Bootstrapping a Phrase-based Statistical Machine Translation System

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

Abstract

One of the main causes of errors in statistical machine translation are the erroneous phrase pairs that can find their way into the phrase table. These phrases are the result of poor word-to-word alignments during the training of the translation model. These word alignment errors in turn cause errors during the phrase extraction phase, and these erroneous bilingual phrase pairs are then used during the decoding process and appear in the output of the machine translation system. This paper presents a technique in which preliminary machine translation systems are built with the sole purpose of indicating those sentence pairs in the training corpus that the systems are able to generate using their models, the hypothesis being that these sentence pairs are likely to make good training data for an SMT system of the same type. These sentences are then used to bootstrap a second SMT system, and those sentences identified as good training data are given additional weight during the training process for building the translation models. Using this technique we were able to improve the performance of a Japanese-to-English SMT system by 1.2-1.5 BLEU points on unseen evaluation data.

1. Introduction

The phrase-based statistical machine translation training process by its very nature relies on the assumption that the tokens comprising source and target sentences can be first aligned in both a one-to-many and many-to-one fashion [1], and then pairs of contiguous sequences of tokens from both source and target can be extracted using heuristics resulting in a table of source-target bilingual phrase pairs [11]. These phrase pairs are then used as the building blocks from which to derive the translation of the source sentence during the decoding process. When corpora contain sentence pairs that are poor translations of each other, non-literal translations (at the sentence level) or even erroneous entries that are not translations, the alignment component of the SMT training process can fail leading to the extraction of incorrect phrase pairs. The erroneous phrase pairs that are extracted can have a detrimental impact on machine translation quality, since they may be used to incorrectly translate word sequences in the source sentence. Previous approaches have relied on the relative frequencies of occurrence of these phrase pairs to either filter the phrase table, or decrease their importance in the decoding through the introduction of a new feature into the log-linear model [12]. In our approach we grade the quality of our training data using an SMT system. We train models and grade the training data according to whether or not we can generate the target word sequence from the source word sequence using the statistical models. If we can generate target from source we will denote the data as *decodable*, otherwise we denote it *undecodable*. Our hypothesis is that the decodable data more desirable because it will be easier to align automatically, and should therefore play a more important part in the word-alignment process. Giving this data a higher weight during training should lead to a better overall word alignment, more reliable phrases, and should cause phrases from this data to be preferred in the decoding process over competing less reliable phrases from the undecodable part of the training corpus since they will receive a greater translation probability.

2. Related Work

Recently research on identifying and utilizing bilingual phrase pairs that are used during the decoding of unseen data has become popular. In [17] data from the n-best lists of decoded monolingual source-language data are reintroduced as additional training data into the system after a process of re-scoring. In [4] a method of grading bilingual phrase pairs in accordance with statistics based on their usage is presented. First a phrase-table is built from the training data in the usual manner, then the training data are decoded using the models and statistics collected on how frequently the phrase pairs were used in the decoding of the training data. Statistics measuring how often phrase pairs were considered during the decoding process, and also how often they appeared in the best translation hypothesis were combined into a single score. The phrase table is then filtered in accordance with this score. Our approach differs in that we decode with stronger con-

straints. The approach used in [4] requires that the bilingual phrase-pairs fit the source, the target being generated freely, whereas in our approach the phrase-pairs are constrained to match both source and target. In [6] unreliable phrase pairs are filtered from the phrase table using a process of statistical significance testing. Our approach is related, but instead of filtering unreliable phrase pairs from the model, we bias the training process towards data which contains these reliable phrase pairs. The effect of this is two-fold: firstly the data used in the word alignment process will contain a larger proportion of data which should be easier to align, and one would expect this to lead to better alignments overall; secondly, phrase pairs arising from the decodable data will have a higher relative frequency (relative to the phrase pairs arising from the undecodable data) that will be used to assign translation probabilities to them.

3. Methodology

The overall system architecture of our technique is shown in Figure 1. The process can be divided into two parts:

1. Classifying the training data into decodable and undecodable bilingual sentence pairs
2. Training a translation model for a new SMT system by weighting the decodable and undecodable portions of the training set

The data that is split into decodable and undecodable parts is used only in training the translation model component of the final phrase-based SMT system. The language model gains nothing from grading the data in this way and is built in the usual manner from the target sentences of the training data (and optionally from larger amounts of additional external corpora).

3.1. Grading the Training Data

In order to classify the data into "decodable" or "undecodable" classes, a phrase-based SMT decoder [11] [9] was employed.

If the source sentence is unable to be covered by phrases from the phrase table, or if the source sentence is able to be covered, but the target is unable to be generated from the target-side phrases corresponding to the source phrases used to cover the source sentence, then the sentence is deemed "undecodable".

The grading process operates on the hypothesis that if a target word sequence can be generated from a source word sequence by an SMT system, then the word sequences are likely to be *good* translations of each other. By *good* here it is meant that the sentence pairs are not only translations but also translations that can be performed by means of the same mechanism by which the phrase-based translation system operates.

3.2. Jackknifing

We used a jackknifing procedure to ensure that the training data were decoded as unseen data by the SMT systems used to grade the data. If the SMT system had been applied to its own training data, erroneous phrases that were extracted during the training process would be used to successfully decode the sentences they were derived from. We chose to do 10-fold jackknifing in our experiments. Every sentence in the training set was decoded in the jackknifing process, thus it was possible to identify those sentences (unseen) that could be generated by an SMT system trained on similar data.

3.3. Weighting

Once we have partitioned the training data sets of "decodable" and "undecodable" sentence pairs, we need some way of biassing the training process of the translation models towards the "decodable" data. We do this simply by adding two copies of the decodable data to the training set. In this way the alignment of the decodable data will contribute more during the word alignment phase and moreover, the translation probabilities of bilingual phrase pairs derived from this data will be higher. The weighting was only used during the training of the translation models, and therefore only affected the processes of word alignment, and subsequent phrase-table construction. The language model from the baseline system (which was built from unweighted data) was used in the final SMT system.

4. Experiments

4.1. Experimental Data

We used all of the first ATR Basic Travel Expression Corpus (BTEC1) [8] 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 [16] J-E open track.

The sentences are relatively short (an average of 6 or 7 tokens) with a simple structure and a fairly narrow range of vocabulary due to the limited domain. We used only the Japanese and English portion of the corpus, and all the systems built for the experiments were translating from Japanese (the source language) into English (the target language).

The corpus consisted of 157317 sentence pairs with approximately 1 million English tokens 1.15 million tokens in Japanese. The default BTEC tokenization scheme was used. 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). The evaluation corpus consisted of 510 sentences drawn from the same sample as the training

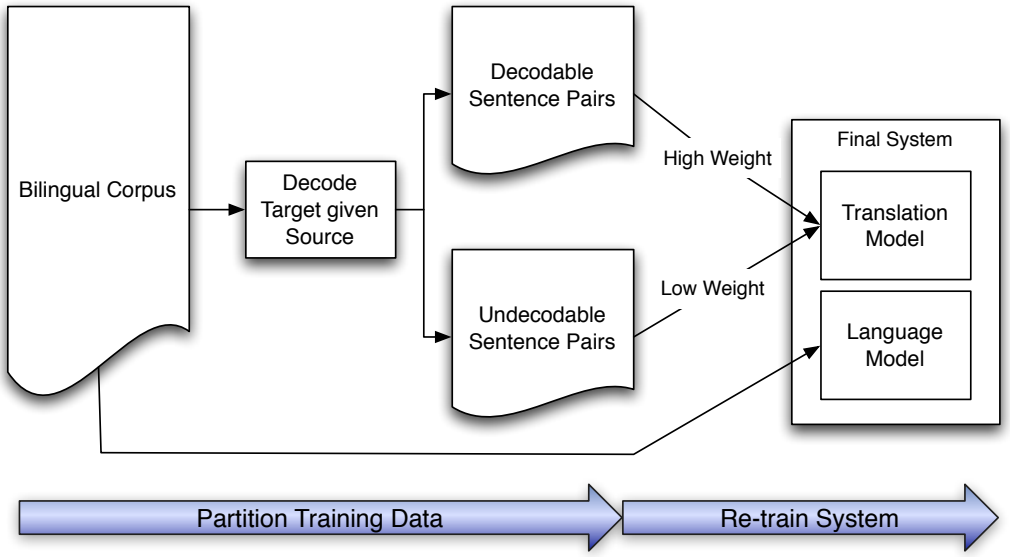

Figure 1. The overall system architecture

data. There were 16 reference sentences for each evaluation sentence. The evaluation corpus used was identical to that used as the IWSLT05 evaluation set [3].

4.2. Translation System

We used a variant of the NICT-ATR CleopATRa decoder in our experiments. The decoder was modified in order to be able to decode with respect to a supplied target word sequence. The decoder is a standard phrase-based machine translation decoder that operates according to the same principles as the publicly available PHARAOH [9] and MOSES [10] 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 [13] with respect to the BLEU score [15] on held-out development data to optimize the decoding parameters. The decodable training data was given a greater weight in the training process by using two copies of the data during training. Word alignment was done using the tools provided in the GIZA++ SMT toolkit [14].

5. Results and Discussion

The results of the evaluation of our proposed technique are shown in Table 1. In an attempt to reduce variance in our results, we conducted experiments both with and without MERT, and found similar gains in performance under both conditions. The baseline system consisted of an SMT system trained on all of the training data. The proposed system, consisted of the system trained on exactly the same data, but with two copies of the decodable data in the training set. Making two copies of the data corresponds to the "High weight" in Figure 1. The decodable data was determined using the jackknifing process described in Section 3.2. A language model for the final system was built using all of the target side of the training corpus. The baseline system achieved a BLEU score of 0.620 before MERT optimization, and 0.645 after MERT. The proposed system had a BLEU score of 0.632 before MERT and 0.660 after MERT. The gains before the MERT optimization process are the most telling since they do not include any variance due to differences in success of the MERT search process.

System	BLEU (no MERT)	BLEU (with MERT)
Baseline	0.620	0.645
Proposed	0.632	0.660
Gain	0.012	0.015

Table 1. Results

6. Conclusion

The results from our experiments are encouraging. We were able to show significant gains in performance in the Japanese-to-English translation task. This data is the cleanest and most worked over data in the BTEC1 collection, and as such represents very difficult data for our technique to show an improvement on. Nonetheless our approach improved over a baseline system trained in a standard fashion by over 1 BLEU point in all our experiments. In future experiments we would like to apply the technique to other language pairs, especially those with less mature corpora. We are hopeful that this technique

will be even more effective on these corpora since incorrect translations and other noisy data will fall into the undecodable category and its contribution to the models will be diminished. We feel that the means of weighting the decodable data during the training of the final translation model could be developed further. Instead of weighting the data by means of data duplication, it would be preferable to introduce a soft weight directly into the EM algorithm used to perform the word-level alignment. This weight could be set by means of the same minimum error rate criterium used to set the model weights during the training process [13].

References

[1] 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.

[2] S. F. Chen. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 9–16, Morristown, NJ, USA, 1993. Association for Computational Linguistics.

[3] M. Eck and C. Hori. Overview of the iwslt 2005 evaluation campaign. In *In Proceedings of International Workshop on Spoken Language Translation (IWSLT 2005)*, pages 11–32, 2005.

[4] M. Eck, S. Vogel, and A. Waibel. Translation model pruning via usage statistics for statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 21–24, Rochester, New York, April 2007. Association for Computational Linguistics.

[5] M. A. Fattah, D. B. Bracewell, F. Ren, and S. Kuroiwa. Sentence alignment using p-nnt and gmm. *Comput. Speech Lang.*, 21(4):594–608, 2007.

[6] Y.-S. Hwang, A. M. Finch, and Y. Sasaki. Improving statistical machine translation using shallow linguistic knowledge. *Computer Speech & Language*, 21(2):350–372, 2007.

[7] M. Kay and M. Röscheisen. Text-translation alignment. *Comput. Linguist.*, 19(1):121–142, 1993.

[8] 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.

[9] 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.

[10] 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.

[11] P. Koehn, F. J. Och, , and D. Marcu. Statistical phrase-based translation. In *In Proceedings of the Human Language Technology Conference*, Edmonton, Canada, 2003.

[12] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney. The rwth statistical machine translation system for the iwslt 2006 evaluation. In *Proc. of the International Workshop on Spoken Language Translation*, pages 103–110, Kyoto, Japan, 2006.

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

[14] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[15] 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.

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

[17] N. Ueffing, G. Haffari, and A. Sarkar. Transductive learning for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[18] T. Utsuro, H. Ikeda, M. Yamane, Y. Matsumoto, and M. Nagao. Bilingual text, matching using bilingual dictionary and statistics. In *Proceedings of the 15th conference on Computational linguistics*, pages 1076–1082, Morristown, NJ, USA, 1994. Association for Computational Linguistics.