Whitepaper for Shared Task on Learning Reordering from Word Alignments at RSMT 2012

Mitesh M. Khapra[1] *Ananthakrishnan Ramanathan*[1]
Karthik Visweswariah[1]
(1) IBM Research India
{mikhapra,aramana5,v-karthik}@in.ibm.com

Abstract

Several studies have shown that the task of reordering source sentences to match the target order is crucial to improve the performance of Statistical Machine Translation, especially when the source and target languages have significantly divergent grammatical structures. In fact, it is now become a standard practice to include reordering as a pre-processing step or as an integrated module (within the decoder). However, despite the importance of this sub-task, there is no forum dedicated for its evaluation. The objective of the proposed Shared Task is to provide a common benchmarking platform to evaluate state of the art approaches for reordering.

Keywords: Reordering, Machine Translation.

1 Task Description

The task is to develop a reordering engine to reorder source English sentences to match the order of the target language. For example, the English (SVO language) sentence “Ram drinks water” is translated into Hindi (SOV language) as “Ram paanii piitaa hai (Ram water drinks)”. Thus, the correct reordering of this English sentence which matches the target (Hindi) order is “Ram water drinks”. The task organizers will release high-quality word-alignments (annotated by hand) between English and 3 languages (*viz.*, Urdu, Farsi and Italian). The participants can use this training and development data to develop a reordering engine for the mentioned source target language pairs. At evaluation time, a list of source sentences will be provided on which the participants will have to run their systems and submit the best reordering for each sentence as output by their system. For every language pair, the participants must submit at least one run which uses only the data provided by the task organizers. This will be called a “standard” run. Participants can submit more than one standard run. In addition, participants can also submit several “non-standard” runs for each language pair which use data other than that provided by the task organizers. The task organizers **will** differentiate between “standard” and “non-standard” runs while preparing the task report.

2 Important Dates

Research Paper Submission Deadline	08-Oct-2012 (23:59 PST)
Shared task	
Release Training/Development Data	10-Sep-2012
Release Test Data	04-Oct-2012
Results Submission Due	08-Oct-2012 (23:59 PST)
Results Announcement	10-Oct-2012
Task (short) Papers Due	15-Oct-2012
For All Submissions	
Acceptance Notification	01-Nov-2012
Camera-Ready Copy Deadline	10-Nov-2012 (23:59 PST)
Workshop Date	09-Dec-2012 (14:00 IST)

3 Data

Participants can register for the task by sending a mail to mikhapra@in.ibm.com and specifying the language pairs that they are interested in. The requested data containing the following files will be then mailed to the participants.

src_tgt.src.[trn|dev].conll : This file is in the standard CoNLL-X format with one word per line and a blank line separating two sentences. Some of the columns have been redefined to suit the reordering task. The columns are as follows:

1. Original index: The index of the word in the original unreordered source sentence
2. word : The lexical form of the word
3. empty : dummy column
4. CPOSTAG: Coarse-grained part-of-speech tag (tagset depends on the language)
5. POSTAG: Fine-grained part-of-speech tag (tagset depends on the language)
6. empty: dummy column

7. Previous Index: The index of the word which precedes this word in the reordered source sentence
8. empty : dummy column
9. empty : dummy column
10. empty : dummy column

Note that the words in the source sentence which do not align to any word in the target sentence will be dropped from the conll file. For example, if the source sentence is “I am going home” and if the word “a” is not aligned to any word in the target sentence then this word will be dropped from the conll file as shown below:

1	I	-	P	PRP	-	0	-	-	-
2	going	-	V	VBG	-	3	-	-	-
3	home	-	N	NOUN	-	1	-	-	-

src_trn.src.[trn|dev].txt : This file contains the complete source sentence (including words which were left unaligned) Example: I am going home.

src_trn.src.[trn|dev].pos : This file contains the pos tags for the complete source sentence (including words which were left unaligned). Example: VRB (I) VMZ (am) VBG (going) NN (home).

src_trn.src.[trn|dev].parse : This file contains a parse for the complete source sentence (including words which were left unaligned). The parse was generated by a state-of-the-art in-house parser.

src_trn.src.[trn|dev].align.info : This file contains indices of only those words which were aligned to some word in the target sentence. Example: 0(I) 2(going) 3 (home)

Note that src_tgt.src.[trn|dev].conll starts at index 1 whereas src_trn.src.[trn|dev].align.info starts at index 0. The participants can use src_tgt.src.[trn|dev].conll and src_trn.src.[trn|dev].align.info to find the words which were left unaligned.

3.1 Language pairs

Table 1 lists the language pairs that will be included in the Shared Task and the amount of hand aligned data that will be released for each language pair (**En**: English, **Fa**: Farsi, **Ur**: Urdu, **It**: Italian).

Language Pair	Train	Dev	Test
En-Fa	5K	500	500
En-Ur	5K	500	500
En-It	4K	500	500

Table 1: Language Pairs included in the Shared Task

3.2 Terms of usage

By requesting for the data the participants agree to the following:

1. using the dataset only for research purposes and not for any non-research/commercial purposes
2. submitting at least one run for the requested language pair
3. submitting a short paper describing their approach/system

Also note that the participants cannot redistribute the dataset in part or in whole nor can they republish it on any other site.

4 Submissions

At evaluation time, participants will be provided with test data containing the 4 files (conll, txt, pos, parse) described above. One "standard" run must be submitted by each group for each language pair. Additional "standard" runs (upto 4 in total can also be submitted). The best of the submitted "standard" runs will be used for reporting performance summary. In addition to "standard" runs the participants can also submit upto 4 "non-standard" runs. The results must be submitted in CoNLL-X format with the 7th column containing the previous index for each word as predicted by the participants system. There should be one conll file for every run. All the conll files should be zipped into a single zip file and mailed to mikhapra@in.ibm.com. The "standard" and "non-standard" runs must be labeled clearly.

4.1 Short Papers on Task

Each participating group is required to submit a short paper describing their approach. Participants should follow COLING 2012 paper submission policy including paper format, blind review policy and title and author format convention. The task paper should be a short paper containing 8 A5 sized pages with any number of reference pages.

5 Evaluation Metrics

The output reorderings will be evaluated using two metrics:

- **BLEU** (Papineni et al., 2001): In the past decade, BLEU has been the most widely used metric for MT evaluation. BLEU compares N-grams in the output translation and the reference translation(s), and uses a "brevity penalty" to prevent outputs that are accurate in terms of N-gram match, but too short.

  For reordering, we use the BLEU metric by comparing candidate reorderings with the reference reorderings that we create from the hand-alignments.

  BLEU is calculated as:

$$log(BLEU) = min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} \frac{1}{N} log(p_n)$$

$$\begin{aligned} where, N &= 4 \textit{(unigrams, bigrams, trigrams, and 4-grams are matched)} \\ r &= \textit{length of reference reordering} \\ c &= \textit{length of candidate reordering} \end{aligned}$$

  and

$$p_n = \frac{\sum_{C \in Candidates} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C \in Candidates} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}$$

where C runs over the entire set of candidate reorderings, and $Count_{clip}$ returns the number of n-grams that match in the reference reordering.

- **LRscore** (Birch and Osborne, 2010):

  LRscore was introduced a couple of years ago as a metric to directly measure reordering performance. LRscore uses a distance score in conjunction with BLEU to help evaluate the word order of MT outputs better. Experiments show that this combined metric correlates better with human judgments than BLEU alone (Birch and Osborne, 2010). Since we do not need a lexical metric, we use only the distance metric from LRscore. We will evaluate reordering distance using the following two scores:

  – Hamming distance: This measures the number of disagreements between two permutations:

  $$d_H(\pi, \rho) = 1 - \frac{\sum_{i=1}^{n} x_i}{n}, \text{ where } x_i = \begin{cases} 0 & \text{if } \pi(i) = \rho(i), \\ 1 & \text{otherwise,} \end{cases}$$

  – Kendall's Tau Distance: This measures the minimum number of transpositions of two adjacent symbols necessary to transform one permutation into another:

  $$d_r(\pi, \rho) = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij}}{Z}, \text{ where } z_{ij} = \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ and } \rho(i) > \rho(j) \\ 0 & \text{otherwise} \end{cases}$$

  $$Z = \frac{(n^2 - n)}{2}$$

  These two distance metrics will be combined with a brevity penalty (as defined in the description of BLEU above).

Links to these evaluation scripts are provided on the workshop webpage[1].

6 Baseline

The baseline score will be obtained by comparing the unreordered source sentence with the reference.

7 Some Pointers

We encourage participation from researchers in other areas such as parsing and language modeling, who may find reordering to be a good application area and extrinsic evaluation of their work. For such participants and others new to the problem of reordering and MT, the following pointers may be useful to get started with the task.

[1] https://sites.google.com/site/rsmt2012/Shared-Task/evaluation-scripts

- **Statistical MT toolkits**: Reordering can be thought of as translation from unreordered to reordered text. Setting up the unreordered text as the source corpus and the reordered text as the target corpus with publicly available MT toolkits like Moses (phrase-based MT toolkit: www.statmt.org/moses/), Hiero and Joshua could be a simple starting point for the task. We have observed improvements using Moses with the above setup, using the default settings, and only the target reordered corpus as the LM data. It should be possible to improve further with better features. For example, we could use a POS factor and a POS LM, or we could do some morphological processing to work with the roots and suffixes. It may also be important to tune various parameters, such as the distortion model parameters, which may be quite sensitive to the language pair.
- **Parser-based reordering**: If the source-language has a parser (we provide parses for the input sentences for the shared task), a few rules could be written for reordering (Collins et al., 2005) or rules could be automatically learnt based on the alignments (Visweswariah et al., 2010).
- **Reordering without a parser**: Some recent work has focussed on reordering without a parser. Examples are the word reordering models in Visweswariah et al. (2011) and Tromble et al. (2009). DeNero and Uszkoreit (2011) describe a technique to induce parse trees from alignments, and use these parses for reordering.

8 Contact Us

Name	Email
Mitesh M. Khapra	mikhapra@in.ibm.com
Ananthakrishnan Ramanathan	aramana5@in.ibm.com
Karthik Visweswariah	v-karthik@in.ibm.com

References

Birch, A. and Osborne, M., (2010). LRScore for evaluating lexical and reordering quality in MT. *Proceedings of the Fifth Workshop on Statistical Machine Translation.*

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of ACL*, pages 531–540, Morristown, NJ, USA. Association for Computational Linguistics.

DeNero, J. and Uszkoreit, J. (2011). Inducing sentence structure from parallel corpora for reordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 193–203, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Tromble, R., and Eisner, J., (2009). Learning linear ordering problems for better translation. In *Proceedings of EMNLP*.

Visweswariah, K., Navratil, J., Sorensen, J., Chenthamarakshan, V., and Kambhatla, N. (2010). Syntax based reordering with automatically derived rules for improved statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics.*

Visweswariah, K., Rajkumar, R., Gandhe, A., Ramanathan, A., and Navratil, J., (2011). A word reordering model for improved machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 486–496, Stroudsburg, PA, USA. Association for Computational Linguistics.