Using linear interpolation and weighted reordering hypotheses in the Moses system

Marta R. Costa-jussà[*], **José R. Fonollosa**[†]

[*]Barcelona Media Research Center
Barcelona, Spain
marta.ruiz@barcelonamedia.org
[†] Universitat Politècnica de Catalunya
Barcelona, Spain
adrian@gps.tsc.upc.edu

Abstract

This paper proposes to introduce a novel reordering model in the open-source Moses toolkit. The main idea is to provide weighted reordering hypotheses to the SMT decoder. These hypotheses are built using a first-step Ngram-based SMT translation from a source language into a third representation that is called *reordered source language*. Each hypothesis has its own weight provided by the Ngram-based decoder. This proposed reordering technique offers a better and more efficient translation when compared to both the distance-based and the lexicalized reordering. In addition to this reordering approach, this paper describes a domain adaptation technique which is based on a linear combination of an specific in-domain and an extra out-domain translation models. Results for both approaches are reported in the Arabic-to-English 2008 IWSLT task. When implementing the weighted reordering hypotheses and the domain adaptation technique in the final translation system, translation results reach improvements up to 2.5 BLEU compared to a standard state-of-the-art Moses baseline system.

1. Introduction

Statistical machine translation (SMT) constitutes a research sub-area of machine translation (MT) that has recently gained much popularity. In fact, this technology has experienced real growth motivated by the development of computer resources needed to implement translation algorithms based on statistical methods (Brown et al., 1993).

Nowadays, one of the most popular SMT approaches is the phrase-based system (Zens et al., 2002) using a combination of feature functions. The Moses system (Koehn et al., 2007) is an implementation of this phrase-based machine translation approach. An input sentence is first split into text chunks (so-called phrases), which are then mapped one-to-one to target phrases using a large phrase translation table. Phrases may be reordered, but typically a reordering limit is used.

Our ongoing efforts are mainly dedicated to finding the best way to reorder the source side of the bilingual corpus aiming to decrease the divergences in word order of the source and target languages. This is especially important when the translation is performed between pairs of languages with non-monotonic word order, like Arabic and English. Recent techniques propose modular approaches where reordering is faced before translation. This makes it possible to easily change the reordering strategy and to speed up translation because a montonic search is used. Using a statistical Ngram-based system, we propose to generate weighted reordering hypotheses and to introduce them as input graphs to the Moses system.

Another promising way to improve the quality of MT output is to involve additional out-of-domain parallel information into bilingual modeling. (Koehn and Schroeder, 2007) perform a log-linear combination of translation models. Inspired by the results presented in (Foster and Kuhn, 2007). we interpolate a principal translation model (TM) with a secondary one, adjusting the weight coefficients according to the corresponding monolingual language models.

This paper is organized as follows. Section 2. briefly describes several reordering approaches related to the one that we are proposing. Section 3. introduces the phrase-based system used throughout this paper. Section 4. reports the proposed Ngram-based reordering technique. Section 5. reports the experiments using the Arabic-to-English task and, finally, section 6. presents the conclusions.

2. Related reordering work

Many alternatives have been proposed on facing the reordering challenge. One simple model is a *'weak'* distance-based distortion model that was initially used to penalize the longest reorderings, only allowed if sufficiently promoted by the rest of models(Och and Ney, 2004; Koehn et al., 2003).

In view of content-independence of the distortion and flat reordering models, several researchers (Tillmann, 2004; Koehn et al., 2005) proposed a more powerful model called lexicalized reordering model that is phrase dependent. Lexicalized reordering model learns local orientations (monotone or non-monotone) with probabilities for each bilingual phrase from training data. During decoding the model attempts to finding a Viterbi local orientation sequence. Performance gains have been reported for systems with lexicalized reordering model. However, since reorderings are related to concrete phrases, researchers have to design their systems carefully in order not to cause other problems, e.g. the data sparseness problem.

Recently, (Crego and Mariño, 2007) employ POS tags to automatically learn reorderings in training. They allow all possible learned reorderings to be used to create a lattice that is input to the decoder, which further improves translation accuracy. (Zhang et al., 2007) describe a similar approach using unlexicalized context-free chunk tags (XPs) to learn reordering rules for Chinese-English SMT. Similarly, to the last two approaches we employ a word graph for coupling reordering and decoding. In our approach, this word graph is built using an Ngram-based reordering technique (Costa-jussà and Fonollosa, 2009) which confronts the reordering challenge using the powerful statistical machine translation techniques.

3. Phrase-based SMT system

The basic idea of phrase-based translation is to segment the given source sentence into units (hereafter called phrases), then translate each phrase and finally compose the target sentence from these phrase translations.

Basically, a bilingual phrase is a pair of m source words and n target words. For extraction from a bilingual word aligned training corpus, two additional constraints are considered:

1. the words are consecutive, and,

2. they are consistent with the word alignment matrix.

Given the collected phrase pairs, the phrase translation probability distribution is commonly estimated by relative frequency in both directions.

The translation model is combined together with six additional feature models: the target language model, the word and the phrase bonus and the source-to-target and target-to-source lexicon model and the reordering model. These models are optimized in the decoder following the procedure described in *http://www.statmt.org/moses/*.

3.1. Translation model interpolation

Due to a small amount of available in-domain data (IWSLT training material), we have used an out-of-domain 130K-line subset from the NIST 2008 parallel corpus (*VIOLIN*) (Habash, 2007) to increase the final translation and language model. Both corpus statistics can be found in Table 1.

3.1.1. Combined training data

The straightforward way is to simply concatenate the two training corpora and use the combined data for both translation model and language model training. However, in case the in-domain data is a much smaller set than the out-domain corpus, the gain expected through a simply concatenation is not much. The result can be even worse when the in-domain data has a very particular style.

3.1.2. Translation and language model interpolation

Another proposal is to implement a TM interpolation strategy following the ideas proposed in (Schwenk and Estève, 2008), where the authors present a promising technique of target language models linear interpolation. These findings open the way to involve additional monolingual information into the translation process, and also gives a motivation to interpolate the translation tables in a linear way.

Instead of time-consuming iterative TM reconstruction and using the highest BLEU score as an maximization criterion, we follow the next procedure:

1. Find the optimal weights on the language model derived from the in-domain data and the language model derived from the out-of-domain data, using perplexity as the criterion.

2. Use those exact same two weights for the corresponding two translation models and two reordering models.

The word-to-word alignment was obtained from the joint database (IWSLT + VIOLIN). Then, we separately computed the translation tables corresponding to the IWSLT and VIOLIN parts of the joint alignment. The final tables, as well as the final target language model were obtained using linear interpolation. The weight coefficients (IWSLT weight = 0.95, VIOLIN weight = 0.05) were selected using a minimum perplexity criterion estimated on the corresponding interpolated combination of the target-side LMs.

4. Ngram-based Reordering Approach

As mentioned in the introduction, the weighted reordering hypotheses are generated with an Ngram-

	IWSLT		VIOLIN	
	Arabic	English	Arabic	English
Sentences	24.45 K	24.45 K	130.59 K	130.59 K
Words	170.24 K	188.54 K	4.12 M	4.44 M
Average sentence length	6.96	7.71	31.52	34.01
Vocabulary	10.89 K	6.92 K	72.9 K	65.9 K

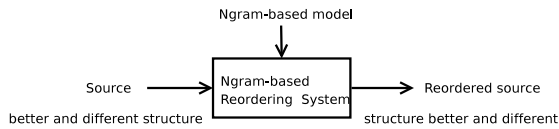Table 1: *The main and additional basic corpora statistics.*

Figure 1: *Ngram-based reordering approach.*

based statistical approach. The aim of the Ngram-based reordering approach (also called statistical machine reordering, SMR) consists in using an SMT system to deal with reordering problems. Therefore, the reordering hypotheses are built with an SMT system which translates from an original source language (*S*) to a reordered source language (*S'*), given a target language (*T*).

Figure 1 shows an example of the Ngram-based reordering system which translates from English to a reordered English given the Spanish as target language.

4.1. Introduction to the Ngram-based translation model

The Ngram-based model used for reordering is inspired by the Ngram-based translation model (Mariño et al., 2006). This section is dedicated to make a brief summary of the Ngram-based translation model.

Differently to the phrase-based translation model, the Ngram-based translation model is trained on bilingual n-grams. This model constitutes a language model of a particular *"bi-language"* composed of bilingual units (translation units) which are referred to as tuples. In this way, the translation model probabilities at the sentence level are approximated by using n-grams of tuples, such as described by the following equation:

$$\hat{t}_1^I = \arg\max_{t_1^I}\{p(s_1^J, t_1^I)\} = \cdots = \qquad (1)$$

$$\arg\max_{t_1^I}\{\prod_{n=1}^{N} p((s,t)_n | (s,t)_{n-x+1}, \cdots, (s,t)_{n-1})\} \qquad (2)$$

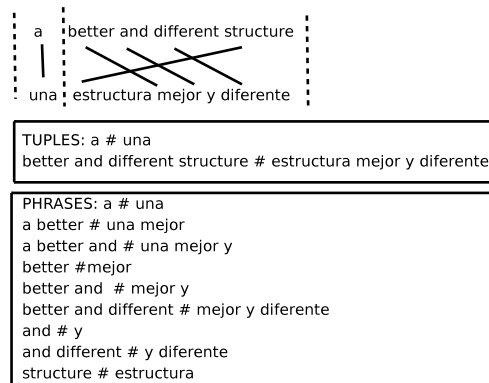where the *n*th tuple of a sentence pair is referred as $(s,t)_n$.

Figure 2: *Regular tuple extraction compared to phrase extraction.*

As any standard n-gram language model, the bilingual translation model is estimated over a training corpus composed of sentences in the language being modeled. In this case, we consider sentences in the *"bi-language"* previously introduced.

The Ngram-based approach is monotonic in that its model is based on the sequential order of tuples during training. Tuples are extracted from a word-to-word aligned corpus (see an example in Figure 2) in such a manner that a unique segmentation of the bilingual corpus is achieved.

- a monotonic segmentation of each bilingual sentence pair is produced,

- no word inside the tuple is aligned to words outside the tuple, and

- no smaller tuples can be extracted without violating the previous constraints.

4.2. Ngram-based reordering approach

Given a regular segmentation into tuples, the Ngram-based model is trained on reordering tuples (or reordering bilingual units). Notice that we make a difference between the Ngram-based model which is the model computed on the reordering tuples and the Ngram-based translation model which is computed on the regular tuples. Hereafter, in this paper, we are only
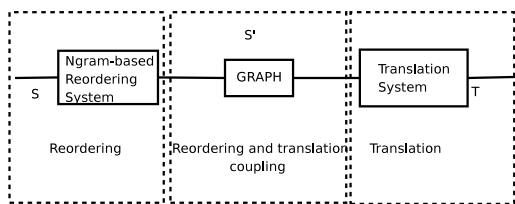
Figure 5: *Reordering and translation coupling.*

using the Ngram-based model which is used to compute reordering hypotheses.

The reordering information is extracted from the alignment as shown in Figure 3 (b). Finally, source words are replaced by source word classes as shown in Figure 3 (c). Therefore, the Ngram-based model is composed of bilingual n-grams and these are composed of the source word classes and the new reordering positions (Costa-jussà and Fonollosa, 2009). Translation from *S* to *S'* is computed as follows:

1. Source words are replaced by their corresponding word classes. In this paper, statistical classes (Och, 1999) are used.

2. Decoding using an Ngram-based model from the source word classes into reordered source positions that are the reordering hypotheses.

At work, the decoder builts a search graph. This graph offers several weighted reordering hypotheses of the source sentence. Note that each arch of the graph contains the weight given by the Ngram-based model. Figure 4 shows an example of reordering graph given the Spanish as a source language.

Finally, the weighted reordering graph is used as input of the monotonic SMT system. The final translation system coupled with the reordering system is shown in Figure 5.

5. Experiments

Experiments were run using the Basic Traveling Expression Corpus (BTEC) Arabic to English translation task used in the 2008 IWSLT Evaluation Campaign. Model weights were tuned with the 2006 IWSLT development corpus, containing 489 sentences and 6 reference translations. Experiments were tested on the 2008 official IWSLT evaluation test set both the CRR (Correct Recognition Results) and ASR (Automatic Speech Recognition) output [1], containing 507 sentences and 6 reference translations.

The phrase-based system used in this paper is based on the well-known Moses toolkit, which is nowadays

considered as a state-of-the-art SMT system (Koehn et al., 2007). The training and weights tuning procedures are explained in details in the above-mentioned publication, as well as, on the Moses web page: *http://www.statmt.org/moses/*.

5.1. Arabic data preprocessing

We used a similar approach to that shown in (Habash and Sadat, 2006), namely the MADA+TOKAN system for disambiguation and tokenization. For disambiguation only diacritic unigram statistics were employed. For tokenization we used the D3 scheme with -TAGBIES option. The scheme splits the following set of enclitics: w+, f+, b+, k+, l+, Al+ and pronominal enclitics. The -TAGBIES option produces Bies POS tags on all taggable tokens.

5.2. Reordering parameters

This section provides the details of the reordering parameters used for each technique.

The maximum distance of the words to be reordered was set to 6 in the distance-based reordering.

The parameters of the lexicalized reordering were settled as follows:

- The lexicalized distortion model was defined as *msd-bidirectional-fe*. *Msd* means the reordering types can be monotone, swap and discontinuous (Tillmann, 2004). *Bidirectional* means that certain phrases may not only flag, if they themselves are moved out of order, but also if subsequent phrases are reordered. *fe* concerns out of sparse data and the probability distribution conditions on the foreign phrase *f* and on the English phrase *e*.

- The maximum number of words to be reordered (max-skip) was set to 6.

Finally, the Ngram-based reordering approach used 100 statistical classes (Och, 1999). The Ngram-based model used a context of 3 bilingual units. The decoding limited the beam search to 5.

5.3. Baseline and interpolation performance

Table 2 shows results of the baseline and the interpolation experiments on the CRR and ASR sets. The baseline system is trained on the IWSLT data. Regarding the CRR task, training on all the data decreases performance (-0.4 BLEU points) for the CRR set and shows a slight improvement in the ASR task. Finally, the results show that the linear combination only in the language model is a useful tool for domain adaptation (+0.6 BLEU points) and even more useful is to

---

(A) BILINGUAL S2T REGULAR TUPLE:

better and different structure # estructura mejor y differente # 1-2 2-3 3-4 4-1

(B) BILINGUAL S2S' REORDERING TUPLE:

better and different structure # 4 1 2 3

(C) CLASS REPLACING:

C36 C88 C185 C176 # 4 1 2 3

Figure 3: *Example of the extraction of reordering bilingual units. In (a) '#' divides the fields: source, target and word alignment, which includes the source and final position separated by '-'. In (b) and (c)'#' divides source and reordering positions.*
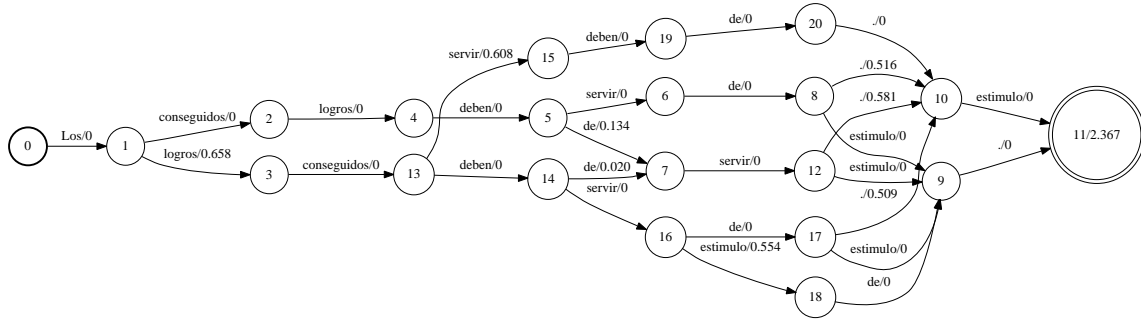

Figure 4: *Weighted reordering graph. The source sentence is:* Los logros conseguidos deben servir de estímulo. *The target sentence could be:* The achieved goals should be an encouragement.

System	BLEU	METEOR	NIST
CRR			
In-d	52.6	68.5	8.59
Ctd	52.2	68.1	8.52
Int lm	53.2	69.01	8.69
Int	**54.2**	**69.7**	**8.87**
ASR			
In-d	43.5	62.9	7.28
Ctd	43.8	62.6	7.28
Int lm	44.20	62.93	7.31
Int	**45.6**	**63.8**	**7.61**

Table 2: *Results of domain adaptation experiments: in-domain data (In-d), combined training data (Ctd), language model interpolation (Int lm) and translation and language model interpolation (Int).*

Reord.	BLEU	METEOR	NIST	SPEED
CRR				
db	52.5	68.4	8.53	27.8
lex	52.6	68.5	**8.59**	19.9
NbR	**53.1**	**68.7**	**8.59**	**31.0**
ASR				
db	43.9	62.8	7.19	27.8
lex	43.5	62.9	7.28	19.9
NbR	**44.4**	**63.3**	**7.35**	**31.0**

Table 3: *Results of using different reordering techniques: distance-based (db), lexicalized (lex) and Ngram-based (NbR). Speed is shown in words per second.*

use the linear combination both in the translation and language model (+1.6 BLEU points).

Regarding the ASR task, the results with linear interpolation of the translation and language model are even better (+2.1 BLEU points).

5.4. Reordering performance

Table 3 shows results of the reordering approach experiments on the CRR and ASR sets. When comparing to the distance-based reordering, the Ngram-based reordering technique achieves an improvement

of 0.6 BLEU on the CRR task and 0.5 BLEU on the ASR task. When comparing to the lexicalized reordering, the Ngram-based reordering technique achieves an improvement of 0.5 BLEU on the CRR task and 0.9 BLEU on the ASR task.

Table 6 shows some differences regarding the word ordering when using different reordering techniques.

One main advantage of the Ngram-based reordering model is being capable of generalizing reorderings that were not seen during training because it uses word classes. This may have a higher influence on the ASR task.

Furthermore, the SMT translation is much more effi-

System	BLEU	METEOR	NIST
CRR			
lex+int	54.2	69.7	8.81
NbR+int	**54.5**	**69.9**	**8.93**
ASR			
lex+int	45.6	63.8	**7.61**
NbR+int	**46.0**	**64.0**	7.59

Table 4: *Results of domain adaptation experiments using different reordering techniques.*

cient when using the novel approach because the input reordering graph can be highly pruned by using a more constrained search in the Ngram-based decoder (in our case a beam search of 5) without affecting the translation quality.

Table 4 shows results on the CRR and ASR sets of the interpolation using the lexicalized and the Ngram-based reordering techniques.

Using the Ngram-based reordering technique reaches an improvement of 0.3 and 0.4 BLEU points.

6. Conclusions

This paper presented two main contributions.

First, a novel reordering approach based on the generation of weighted reordering hypotheses that was implemented in the open Moses toolkit. Translation results show that this new technique outperforms the lexicalized reordering approach both in translation quality and efficiency. NbR yields an improvement of 0.3-0.9 BLEU points over the lexicalized reordering (implemented in Moses) in the Arabic-to-English 2008 IWSLT task.

Second, a domain adaptation technique based on linear interpolation of the translation and language models. This leads to better translation performance (up to 2 BLEU) when interpolating the IWSLT and a subset from the NIST parallel corpus.

Both techniques improvements almost add up reaching a global increment of 2.5 BLEU on the ASR task.

7. Acknowledgments

8. References

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

M.R. Costa-jussà and J.A.R. Fonollosa. 2009. An ngram-based reordering model. *Computer Speech and Language*, 23:362–375.

J.M. Crego and J.B. Mariño. 2007. Improving SMT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

G. Foster and R. Kuhn. 2007. Mixture-model adaptation for smt. In *Annual Meeting of the Association for Computational Linguistics: Proc. of the Second Workshop on Statistical Machine Translation (WMT)*, Prague, Czech Republic, June.

N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52, New York City, USA, June.

N. Habash. 2007. Syntactic preprocessing for statistical machine translation. *Proc. of the MT-Summit XI*, September.

P. Koehn and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics: Proc. of the Second Workshop on Statistical Machine Translation (WMT)*, pages 224–227, Prague, June.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Human Language Technology Conf., HLT-NAACL'03*, pages 48–54, Edmonton, Canada, May.

P. Koehn, A. Amittai, A. Birch, C. Callison-Burch, M. Osborne, D. Talbot, and M. White. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proc. of International Workshop on Spoken Languages Translation*, Pittsburgh, October.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic, June.

J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December.

F.J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, December.

DISTANCE-BASED:	Is there any place shopping?
LEXICALIZED:	Is there any place shopping?
Ngram-based:	Is there **any place to go shopping?**
DISTANCE-BASED:	Could you tell me how to get to the market Central?
LEXICALIZED:	Could you tell me how to get to the market Central?
NGRAM-BASED:	Could you tell me how to get to the **Central market**?
DISTANCE-BASED:	Are there any hospital emergency around?
LEXICALIZED:	Are there any hospital emergency around?
NGRAM-BASED:	Are there any **emergency hospital** around?

Figure 6: *Translation examples using different reordering techniques.*

F.J. Och. 1999. An efficient method for determining bilingual word classes. In *Proc. of the 9th Conf. of the European Chapter of the Association for Computational Linguistics*, pages 71–76, June.

H. Schwenk and Y. Estève. 2008. Data selection and smoothing in an open-source system for the 2008 NIST machine translation evaluation. In *Proceedings of the Interspeech'08*, Brisbane, Australia.

C. Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proc. of the Human Language Technology Conf., HLT-NAACL'04*, pages 101–104, Boston, May.

R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In M. Jarke, J. Koehler, and G. Lakemeyer, editors, *KI - 2002: Advances in artificial intelligence*, volume LNAI 2479, pages 18–32. Springer Verlag, September.

Y. Zhang, R. Zens, and H. Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL'06):Proc. of the Workshop on Syntax and Structure in Statistical Translation (SSST)*, pages 1–8, Rochester, April.