Reordering Modeling using Weighted Alignment Matrices

Wang Ling, Tiago Luís, João Graça, Luísa Coheur and Isabel Trancoso
L^{2}F Spoken Systems Lab
INESC-ID Lisboa
{wang.ling,tiago.luis,joao.graca}@inesc-id.pt
{luisa.coheur,isabel.trancoso}@inesc-id.pt

Abstract

In most statistical machine translation systems, the phrase/rule extraction algorithm uses alignments in the 1-best form, which might contain spurious alignment points. The usage of weighted alignment matrices that encode all possible alignments has been shown to generate better phrase tables for phrase-based systems. We propose two algorithms to generate the well known MSD reordering model using weighted alignment matrices. Experiments on the IWSLT 2010 evaluation datasets for two language pairs with different alignment algorithms show that our methods produce more accurate reordering models, as can be shown by an increase over the regular MSD models of 0.4 BLEU points in the BTEC French to English test set, and of 1.5 BLEU points in the DIALOG Chinese to English test set.

1 Introduction

The translation quality of statistical phrase-based systems (Koehn et al., 2003) is heavily dependent on the quality of the translation and reordering models generated during the phrase extraction algorithm (Ling et al., 2010). The basic phrase extraction algorithm uses word alignment information to constraint the possible phrases that can be extracted. It has been shown that better alignment quality generally leads to better results (Ganchev et al., 2008). However the relationship between the word alignment quality and the results is not straightforward, and it was shown in (Vilar et al., 2006) that better alignments in terms of F-measure do not always lead to better translation quality.

The fact that spurious word alignments might occur leads to the use of alternative representations for word alignments that allow multiple alignment hypotheses, rather than the 1-best alignment (Venugopal et al., 2009; Mi et al., 2008; Christopher Dyer et al., 2008). While using n-best alignments yields improvements over using the 1-best alignment, these methods are computationally expensive. More recently, the method described in (Liu et al., 2009) produces improvements over the methods above, while reducing the computational cost by using weighted alignment matrices to represent the alignment distribution over each parallel sentence. However, their results were limited by the fact that they had no method for extracting a reordering model from these matrices, and used a simple distance-based model.

In this paper, we propose two methods for generating the MSD (Mono Swap Discontinuous) reordering model from the weighted alignment matrices. First, we test a simple approach by using the 1-best alignment to generate the reordering model, while using the alignment matrix to produce the translation model. This reordering model is a simple adaptation of the MSD model to read from alignment matrices. Secondly, we develop two algorithms to infer the reordering model from the weighted alignment matrix probabilities. The first one uses the alignment information within phrase pairs, while the second uses contextual information of the phrase pairs.

This paper is organized as follows: Section 2 describes the MSD model; Section 3 presents our two algorithms; in Section 4 we report the results from the experiments conducted using these algorithms,

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 450–454,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

and comment on the results; we conclude in Section 5.

2 MSD models

Moses (Koehn et al., 2007) allows many configurations for the reordering model to be used. In this work, we will only refer to the default configuration (msd-bidirectional-fe), which uses the MSD model, and calculates the reordering orientation for the previous and the next word, for each phrase pair. Other possible configurations are simpler than the default one. For instance, the monotonicity model only considers monotone and non-monotone orientation types, whereas the MSD model also considers the monotone orientation type, but distinguishes the non-monotone orientation type between swap and discontinuous. The approach presented in this work can be adapted to the other configurations.

In the MSD model, during the phrase extraction, given a source sentence S and a target sentence T, the alignment set A, where a_i^j is an alignment from i to j, the phrase pair with words in positions between i and j in S, S_i^j, and n and m in T, T_n^m, can be classified with one of three orientations with respect to the previous word:

- The orientation is monotonous if only the previous word in the source is aligned with the previous word in the target, or, more formally, if $a_{i-1}^{n-1} \in A \wedge a_{j+1}^{n-1} \notin A$.
- The orientation is swap, if only the next word in the source is aligned with the previous word in the target, or more formally, if $a_{j+1}^{n-1} \in A \wedge a_{i-1}^{n-1} \notin A$.
- The orientation is discontinuous if neither of the above are true, which means, $(a_{i-1}^{n-1} \in A \wedge a_{j+1}^{n-1} \in A) \vee (a_{i-1}^{n-1} \notin A \wedge a_{j+1}^{n-1} \notin A)$.

The orientations with respect to the next word are given analogously. The reordering model is generated by grouping the phrase pairs that are equal, and calculating the probabilities of the grouped phrase pair being associated each orientation type and direction, based on the orientations for each direction that are extracted. Formally, the probability of the phrase pair p having a monotonous orientation is given by:

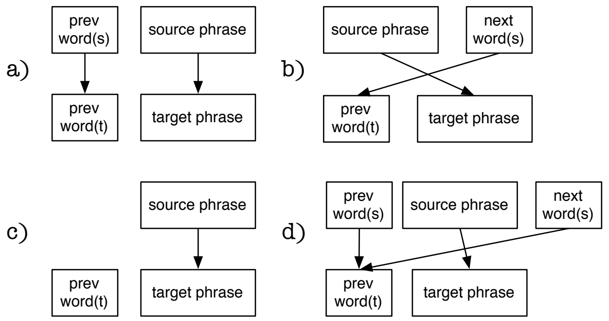

Figure 1: Enumeration of possible reordering cases with respect to the previous word. Case a) is classified as monotonous, case b) is classified as swap and cases c) and d) are classified as discontinuous.

$$P(p, mono) = \frac{C(mono)}{C(mono)+C(swap)+C(disc)} \qquad (1)$$

Where $C(o)$ is the number of times a phrase is extracted with the orientation o in that group of phrase pairs. Moses also provides many options for this stage, such as types of smoothing. We use the default smoothing configuration which adds the fixed value of 0.5 to all $C(o)$.

3 Weighted MSD Model

When using a weighted alignment matrix, rather than working with alignments points, we use the probability of each word in the source aligning with each word in the target. Thus, the regular MSD model cannot be directly applied here.

One obvious solution to solve this problem is to produce a 1-best alignment set along with the alignment matrix, and use the 1-best alignment to generate the reordering model, while using the alignment matrix to produce the translation model. However, this method would not be taking advantage of the weighted alignment matrix. The following subsections describe two algorithms that are proposed to make use of the alignment probabilities.

3.1 Score-based

Each phrase pair that is extracted using the algorithm described in (Liu et al., 2009) is given a score based on its alignments. This score is higher if the alignment points in the phrase pair have high probabilities, and if the alignment is consistent. Thus, if an extracted phrase pair has better quality, its orientation should have more weight than phrase pairs with worse quality. We implement this by changing the $C(o)$ function in equation 1 from being the number of the phrase pairs with the orientation o, to the sum of the scores of those phrases. We also need to normalize the scores for each group, due to the fixed smoothing that is applied, since if the sum of the scores is much lower (e.g. 0.1) than the smoothing factor (0.5), the latter will overshadow the weight of the phrase pairs. The normalization is done by setting the phrase pair with the highest value of the sum of all MSD probabilities to 1, and readjusting other phrase pairs accordingly. Thus, a group of 3 phrase pairs that have the MSD probability sums of 0.1, 0.05 and 0.1, are all set to 1, 0.5 and 1.

3.2 Context-based

We propose an alternative algorithm to calculate the reordering orientations for each phrase pair. Rather than classifying each phrase pair with either monotonous (M), swap (S) or discontinuous (D), we calculate the probability for each orientation, and use these as weighted counts when creating the reordering model. Thus, for the previous word, given a weighted alignment matrix W, the phrase pair between the indexes i and j in S, S_i^j, and n and m in T, T_n^m, the probability values for each orientation are given by:

- $P_c(M) = W_{i-1}^{n-1} \times (1 - W_{j+1}^{n-1})$
- $P_c(S) = W_{j+1}^{n-1} \times (1 - W_{i-1}^{n-1})$
- $P_c(D) = W_{i-1}^{n-1} \times W_{j+1}^{n-1}$ $+ (1 - W_{i-1}^{n-1}) \times (1 - W_{j+1}^{n-1})$

These formulas derive from the adaptation of conditions of each orientation presented in 2. In the regular MSD model, the previous orientation for a phrase pair is monotonous if the previous word in the source phrase is aligned with the previous word in the target phrase and not aligned with the next word. Thus, the probability of a phrase pair to have a monotonous orientation $P_c(M)$ is given by the probability of the previous word in the source phrase being aligned with the previous word in the target phrase W_{i-1}^{n-1}, and the probability of the previous word in the source to not be aligned with the next word in the target $(1 - W_{j+1}^{n-1})$. Also, the sum of the probabilities of all orientations ($P_c(M)$, $P_c(S)$, $P_c(D)$) for a given phrase pair can be trivially shown to be 1. The probabilities for the next word are given analogously. Following equation 1, the function $C(o)$ is changed to be the sum of all $P_c(o)$, from the grouped phrase pairs.

4 Experiments

4.1 Corpus

Our experiments were performed over two datasets, the BTEC and the DIALOG parallel corpora from the latest IWSLT evaluation 2010 (Paul et al., 2010). BTEC is a multilingual speech corpus that contains sentences related to tourism, such as the ones found in phrasebooks. DIALOG is a collection of human-mediated cross-lingual dialogs in travel situations. The experiments performed with the BTEC corpus used only the French-English subset, while the ones perfomed with the DIALOG corpus used the Chinese-English subset. The training corpora contains about 19K sentences and 30K sentences, respectively. The development corpus for the BTEC task was the CSTAR03 test set composed by 506 sentences, and the test set was the IWSLT04 test set composed by 500 sentences and 16 references. As for the DIALOG task, the development set was the IWSLT09 devset composed by 200 sentences, and the test set was the CSTAR03 test set with 506 sentences and 16 references.

4.2 Setup

We use weighted alignment matrices based on Hidden Markov Models (HMMs), which are produced by the the PostCAT toolkit[1], based on the posterior regularization framework (V. Graça et al., 2010). The extraction algorithm using weighted alignment matrices employs the same method described in (Liu et al., 2009), and the phrase pruning threshold was set to 0.1. For the reordering model, we use the distance-based reordering, and compare the results with the MSD model using the 1-best alignment. Then, we apply our two methods based on alignment matrices. Finally, we combine our two methods above by adapting the function $C(o)$, to be the

[1] http://www.seas.upenn.edu/ strctlrn/CAT/CAT.html

sum of all $P_c(o)$, weighted by the scores of the respective phrase pairs. The optimization of the translation model weights was done using MERT, and each experiment was run 5 times, and the final score is calculated as the average of the 5 runs, in order to stabilize the results. Finally, the results were evaluated using BLEU-4, METEOR, TER and TERp. The BLEU-4 and METEOR scores were computed using 16 references. The TER and TERp were computed using a single reference.

4.3 Reordering model comparison

Tables 1 and 2 show the scores using the different reordering models. Consistent improvements in the BLEU scores may be observed when changing from the MSD model to the models generated using alignment matrices. The results were consistently better using our models in the DIALOG task, since the English-Chinese language pair is more dependent on the reordering model. This is evident if we look at the difference in the scores between the distance-based and the MSD models. Furthermore, in this task, we observe an improvement on all scores from the MSD model to our weighted MSD models, which suggests that the usage of alignment matrices helps predict the reordering probabilities more accurately.

We can also see that the context based reordering model performs better than the score based model in the BTEC task, which does not perform significantly better than the regular MSD model in this task. Furthermore, combining the score based method with the context based method does not lead to any improvements. We believe this is because the alignment probabilities are much more accurate in the English-French language pair, and phrase pair scores remain consistent throughout the extraction, making the score based approach and the regular MSD model behave similarly. On the other hand, in the DIALOG task, score based model has better performance than the regular MSD model, and the combination of both methods yields a significant improvement over each method alone.

BTEC	BLEU	METEOR	TERp	TER
Distance-based	61.84	65.38	27.60	22.40
MSD	62.02	65.93	27.40	22.80
score MSD	62.15	66.18	27.30	22.20
context MSD	**62.42**	**66.29**	**27.00**	**22.00**
combined MSD	**62.42**	66.14	27.10	22.20

Table 1: Results for the BTEC task.

DIALOG	BLEU	METEOR	TERp	TER
Distance-based	36.29	45.15	49.00	41.20
MSD	39.56	46.85	47.20	39.60
score MSD	40.2	47.16	46.52	38.80
context MSD	40.14	47.14	**45.88**	39.00
combined MSD	**41.03**	**47.69**	46.20	**38.20**

Table 2: Results for the DIALOG task.

Table 3 shows a case where the context based model is more accurate than the regular MSD model. The alignment is obviously faulty, since the word “two” is aligned with both “deux”, although it should only be aligned with the first occurrence. Furthermore, the word “twin” should be aligned with “à deux lit”, but it is aligned with “chambres”. If we use the 1-best alignment to compute the reordering type of the sentence pair “Je voudrais réserver deux” / “I’d like to reserve two”, the reordering type for the following orientation would be monotonous, since the next word “chambres” is falsely aligned with “twin”. However, it should clearly be discontinuous, since the right alignment for “twin” is “à deux lit”. This problem is less serious when we use the weighted MSD model, since the orientation probability mass would be divided between monotonous and discontinuous since the probability weighted matrix for the wrong alignment is 0.5. On the BTEC task, some of the other scores are lower than the MSD model, and we suspect that this stems from the fact that our tuning process only attempts to maximize the BLEU score.

5 Conclusions

In this paper we addressed the limitations of the MSD reordering models extracted from the 1-best alignments, and presented two algorithms to extract these models from weighted alignment matrices. Experiments show that our models perform better than the distance-based model and the regular MSD model. The method based on scores showed a good performance for the Chinese-English language pair, but the performance for the English-French pair was similar to the MSD model. On the other hand, the method based on context improves the results on

Alignment	Je	voudrais	réserver	deux	chambres	à	deux	lits	.
I	1								
'd		0.7							
like		0.7							
to									
reserve			1						
two				1			0.5		
twin					0.5			0.5	
rooms					1				
.									1

Table 3: Weighted alignment matrix for a training sentence pair from BTEC, with spurious alignment probabilities. Alignment points with 0 probabilities are left empty.

both pairs. Finally, on the Chinese-English test, by combining both methods we can achieve a BLEU improvement of approximately 1.5%. The code used in this work is currently integrated with the Geppetto toolkit[2] , and it will be made available in the next version for public use.

6 Acknowledgements

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007. The PhD thesis of Tiago Luís is supported by FCT grant SFRH/BD/62151/2009. The PhD thesis of Wang Ling is supported by FCT grant SFRH/BD/51157/2010. The authors also wish to thank the anonymous reviewers for many helpful comments.

References

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing Word Lattice Translation. Technical Report LAMP-TR-149, University of Maryland, College Park, February.

Kuzman Ganchev, João V. Graça, and Ben Taskar. 2008. Better alignments = better translations? In *Proceedings of ACL-08: HLT*, pages 986–993, Columbus, Ohio, June. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, 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 Ondrej Bojar. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Wang Ling, Tiago Luís, Joao Graça, Luísa Coheur, and Isabel Trancoso. 2010. Towards a general and extensible phrase-extraction algorithm. In *IWSLT '10: International Workshop on Spoken Language Translation*, pages 313–320, Paris, France.

Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. 2009. Weighted alignment matrices for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 1017–1026, Morristown, NJ, USA. Association for Computational Linguistics.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL-08: HLT*, pages 192–199, Columbus, Ohio, June. Association for Computational Linguistics.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *IWSLT '10: International Workshop on Spoken Language Translation*, pages 3–27.

João V. Graça, Kuzman Ganchev, and Ben Taskar. 2010. Learning Tractable Word Alignment Models with Complex Constraints. *Comput. Linguist.*, 36:481–504.

Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2009. Wider pipelines: N-best alignments and parses in MT training.

David Vilar, Maja Popovic, and Hermann Ney. 2006. Aer: Do we need to "improve" our alignments? In *International Workshop on Spoken Language Translation (IWSLT)*, pages 205–212.

[2] http://code.google.com/p/geppetto/