A Multi-system Translation Word-order Data Set

Colin Cherry and Xiaodong He
{colinc,xiaohe}@microsoft.com

May 2008

Technical Report
MSR-TR-2008-73

System combination is emerging as a powerful tool in statistical machine translation. In this document, we describe a data set designed to enable the study of a sub-problem of system combination: that of combining the word-order decisions made by several translation systems. We outline a data set derived from the MSR-NRC-SRI joint entry to the 2008 NIST machine translation evaluation, and briefly describe each translation system that contributes to the set.

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

1 Introduction

System combination is emerging as a powerful tool in statistical machine translation. We are interested in studying (and enabling the study) of an interesting sub-problem in this area: how can one combine the re-ordering decisions made by various translation systems? During translation, each system implicitly makes decisions on how the source sentence should be rearranged in target order, and invariably, different systems will disagree. Current system combination frameworks (Rosti et al., 2007; He et al., 2008) select one or more backbone translations, which then dictate word order for all of the other systems to be combined. Generally the backbone is selected to be the translation that is closest, on average, to the other systems. This approach follows the notion of Minimum Bayes Risk; however, alternatives are available. There is interest in the machine learning community in studying distributions and statistics measured over collections of permutations (Klementiev et al., 2007). Up until this point, these techniques have been applied to combining statistical ranking systems, but they could also benefit situations where the permutations in question are descriptions of how words move during translation. These techniques may allow a backbone to be selected more precisely by focusing specifically on word order, and they may enable the generation of a mean backbone, that does not exactly match any of the candidates being combined, but provides a synthesis of all the systems.

This document has three main purposes. The first is to describe the data provided in the multi-system word-order data set, and the second is to describe the various systems used in the creation of the machine translations. Finally and most importantly, these translations represent a large amount of work by a large body of contributors, and this document exists to ensure that all involved in the creation of this set receive credit for their part. We begin by describing the data, and then describe systems and contributors together.

2 Data Description

This data set describes the ordering decisions made by 8 single-system machine translation engines on the NIST 2008 Chinese-to-English Machine Translation Evaluation. The 8 systems are described in Section 3, and correspond to the 8 systems used in the MSR-NRC-SRI system combination entry to the NIST competition. For each system we provide a complete word alignment between system output and source sentence for each sentence in the test set. Word alignments are created using an HMM with word-dependent transition probabilities (He, 2007), trained on the entirety of the NIST 2008 Chinese-English constrained training data. To indicate bilingual connections between source and system translation, we opted to use a third-party word-alignment tool rather than system-specific back-tracking information because of the diversity of the systems involved.

Each alignment appears on its own line, with each link in the alignment

enclosed in round brackets, and separated by a single whitespace character. For example, the French-English sentence pair:

English	I [1] gave [2] it [3] to [4] him [5]
French	Je [1] le [3] lui [5] ai [2] donné [2]
Gloss	I it him past give

would appear in the data as:

```
(1,1) (2,3) (3,5) (4,2) (5,2)
```

which indicates that target position 1 was linked to source position 1, target position 2 was linked to source position 3, and so on.

We also provide sentence-specific target vocabulary IDs, so that researchers can tell when systems differ in lexical choice. We provide local, rather than global lexicon IDs to eliminate the chance that this data set could be used to reconstruct the source sentences or reference translations, which would violate LDC and NIST license agreements. Returning to our example, if we had two candidate translations with the same word order, but differing in the translation of "gave", the data would contain:

```
(1,1,[1]) (2,3,[2]) (3,5,[3]) (4,2,[4]) (5,2,[5])
(1,1,[1]) (2,3,[2]) (3,5,[3]) (4,2,[4]) (5,2,[6])
```

with the number in square brackets providing a sentence-specific vocabulary ID. The next source sentence would re-index the target vocabulary specifically according to the outputs provided, starting at 1.

For each system, we include alignments corresponding to each translation in its 10-best candidate list for the source sentence. In this way, the data consumer can decide whether 1-best or 10-best outputs are best for analyzing ordering information.

We also include alignments inferred from the reference translations as a distinct reference system. This system provides alignments corresponding to 4 different references. These are presented as a 4-best list for consistency, but the order of the references does not correspond to reference quality.

3 System Descriptions

3.1 Microsoft Research (MSR) [3]

The MSR team is: Xiaodong He, Jianfeng Gao, Chris Quirk, Patrick Nguyen, Arul Menezes, Robert Moore, Kristina Toutanova, Mei Yang, and William Dolan.

3.1.1 MSR treelet system

The MSR Tree-to-String system uses a syntax-based decoder (Menezes and Quirk, 2007), informed by a source language dependency parse (Chinese). The

Chinese text is segmented using a Semi-CRF Chinese word breaker trained on the Penn Chinese Treebank (Andrew, 2006), then POS-tagged using a feature rich Maximum Entropy Markov Model, and parsed using a dependency parser trained on the Chinese Treebank (Corston-Oliver et al., 2006). The English side is segmented to match the internal tokenization of the reference BLEU script. Sentences are word aligned using an HMM with word-based distortion (He, 2007), and the alignments are combined using the grow-diag-final method. Treelets, templates, and order model training instances are extracted from this aligned set; treelets are annotated with relative frequency probabilities and lexical weighting scores. The decoder uses three language models: a small trigram model built on the target side of the training data, a medium sized LM built on only the Xinhua portion of the English Gigaword corpus, and a large LM built on the whole English Gigaword corpus using a scalable LM toolkit (Nguyen et al., 2007). It also has treelet count, word count, order model logprob, and template logprob features. At decoding time, the 32-best parses for each sentence are packed into a forest; packed forest transduction is used to find the best translation.

3.1.2 MSR phrase-based system

The second MSR system is a single-pass phrase-based system. The decoder uses a beam search to produce translation candidates left-to-right, incorporating future distortion penalty estimation and early pruning to limit the search (Moore and Quirk, 2007). The data is segmented and aligned in the same manner as above. Phrases are extracted and provided with conditional model probabilities of source given target and target given source (estimated with relative frequency), as well as lexical weights in both directions. In addition, word count, phrase count, and a simple distortion penalty are included as features.

3.1.3 MSR syntactic source reordering system

The MSR syntactic source reordering MT system is essentially the same as the second MSR system except that we apply a syntactic reordering system as a preprocessor to reorder Chinese sentences in training and test data in such a way that the reordered Chinese sentences are much closer to English in terms of word order. For a Chinese sentence, we first parse it using the Stanford Chinese Syntactic Parser (Levy and Manning, 2003), and then reorder it by applying a set of reordering rules, proposed by (Wang et al., 2007a), to the parse tree of the sentence.

3.2 Microsoft Research Asia (MSRA) [3]

The MSRA team is: Mu Li, Chi-Ho Li, Dongdong Zhang, Long Jiang, and Ming Zhou.

3.2.1 MSRA syntax-based pre-ordering system

The MSRA syntax-based pre-ordering based MT system uses a syntax-based pre-ordering model as described in (Li et al., 2007). Given a source sentence and its parse tree, the method generates, by tree operations, an n-best list of reordered inputs, which are then fed to a standard phrase-based decoder to produce the optimal translation. In implementation, the Stanford parser (Levy and Manning, 2003) is used to parse the input Chinese sentences. In the system, GIZA++ (Och and Ney, 2003) is used for word alignment and a modified version of MSRSeg tool (Gao et al., 2005) is used to perform Chinese segmentation. Moreover, we recognize certain named entities such as number, data, time, person / location names. For those named entity, translations are generated by rules or lexicon look-up. These translations serve as part of the hypotheses of the translation of the entire sentence. The decoder is a lexicalized maxent-based decoder. Note that non-monotonic translation is used here since the distance-based model is needed for local reordering. A 5-gram language model is used, which is trained on the Xinhua part of English Gigaword version 3 using an MSRA LM training tool. In order to obtain the translation table, GIZA++ is run over the training data in both translation directions, and the two alignment matrices are integrated by the grow-diag-final method into one matrix, from which phrase translation probabilities and lexical weights of both directions are obtained. Regarding to the distortion limit, our experiments show that the optimal distortion limit is 4, which was therefore selected for all our later experiments.

3.2.2 MSRA hierarchical phrase-based system

This is a re-implementation of hierarchical phrase-based system as described by Chiang (2007). It uses a statistical phrase-based translation model that uses hierarchical phrases. The model is a synchronous context-free grammar and it is learned from parallel data without any syntactic information. In this system, the same word segmentation and word alignment process as described in Section 3.2.1 were adopted, as well as the language models and the handling of named entities.

3.2.3 MSRA lexicalized re-ordering system

This system uses a lexicalized re-ordering model similar to the one described by (2006). It uses a maximum entropy model to predicate reordering of neighbor blocks (phrase pairs). As previous MSRA systems, the same word segmentation, word alignment, language model and the handling of named entities were adopted as described in Section 3.2.1.

3.3 National Research Council of Canada (NRC) [1]

The NRC team is: George Foster and Roland Kuhn.

3.3.1 NRC system

The NRC system uses a standard two-pass phrase-based approach. Major features in the first-pass log-linear model include phrase tables derived from symmetrized IBM2 and HMM word alignments, a static 5-gram LM trained on the Giga-word corpus using the SRILM toolkit, and an adapted 5-gram LM derived from the parallel corpus using the technique of Foster and Kuhn (2007). Other features are word count and phrase-displacement distortion. Decoding uses the cube-pruning algorithm of Huang and Chiang (2007), and parameter tuning is performed using the minimum error-rate training algorithm (Och, 2003) with a closest-match brevity penalty. The re-scoring pass uses 5000-best lists, with additional features including various HMM- and IBM- model probabilities; word, phrase, and length posterior probabilities; Google ngrams; reversed and cache LMs; and quote and parenthesis mismatch indicators.

3.3.2 SRI International (SRI) [1]

The SRI team is: Jing Zheng, Wen Wang, Necip Fazil Ayan, Dimitra Vergyri, Nicolas Scheffer, and Andreas Stolcke.

3.3.3 SRI system

SRI's system is a hierarchical phrase-based system that uses a 4-gram language model in the first pass to generate n-best lists, which are re-scored by three additional language models to generate the final translations via re-ranking. The text is tokenized with RWTH's Chinese-English system preprocessor, which uses LDC's word-segmenter to convert character strings to word-strings. The preprocessor also performs rule-based translation for number, date and time expressions, as well as some cleanup. The translation engine is SRI's in-house developed CKY-style decoder, which performs parsing and generation simultaneously guided by a language model and synchronous context free grammars (SCFGs). The SCFGs are extracted from parallel text with word alignments generated by GIZA++, in the similar manner described by Chiang (2007). The three re-scoring language models include a count-based LM from Google Tera-word corpus, an almost parsing class LM based on SARV tags, and an approximated parser based LM (Wang et al., 2007b).

4 Conclusion

We hope that by releasing this data, we will enable others with expertise in modeling collections of permutations, and other related fields, to play in the translation domain, where we feel there is both an opportunity to make an impact on a real application, and an opportunity to test and improve these modeling techniques on what is sure to be a unique sort of permutation phenomenon.

References

Galen Andrew. 2006. A hybrid Markov/semin-Markov conditional random field for sequence segmentation. In *EMNLP*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, June.

Simon Corston-Oliver, Anothony Aue, Kevin Duh, and Eric Ringger. 2006. Multilingual dependency parsing using bayes point machines. In *HLT-NAACL*.

Goerge Foster and Roland Khun. 2007. Mixture-model adaptation for SMT. In *ACL Workshop on Machine Translation*.

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: a pragmatic approach. *Computational Linguistics*, 31(4).

Xiaodong He, Mei Yang, Jianfeng Gao, and Patrick Nguyen. 2008. A study on combining multiple statistical machine translation systems. Technical report, Microsoft Research.

Xiaodong He. 2007. Using word-dependent transition models in HMM-based word alignment for statistical machine translation. In *ACL Workshop in Machine Translation*.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *ACL*.

Alexandre Klementiev, Kevin Small, and Dan Roth. 2007. An unsupervised learning algorithm for rank aggregation. In *ECML*.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *ACL*.

Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. 2007. A probabilistics approach to syntax-based reordering for statistical machine translation. In *ACL*.

Arule Menezes and Chirs Quirk. 2007. Using dependency order templates to improve generality in translation. In *ACL Workshop on Machine Translation*.

Robert Moore and Chris Quirk. 2007. Faster beam-search decoding for phrasal statistical machine translation. In *MT Summit XI*.

Patrick Nguyen, Jianfeng Gao, and Milind Mahajan. 2007. MSRLM: A scalable language modeling toolkit. Technical Report MSR-TR-2007-144, Mircosoft Research.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

Franz J. Och. 2003. Minimum error rate training for statistical machine translation. In *ACL*, pages 160–167.

Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007. Combining outputs from multiple machine translation systems. In *HLT-NAACL*.

Chao Wang, Michael Collins, and Philipp Koehn. 2007a. Chinese syntactic reordering for statistical machine translation. In *EMNLP*, pages 737–745.

Wen Wang, Andreas Stolcke, and Jing Zheng. 2007b. Reranking machine translation hypotheses with structured and web-based language models. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *COLING-ACL*, pages 521–528, Sydney, Australia.