Discriminative Alignment Training without Annotated Data for Machine Translation

Patrik Lambert, Rafael E. Banchs and Josep M. Crego
TALP Research Center
Jordi Girona Salgado 1–3
08034 Barcelona, Spain
{lambert, rbanchs, jmcrego}@gps.tsc.upc.edu

Abstract

In present Statistical Machine Translation (SMT) systems, alignment is trained in a previous stage as the translation model. Consequently, alignment model parameters are not tuned in function of the translation task, but only indirectly. In this paper, we propose a novel framework for discriminative training of alignment models with automated translation metrics as maximization criterion. In this approach, alignments are optimized for the translation task. In addition, no link labels at the word level are needed. This framework is evaluated in terms of automatic translation evaluation metrics, and an improvement of translation quality is observed.

1 Introduction

In the first SMT systems (Brown et al., 1993), word alignment was introduced as a hidden variable of the translation model. When word-based translation models have been replaced by phrase-based models (Zens et al., 2002), alignment[1] and translation model training have become two separated tasks.

The system of Brown *et al.* was based on the noisy channel approach. Present SMT systems use a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (Och and Ney, 2002). Within this new framework translation quality can be tuned by adjusting the weight of each feature function in the log-linear combination. In order to improve translation quality, this tuning can be effectively performed by minimizing translation error over a development corpus for which manually translated references are available (Och, 2003). As a separate first stage of the process, alignment is not in practice directly tuned in function of the machine translation task.

Tuning alignment for an MT system is subject to practical difficulties. Unsupervised systems (Och and Ney, 2003; Liang et al., 2006) are based on generative models trained with the EM algorithm. They require large computational resources, and incorporating new features is difficult. In contrast, adding new features to some supervised systems (Liu et al., 2005; Moore, 2005; Ittycheriah and Roukos, 2005) is easy, but the need of annotated data is a problem.

A more general difficulty, however, is that of finding an alignment evaluation metric favoring alignments which benefit Machine Translation. The fact that the required alignment characteristics depend on each particular system makes it even more difficult. It seems that high precision alignments are better for phrase-based SMT (Chen and Federico, 2006; Ayan and Dorr, 2006), whereas high recall alignments are more suited to N-gram SMT (Mariño et al., 2006). In this context, alignment quality improvements does not necessarily imply translation quality improvements. This is in agreement with the observation of a poor correlation between word alignment error rate (AER (Och and Ney, 2000)) and automatic translation evaluation metrics (Ittycheriah and Roukos, 2005; Vilar et al., 2006).

[1] Hereinafter, alignment will refer to word alignment, unless otherwise stated.

Proceedings of NAACL HLT 2007, Companion Volume, pages 85–88,
Rochester, NY, April 2007.

Recently some alignment evaluation metrics have been proposed which are more informative when the alignments are used to extract translation units (Fraser and Marcu, 2006; Ayan and Dorr, 2006). However, these metrics assess translation quality very indirectly.

In this paper, we propose a novel framework for discriminative training of alignment models with automated translation metrics as maximization criterion. Thus we just need a reference aligned at the sentence level instead of link labels at the word level.

The paper is structured as follows. Section 2 explains the models used in our word aligner, focusing on the features designed to account for the specificities of the SMT system. In section 3, our minimum error training procedure is described and experimental results are shown. Finally, some concluding remarks and lines of further research are given.

2 Bilingual Word Aligner

For versatility and efficiency requirements, we implemented BIA, a BIlingual word Aligner similar to that of Moore (2005). BIA consists in a beam-search decoder searching, for each sentence pair, the alignment which minimizes the cost of a linear combination of various models. The differences with the system of Moore lie in the features, which we specially designed to suit our translation system (N-gram SMT (Mariño et al., 2006)). Its particularity is the translation model, which is based on a 4-gram language model of bilingual units referred to as tuples. Two issues regarding this translation model can be dealt with at the alignment stage.

Firstly, in order to estimate the bilingual n-gram model, only one monotonic segmentation of each sentence pair is performed. Thus long reorderings cause long and sparse tuples to be extracted. For example, if the first source word is linked to the last target word, only one tuple can be extracted, which contains the whole sentence pair. This kind of tuple is not reusable, and the data between its two extreme words are lost.

Secondly, it occurs very often that unlinked words (*i.e.* linked to NULL) end up producing tuples with NULL source sides. This cannot be allowed since no NULL is expected to occur in a translation input. This problem is solved by preprocessing alignments before tuple extraction such that any unlinked target word is attached to either its precedent or its following word.

Taking theses issues into account, we implemented the following features:

- distinct source and target unlinked word penalties: since unlinked words have a different impact whether they appear in the source or target language, we introduced an unlinked word feature for each side of the sentence pair.
- link bonus: in order to accommodate the N-gram model preference for higher recall alignment, we introduced a feature which adds a bonus for each link in the alignment.
- embedded word position penalty: this feature penalizes situations like the one depicted in figure 1. In this example, the bilingual units s2-t2 and s3-t3 cannot be extracted because word positions s2 and s3 are embedded between links s1-t1 and s4-t1. Thus the link s4-t1 may introduces data sparseness in the translation model, although it may be a correct link. So we want to have a feature which counts the number of embedded word positions in an alignment.

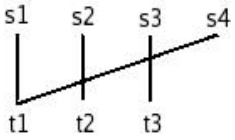

Figure 1: Word positions embedded in a tuple.

In addition to the embedded word position feature, we used the same two distortion features as Moore to penalize reorderings in the alignment (one sums the number of crossing links, and the other one sums the amplitude of crossing links). We also used the ϕ^2 score (Gale and Church, 1991) as a word association model, and as a POS-tags association model.

3 Experimental Work

For these experiments we used the Chinese-English data provided for IWSLT'06 evaluation campaign (Paul, 2006). The training set contains 46000 sentences (of 6.7 and 7.0 average length). Parameters were tuned over the development set (dev4) provided, consisting of 489 sentences of 11.2 words in average, with 7 references. Our test set was a selection of 500 sentences (of 6 words in average, with 16 references) among dev1, dev2 and dev3 sets.

3.1 Optimization Procedure

Once the alignment models were computed, a set of optimal log-linear coefficients was estimated via the optimization procedure depicted in Figure 2.

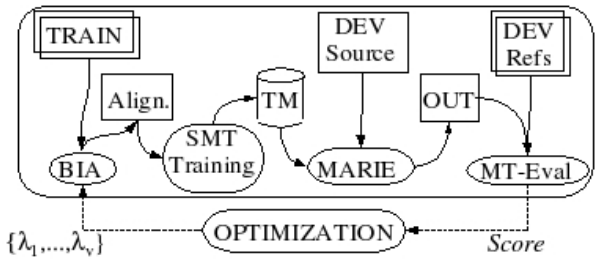

Figure 2: Optimization loop.

The training corpus was aligned with a set of initial parameters $\lambda_1, \ldots, \lambda_7$. This alignment was used to extract tuples and build a bilingual N-gram translation model (TM). A baseline SMT system, consisting of MARIE decoder and this translation model as unique feature[2], was used to produce a translation (OUT) of the development source set. Then, translation quality over the development set is maximized by iteratively varying the set of coefficients.

The optimization procedure was performed by using the SPSA algorithm (Spall, 1992). SPSA is a stochastic implementation of the conjugate gradient method which requires only two evaluations of the objective function. It was observed to be more robust than the Downhill Simplex method when tuning SMT coefficients (Lambert and Banchs, 2006).

Each function evaluation required to align the training corpus and build a new translation model. The algorithm converged after about 80 evaluations, lasting each 17 minutes with a 3 GHz processor. Alignment decoding was performed with a beam of 10 (it took 50 seconds and required 8 MB memory).

Finally, the corpus was aligned with the optimum set of coefficients, and a full SMT system was build, with a target language model (trained on the provided training data), a word bonus model and two lexical models. SMT models weights were optimized with a standard Minimum Error Training (MET) strategy[3] and the test corpus was translated with the full system. To contrast the results, full translation systems were also build extracting tuples from various combinations of GIZA++ alignments (trained with 50 classes and respectively 4,5 and 4 iterations of models 1,HMM and 4). In order to limit the error introduced by MET, we translated the test corpus with three sets of SMT model weights, and took the average and standard deviation.

3.2 Results

Table 1 shows results obtained with the full SMT system on the test corpus, with GIZA++ alignments, and BIA alignments optimized in function of three metrics: BLEU, NIST, and BLEU+4*NIST. The standard deviation is indicated in parentheses. Although results for systems trained with different BIA alignments present more variability than systems trained with GIZA++ alignments, they achieve better average scores, and one of them obtains much higher scores. Unexpectedly, BIA alignments tuned with NIST yield the system with worse NIST score.

4 Conclusions and further work

We proposed a novel framework for discriminative training of alignment models with automated translation metrics as maximization criterion. According to this type of metrics, the translation systems trained from the optimized alignments clearly performed better than the ones trained from Giza++ alignment combinations.

In addition, this first version of the alignment system has very basic models and could be improved. We could certainly improve the association score model, for example adding discount factors or adding more association score types, or dictionaries.

During the alignment coefficient optimization depicted in Figure 2, only the baseline SMT system is used. In future work, we could consider using various SMT features (as would be required for a phrase-based SMT system).

Our approach, as it is, cannot be applied to a large corpus, since it requires to align the whole training corpus at each iteration. Thus an interesting further research would consist in determining whether the

[2] An N-gram SMT system can produce good translations without additional target language model since the target language is modeled inside the bilingual N-gram model.

[3] SMT parameters are not optimized together with alignment parameters for two main reasons. Firstly, translation is more sensitive to variations of SMT parameters. Secondly, alignment is optimized over the full training set, whereas SMT is tuned over the development set.

System	BLEU	NIST	PER	WER
GIZA++ union	42.7 (1.1)	8.82 (0.07)	34.7 (0.2)	43.7 (0.4)
GIZA++ intersection	42.4 (0.9)	8.53 (0.07)	37.0 (0.9)	45.0 (1.3)
GIZA++ Zh→En	43.7 (0.9)	8.90 (0.2)	37.2 (1.4)	45.5 (2.0)
BIA (BLEU)	44.8 (0.4)	9.00 (0.04)	35.7 (0.07)	43.8 (0.09)
BIA (BLEU+4*NIST)	47.0 (1.5)	8.83 (0.4)	32.9 (0.8)	40.9 (0.5)
BIA (NIST)	44.8 (0.1)	8.55 (0.14)	33.0 (0.2)	41.4 (0.5)

Table 1: *Automatic translation evaluation results.*

alignment parameters trained on a part of the corpus are valid for the whole corpus.

Finally, some Giza++ parameters may also be tuned, in the same way as for BIA parameters.

5 Acknowledgments

This work has been partially funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation -(IST-2002-FP6-506738, http://www.tc-star.org) and by the Spanish Government under grant TEC2006-13964-C03 (AVIVAVOZ project).

References

Necip F. Ayan and Bonnie J. Dorr. 2006. Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT. In *Proc. COLING-ACL*, pages 9–16, Sydney, Australia.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Boxing Chen and Marcello Federico. 2006. Improving phrase-based statistical translation through combination of word alignment. In *Proc. FinTAL*, Turku, Finland.

Alexander Fraser and Daniel Marcu. 2006. Semi-supervised training for statistical word alignment. In *Proc. COLING-ACL*, pages 769–776, Sydney, Australia.

W. Gale and K. W. Church. 1991. Identifying word correspondences in parallel texts. In *DARPA Speech and Natural Language Workshop*, Asilomar, CA.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proc. HLT-EMNLP*, pages 89–96, Vancouver, Canada.

Patrik Lambert and Rafael E. Banchs. 2006. Tuning Machine Translation Parameters with SPSA. In *Proc. IWSLT*, pages 190–196, Kyoto, Japan.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proc. the HLT-NAACL*, pages 104–111, New York City, USA.

Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proc. ACL*, pages 459–466, Ann Arbor, Michigan.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A.R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proc. HLT-EMNLP*, pages 81–88, Vancouver, Canada.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proc. COLING*, pages 1086–1090, Saarbrucken,Germany.

F.J. Och and H. Ney. 2002. Dicriminative training and maximum entropy models for statistical machine translation. In *Proc. ACL*, pages 295–302, Philadelphia, PA.

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

F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL*, pages 160–167.

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proc. IWSLT*, pages 1–15, Kyoto, Japan.

James C. Spall. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Control*, 37:332–341.

David Vilar, Maja Popovic, and Hermann Ney. 2006. AER: Do we need to "improve" our alignments? In *Proc. IWSLT*, pages 205–212, Kyoto, Japan.

R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In Springer Verlag, editor, *Proc. German Conf. on Artificial Intelligence (KI)*.