Learning Improved Reordering Models for Urdu, Farsi and Italian using SMT

Rohit GUPTA[1] *Raj Nath PATEL*[1] *Ritesh SHAH*[1]
(1) CDAC Mumbai, Gulmohar Cross Road No. 9, Juhu, Mumbai, India
rohitg@cdac.in, rajnathp@cdac.in, ritesh@cdac.in

ABSTRACT

This paper presents some experiments which have been carried out as part of a shared task for the workshop "Reordering for Statistical Machine Translation" (RSMT, collocated with COLING 2012). The shared task objective is to learn reordering models by making use of a manually word-aligned, bilingual parallel data. We view this task as that of a statistical machine translation (SMT) system which implicitly employ such models. These models are obtained using empirical methods and machine learning techniques. We have therefore used "Moses"; a state of the art SMT system to conduct experiments for the task at hand. The training and the development datasets used for the experiments have been provided by RSMT and we report our work on three pair of languages namely English-Urdu, English- Farsi and English- Italian.

KEYWORDS : reordering, factored, alignment, statistical, SMT, Moses, BLEU, GIZA++, Urdu, Farsi, Persian, Italian

Proceedings of the Workshop on Reordering for Statistical Machine Translation, pages 37–46,
COLING 2012, Mumbai, December 2012.

1 Introduction

The objective of the shared task is to learn reordering models by making use of human-annotated parallel data which is word aligned. We have transformed this task into one of empirical machine translation where model parameters for the system are learnt using parallel training data and machine learning techniques.

A statistical machine translation (SMT) system primarily relies on two models viz. the translation model (TM) and the language model (LM). In essence, it involves learning mutual correspondences using bilingual parallel data and reducing divergences among the source-target language pair. The alignment models which help establish such links, and the reordering models which help reduce the word order differences in the source-target pair constitute a part of the TM and are an implicit part of such an SMT system. Thus, our motivation to use the SMT system for the shared task comes from the alignment model GIZA++ (Och and Ney, 2003) and the basic reordering model (distance based distortion) employed in the Moses (Koehn et al., 2007) framework. Section 2 briefly explains the GIZA++ alignment model and Section 3 elaborates on the reordering model.

Across many language pairs, the existing SMT systems are usually infested with a lack of resources which leads to reduced annotations on the source and/or target side. Use of machine learning techniques, data preprocessing or other heuristics is mostly employed to overcome this lack of information and estimate good translation models. However, the training data provided in this shared task has the necessary alignment information on both sides. Availability of such information initially motivated us to use simple techniques of chunking based on source-target index information thereby modeling large distance word movements. However, we failed in these initial experiments which resulted in even lesser scores than those trained on a phrase based baseline system.

Therefore, the experiments were planned with only a scope of

1. training a baseline phrase based translation model; elaborated in Section 4.
2. training a factored translation model (Koehn and Hoang, 2007) with linguistic annotations as factors; explained in Section 5.

The BLEU (Papineni et al., 2001) score was the evaluation metric chosen to compare various results. The experiments and results are detailed in Section 6, followed by conclusions.

2 Alignment model

GIZA++is an open source toolkit used to train IBM Models 1-5 (Brown et al., 1993) and an HMM word alignment model (Vogel et al., 1996).

Given a source string $s_1^J = s_1, \ldots, s_j, \ldots, s_J$ and a target string $t_1^I = t_1, \ldots, t_i, \ldots, t_I$

An alignment A of the two strings is defined as

$$A \subseteq \{(j, i): j = 1, \ldots, J; i = 0, \ldots, I\}$$

In statistical word alignment, the probability of a source sentence given target sentence is written as:

$$P(s_1^J|t_1^I) = \sum_{a_1^J} P(s_1^J, a_1^J|t_1^I)$$

where a_1^J denotes the alignment across the sentence pair. Expressing the probability in statistical terms leads to $P(s_1^J, a_1^J|t_1^I) = p_\theta(s_1^J, a_1^J|t_1^I)$

The parameters θ can be estimated using maximum likelihood estimation (MLE) on a training corpus. If a corpus has N sentences

$$\hat{\theta} = \text{argmax}_\theta \prod_{n=1}^{N} \sum_a p_\theta(s_n, a|t_s)$$

The best alignment of a sentence pair is given by

$$\hat{a}_1^J = \text{argmax}_{a_1^J} p_{\hat{\theta}}(s_1^J, a_1^J|t_1^I)$$

When estimating the parameters, the Expectation-Maximization (Dempster et al., 1977) algorithm is employed. In the E-step the counts for all the parameters are collected, and the counts are normalized in M-step. Figure 1 shows a high-level view of GIZA++.

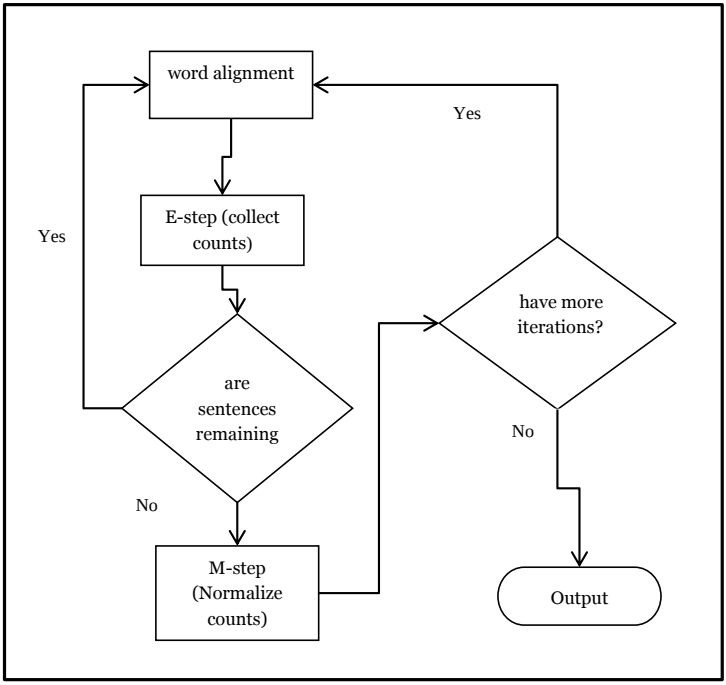

FIGURE 1 - GIZA++ algorithm overview

3 Distortion limit

Moses handles the reordering task using a reordering model (Koehn et al., 2003). This model is language independent and introduces a penalty when phrases are picked out of order. This penalty depends on the number of words skipped and is modeled by a linear distortion model given by

$$D(s,t) = p_f^1 + \sum_{i=2}^{p} d(i)$$

where P is the no. of phrases used to translate source s to target t, p_f^1 is the first word index of the source in first phrase and $d(i)$ is the distortion for phrase i given by

$$d(i) = |p_l^{i-1} + 1 - p_f^i|$$

where p_l^{i-1} is the last index of the source in previous phrase $i-1$ and p_f^i is the first index of the source in the current phrase i.

4 Phrase based model

Moses requires two types of data for training a phrase based model. Sentence aligned bilingual corpus to train the TM and the target side monolingual corpus for the LM. The TM presents itself in the form of a phrase-table which contains phrase entries and probabilities representing their mutual translation scores. The LM, however, represents the target language word order thereby ensuring good scores for a fluent output. A decoder component consults both the models to generate a sequence of phrases for a given test input.

In our experiment, we place this as a baseline system. Particularly, for the shared task data, the phrase based model has the following advantages:

- the same source and target language vocabulary can lead to a lesser sparse and improved translation model
- one to one mapping between source and target language words can result in a better alignment model

5 Factored model

Factored models are an extension of the phrase based models as they allow addition of factors to the training data. These factors could be linguistic annotations such as part-of-speech tags or any other relevant information used to improve the various models.
These factors are combined using a log-linear model given by the following equation.

$$p(t|s) = \frac{1}{Z} e^{\sum_{i=1}^{n} \lambda_i h_i(t,s)}$$

Each h_i is a feature function for a component of the translation and the values λ_i are the corresponding weights for the feature functions.
In the training data, each word is represented as a vector of factors, instead of a simple token. A phrase mapping is decomposed into several steps that either translate input factors to output factors or generate target factors to other target factors.

6 Experiments and results

The focus of this task as mentioned above is to learn the alignment information from the training data. Since the given data was in the CoNLL-X format, some preprocessing was done to obtain sentence aligned source and target data files for all language pairs. A distinct pair of source and target files was created for sentences containing indices, surface word forms and part-of-speech (POS) forms. In order to run trials on the phrase based and factored model, the data was split as per Table 1 below

	Numbers of sentences		
	English-Urdu	**English-Farsi**	**English-Italian**
Train	4500	3500	4500
Test	500	500	500
Development (Tuning)	500	500	500

TABLE 1 - Training, Testing and Development Data

A pair of trials was conducted with phrase based and factor based approach each with default parameters and tuned parameters.

The Moses default setting sets the distortion limit to 6. Therefore, if no. of words skipped is greater than 6 the translation will be pruned. This puts hard constraint and makes the model less suitable for more syntactically divergent languages like Urdu, Hindi, and Marathi etc. According to the choice of parameters, the correct reordering is sometimes improbable for large scale reordering. Thus, we have varied the distortion limits from 3 to 12 and observed the results for all trials.

Surface word form training was done for phrase based TM. We trained this baseline system with the original source sentences and the target reordered sentences.

For the factor based TM, we used a training data containing the surface form word and a POS tag (as an additional factor). Additionally, the training script included a mapping for translation-factors.

Translation-factor mapping:

[source side surface] to [target side surface + target side POS]

The format for factored model training data is as given below:

source format:
"a|DT developed|JJ pakistan|NN began|VBD taking|VBG shape|NN again|RB .|."

target format (reordered):
"a|DT developed|JJ pakistan|NN again|RB shape|NN taking|VBG began|VBD .|."

In terms of language model, surface word form LM was used for phrase based approach.

For the factored model, however, surface word form LM and POS based LM were used because better estimates of the target language order are provided by the POS LM. In comparison with the surface LM, the POS LM proves to be more useful on account of learning from a more generic target word order and richer evidences.

6.1 Experiments: without tuning

The results of the experiments on both approaches were evaluated for two test datasets. The first test dataset (test1) was obtained from splitting the provided data (ref. Table 1) and the second test dataset (test2) is the same on which the task results were announced. For BLEU evaluation and comparison, we requested the reference set for *test2* from the RSMT organizers. The results without tuning and with default parameter settings of Moses are shown in Table 2 below

	BLEU score phrase based model		**BLEU score factored model**	
	test1	**test2**	**test1**	**test2**
Urdu	42.59	42.21	44.54	44.07
Farsi	57.76	54.78	57.95	55.77
Italian	74.05	73.91	73.93	73.37

TABLE 2 - BLEU scores: default settings for phrase based and factored model

6.2 Experiments: with tuning

For tuning, the development data of 500 sentences was used. We evaluated results for varying distortion limit values after tuning. The motivation for this comes from the fact that the distance-based distortion model is placed as a weak model for highly divergent languages and our task is to learn reordering using alignments. Hence, the evaluation results might inform about the extent of reordering expected by each language pair.

	Urdu				**Farsi**				**Italian**			
Distortion limit	**Phrase based**		**Factored**		**Phrase based**		**Factored**		**Phrase based**		**Factored**	
	test1	**test2**	**test1**	**test2**	**test1**	**test2**	**test1**	**test2**	**test1**	**test2**	**test1**	**test2**
3	41.09	41.31	40.97	40.99	59.26	56.51	60.70	**57.89**	75.51	75.34	75.81	75.62
4	42.80	43.00	43.02	43.05	**59.67**	56.72	60.92	57.71	75.58	75.43	75.90	75.75
5	43.77	45.05	45.30	45.17	59.50	**56.87**	60.78	57.69	75.57	75.48	75.94	**75.80**
6 (def.)	44.32	45.12	45.96	45.58	59.51	56.76	61.05	57.55	**75.58**	**75.49**	75.94	75.78
7	45.38	45.60	47.26	46.36	59.56	56.71	**61.07**	57.78	**75.58**	**75.49**	75.94	75.78
8	45.67	45.98	47.51	46.97	59.56	56.74	**61.07**	57.66	**75.58**	75.47	75.94	75.76
9	**46.00**	**46.24**	**47.59**	47.97	59.51	56.64	60.98	57.48	75.56	75.44	**75.97**	75.74
10	45.40	45.76	46.94	**48.22**	59.05	56.35	60.70	57.38	75.46	75.30	75.96	75.63
11	45.23	45.64	46.49	47.95	58.68	55.83	60.39	57.17	75.21	75.00	75.79	75.18
12	44.85	44.78	46.05	47.74	57.78	54.93	60.03	56.51	74.80	74.49	75.45	74.75

TABLE 3-BLEU scores: after tuning and varying distortion limits

The evaluation results in the Table 3 for each pair of languages have been plotted below against varying distortion limit values. The dotted line in the plot represents phrase based values and the solid line represents the factor based values obtained after tuning. For a consistent comparison of the test results with that of the system, the scores obtained on the *test2* dataset are also plotted.

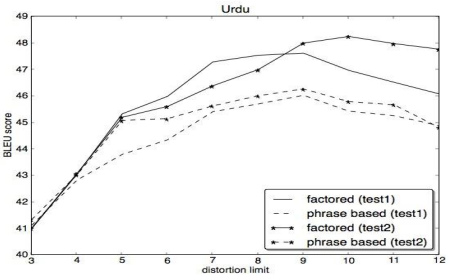

FIGURE 2 – BLEU score variation against distortion limit for Urdu

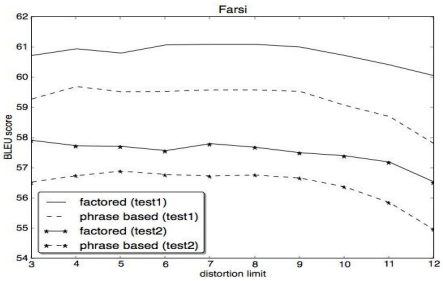

FIGURE 3 – BLEU score variation against distortion limit for Farsi

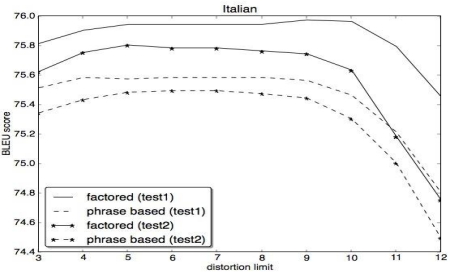

FIGURE 4 – BLEU score variation against distortion limit for Italian

6.3 Submission for the shared task

At the time of task submission the factored model with default settings was the best system we had. The mapping for translation factors was the same as described in Section 6. The system output for the test data provided by the organizers was obtained and eventually converted to the CoNLL-X format using some post-processing scripts. The results as provided by the organizers for the test corpus are given in the Table 4 below.

	Urdu	**Farsi**	**Italian**
BLEU score (our approach)	44.7	55.7	73.0
BLEU score (RSMT workshop baseline)	38.3	50.0	65.1

TABLE 4 - BLEU scores: test data results evaluated by the organizers

7 Conclusion and perspectives

With the default settings (before tuning) and for *test1*, factored model shows improvements for Urdu and Farsi pair only. However, English-Italian pair scores decrease slightly in the factored based approach. The same trend repeats for *test2* also. Apparently, the POS LM does not help the English-Italian pair with the default settings.

The Table 3 scores and graphs shown in Figure 2, 3 and 4 clearly show that the factored model (for *test1* and *test2*) outperforms the phrase based model for all languages after tuning is carried out. Although this varies for each language and the improvements are relatively high for Urdu and Farsi, only marginal improvements for Italian are observed.

More importantly, the plot for Urdu behaves sensitively for varying values of distortion limits. It begins to increase from a distortion limit value of 4 and attains a maximum at a value of 9 (for *test1*) and at 10 (for *test2*). The other languages do not vary highly against the distortion limit changes. Specifically, for *test2* the Urdu plot maintains good improvement even for distortion limit values of beyond 10. Evidently, this shows that Urdu prefers larger reordering and could be relatively more divergent.

The graphs also indicate a downward trend in scores for all languages from a distortion limit value of 10 onwards. The cause for this may be attributed to the increase in the number of translation choices during decoding, thereby increasing the error in selection of the correct hypotheses.

The plots for *test1* and *test2* follow the same trend in all cases except for Urdu factored, where BLEU score for *test2* does not drop heavily with increasing distortion limit values.

The results indicate that the shared task of learning reordering from the alignment information is modeled well by the approach as described above. This also resulted in improved BLEU scores over that of the baseline scores provided by RSMT.

References

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L., (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311

Dempster, A., Laird, N., and Rubin, D.,(1977).Maximum Likelihood From Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1):138

Koehn, P., and Hoang, H., (2007). Factored Translation Models. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 868–876, Prague, June 2007.

Koehn, P., Hoang, H.,Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E., (2007). Moses: Open source toolkit for statistical machine translation. *In ACL*, Demonstration Session.

Koehn, P., Och, F. J., and Marcu, D.,(2003). Statistical phrase based translation. *In NAACL*.

Och, F. J., and Ney, H., (2003).A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

Papineni, K., Roukos, S., Ward, T., and Zhu, W., (2001). BLEU: a method for automatic evaluation of machine translation. *IBM Research Report*, Thomas J. Watson Research Center.

Vogel, S., NeyH., and Tillmann, C., (1996). HMM-based Word Alignment in Statistical Translation. *In COLING: The 16th International Conference on Computational Linguistics,* pp. 836-841, Copenhagen, Denmark.