.8	29.3	28.7	33.7	31.0	29.3	38.0	29.2	71.0	-	37.4	28.8	30.0	29.8	28.8	29.3	40.9
ms	50.4	57.1	53.4	63.7	53.8	50.6	95.7	52.2	32.6	37.7	-	54.5	53.2	55.5	56.7	60.0	44.6
nl	43.9	69.5	75.6	78.7	62.0	58.0	54.3	59.0	27.5	29.3	54.6	-	59.5	54.0	42.7	49.2	34.7
pt	48.6	59.4	57.2	73.0	76.1	59.0	52.7	72.1	29.6	30.2	53.3	59.4	-	52.3	43.3	46.8	35.8
ru	44.6	54.5	54.3	60.3	52.8	48.6	55.4	50.5	27.7	30.4	55.5	54.0	52.5	-	45.3	45.1	36.0
th	40.2	44.7	40.7	51.4	42.5	40.4	57.4	40.2	26.4	28.7	56.6	42.3	43.0	45.1	-	51.1	36.4
vi	39.1	50.8	46.7	58.2	46.7	43.6	59.2	44.2	26.8	29.1	59.9	48.8	46.3	44.5	51.1	-	38.3
zh	34.4	34.8	34.4	37.8	35.8	34.6	45.0	34.6	39.0	41.3	44.8	34.8	35.7	35.8	36.7	38.8	-

Table 2. Overlap in the phrase tables. The figures in the tables represent the percentages of all of the phrases from both of the phrase tables that occur in both tables.

	ar	da	de	en	es	fr	id	it	ja	ko	ms	nl	pt	ru	th	vi	zh
ar	-	0.16	0.36	0.57	0.43	0.57	0.22	0.05	0.16	0.75	0.24	0.56	0.41	0.18	1.00	0.97	0.51
da	0.27	-	0.18	0.09	0.13	0.10	0.19	0.24	0.94	0.64	0.26	0.18	0.46	0.23	0.59	0.56	0.36
de	0.46	0.10	-	0.23	0.50	0.35	0.42	0.05	0.72	0.25	0.64	0.25	0.28	0.28	0.58	0.50	0.73
en	0.43	0.09	0.06	-	0.10	0.04	0.39	0.29	0.87	0.51	0.37	0.11	0.07	0.22	0.68	0.16	0.45
es	0.52	0.36	0.53	0.02	-	0.19	0.01	0.05	0.49	0.41	0.41	0.28	0.01	0.35	0.69	0.75	0.51
fr	0.69	0.13	0.28	0.24	0.15	-	0.38	0.11	0.89	0.21	0.09	0.28	0.36	0.53	0.41	0.43	0.22
id	0.26	0.13	0.31	0.63	0.38	0.56	-	0.46	0.30	0.84	0.03	0.41	0.24	0.41	0.22	0.53	0.18
it	0.25	0.40	0.03	0.17	0.00	0.24	0.56	-	0.74	0.49	0.45	0.12	0.08	0.73	0.41	0.92	0.55
ja	0.25	0.30	0.84	0.24	0.43	0.29	0.00	0.47	-	0.25	0.10	0.50	0.18	0.27	0.85	0.30	0.34
ko	0.28	0.85	0.74	0.02	0.40	0.59	0.26	0.32	0.06	-	0.10	0.31	0.16	0.75	1.02	0.93	0.37
ms	0.32	0.23	0.37	0.50	0.30	0.43	0.07	0.47	0.54	0.79	-	0.60	0.28	0.26	0.49	0.40	0.00
nl	0.58	0.10	0.46	0.03	0.02	0.16	0.68	0.52	0.32	0.28	0.33	-	0.07	0.39	0.33	0.88	0.35
pt	0.54	0.37	0.49	0.22	0.29	0.07	0.12	0.05	0.33	0.40	0.43	0.15	-	0.26	0.16	0.85	0.62
ru	0.45	0.24	0.45	0.41	0.00	0.22	0.49	0.60	0.61	0.77	0.38	0.67	0.07	-	0.92	0.87	0.38
th	1.08	0.71	0.65	0.57	0.56	0.28	0.23	0.75	0.56	0.62	0.46	0.63	0.66	0.64	-	0.15	0.50
vi	0.49	0.21	0.34	0.34	0.51	0.33	0.38	0.65	0.34	0.28	0.27	0.36	0.03	0.42	0.56	-	0.40
zh	0.27	0.53	0.82	0.36	0.48	0.10	0.29	0.27	0.35	0.66	0.39	0.55	0.80	0.30	0.46	0.47	-

Table 3. Gains in BLEU score from using an integrated translation model, over using a single translation model generated in the standard manner with GIZA++ alignments and phrase extraction heuristics from the corpus text in its default word order. The numbers in the cells are the differences in BLEU percentage points between the systems. Shaded cells indicate the cases where the baseline system give the higher score. Source languages are indicated by the column headers, the row headers denoting the target languages.