The MSR-NRC-SRI MT System for NIST Open Machine Translation 2008 Evaluation

AUTHORS AND AFFILIATIONS

MSR: Xiaodong He, Jianfeng Gao, Chris Quirk, Patrick Nguyen, Arul Menezes, Robert Moore, Kristina Toutanova, Mei Yang[1], and William Dolan
MSRA: Mu Li, Chi-Ho Li, Dongdong Zhang, Long Jiang, and Ming Zhou
NRC: George Foster and Roland Kuhn
SRI: Jing Zheng, Wen Wang, Necip Fazil Ayan, Dimitra Vergyri, Nicolas Scheffer, and Andreas Stolcke

1 SITE AFFILIATION

1.1 Site name

MSR-NRC-SRI

1.2 Full names of group members

Microsoft Research (Redmond and Asia)
National Research Council (Canada)
SRI International

2 CONTACT INFORMATION

Xiaodong He	xiaohe@microsoft.com
Mu Li	muli@microsoft.com
Roland Kuhn	roland.kuhn@cnrc-nrc.gc.ca
Jing Zheng	zj@speech.sri.com

3 SUBMISSIONS

We participated in the Chinese-to-English Constrained training data track MT evaluation. We submit one primary submission and two contrastive submissions. They are:

MSR-NRC-SRI_**chinese_constrained**_primary
MSR-NRC-SRI_**chinese_constrained**_contrast1
MSR-NRC-SRI_**chinese_constrained**_contrast2

4 PRIMARY SYSTEM SPEC

4.1 Core MT Engine Algorithmic Approach

4.1.1 The system combination framework

A system combination framework is used for this entry. Within this framework, up to eight individual systems are combined to produce the final MT output.

The system combination approach combining system outputs at the word level is similar to the one described in (Rosti et al., 2007). Compared to the previous work, we developed a new method to generate a better alignment between multiple MT hypotheses from different individual systems, which is used to construct a high-quality confusion network. The details of our method will be elaborated in a future paper (He et al., 2008).

First, a minimum Bayes risk (MBR) based method is used to select a backbone from the multiple hypotheses, then all the hypotheses are aligned to that backbone to form a confusion network, i.e., a word lattice in which each word is aligned to a list of alternative words (including *null*). Then, a set of features, including language model scores, word count, and normalized system voting score, are used to decode the confusion network. In training, a confusion network is constructed based on the multiple hypotheses of each sentence in a dev set. Then the corresponding feature weights are trained using Powell's search to maximize the BLEU score on that dev set. In testing, a confusion network for each sentence in the test set is constructed and these feature weights are applied to decode the final MT output from the confusion network.

In this entry, two language models are used, including a 3-gram LM trained on the English part of the parallel training data, and a 5-gram LM trained on the whole English Gigaword corpus using a scalable LM toolkit (Nguyen et al., 2007).

[1] Mei Yang was an intern with MSR in the summer of 2007

4.1.2 Description of individual systems

There are eight individual systems incorporated in the system combination framework. Among these eight systems, MSR provided three of them, MSRA provided other three systems, and each of NRC and SRI provided one system. In the following sub-sections, we give a brief description of each system.

4.1.2.1 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.

4.1.2.2 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.

4.1.2.3 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. (2007), to the parse tree of the sentence.

4.1.2.4 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++ 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.

4.1.2.5 MSRA hierarchical phrase-based system

This is a re-implementation of hierarchical phrase-based system as described by Chiang (2005). 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 4.1.2.4 were adopted, as well as the language models and the handling of named entities.

4.1.2.6 MSRA lexicalized re-ordering system

This system uses a lexicalized re-ordering model similar to the one described by Xiong et al. (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 4.1.2.4.

The above six systems are also the six individual systems used in the primary submission of the MSR-MSRA entry. Please refer to the system description of that entry for more details.

4.1.2.7 NRC system

NRC contributed one system within the system combination framework. It corresponds to the "*NRC_chinese_constrained_constrast1*" submission that NRC submitted in the NRC-only entry.

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 Och's max-BLEU algorithm with a closest-match brevity penalty. The rescoring 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.

4.1.2.8 SRI system

SRI contributed one system within the system combination framework. It corresponds to the "*SRI_chinese_constrained_constrast1*" submission that SRI submitted in the SRI-only entry.

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 rescored 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 (2005). The three rescoring 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., 2007).

4.1.3 Scalable language model server

Several language models used in this submission were built using our publicly available scalable language modeling toolkit (Nguyen et al, 2007). They were directly available in the first decoding pass in some systems, but also in the subsequent system combination and case restoration. For all cases, a single server handled all requests from up to 40 decoding processes, loading one or two language models entirely into memory. A Gigaword 5-gram model is trained in

about 3 hours on a single machine starting from tokenized text. All language models were 5-grams with a vocabulary size of 120k, count cutoff of 1, and modified absolute discounting (Gao et al., 2001). A typical Gigaword LM contains 30M bigrams, 170M trigrams, 340M 4-grams, and 440M 5-grams. For first pass decoding, we use two LMs: one based on the whole Gigaword corpus, and one based on the Xinhua portion of the Gigaword corpus. For system combination, we only use the Gigaword LM. For case restoration, a case sensitive Gigaword 5-gram LM was built.

4.1.4 Case restoration

The model for case restoration is applied as a final step after system combination. It predicts the true-case forms of words in a target translation, given a lowercase target translation, and a source sentence. The model is a log-linear conditional Markov Model, using syntactic and word-based features from the source and target, and capitalization pattern features from the target (Minkov et al., 2007). This model is combined with a 5-gram LM trained on the Giga-word corpus and a rule-based component for capitalizing headlines. Based on our post-eval investigation, the primary submission gave a case insensitive BLEU-4 score of 0.3244 on the 2008 Chinese-to-English "current" test set, where the case sensitive BLEU-4 score is 0.3089.

4.1.5 MT hypothesis length adaptation

In our system, a simple unsupervised MT hypothesis length adaptation method is used. We model the expected word count ratio between the hypotheses and the source sentences. This is motivated from the assumption that, in general, there exists a relatively stable word count ratio between two languages. When testing, if the MT system generates hypotheses that are too long or too short, we adapt the model (feature weights) to encourage the system to produce hypotheses with reasonable length based on the expected hyp/src ratio.

This expected word count ratio is estimated on the dev set. I.e., after Max-Bleu training, we compute the word count ratio between the MT hypotheses and the source sentences. Then at test, we adapt the length of the MT hypotheses by adjusting the word count weight so that the hypotheses vs. source word count ratio matches the expected hyp/src ratio. We found this length adaptation scheme helps in general, and is especially helpful if there is a severe mismatch between dev and test sets. In the MSR-NRC-SRI entry, we applied this scheme to the primary submission and the first contrastive submission. Please refer to section 5 for more details.

4.1.6 MT08 results

We participated in the NIST MT08 Chinese-to-English constrained training data track MT evaluation. All individual systems are trained using constrained training data corpora prescribed by NIST.

Regarding the system combination model training, the development set is a sampling of all past years' NIST MT test data. For the primary submission, we only sample the newswire data from MT04 to MT06-newswire. In total, we sampled 1002 newswire sentences: 35% from MT04, 55% from MT05, and 10% from MT06-newswire.

As shown in the NIST preliminary results sheet, our primary system achieved a case sensitive BLEU-4 score of 0.3089 on the 2008 "current" test set, where the best individual system out of the eight systems used for system combination is from SRI: "*SRI_chinese_constrained_constrast1*", which gave a case sensitive BLEU-4 score of 0.2624 on the 2008 "current" test set.

4.2 Critical Additional Features and Tools Used

In our system, a regular expression based dateline detection module is used to detect common dateline formats of newswire text. Then, the detected datelines are translated by a set of simple rules. In the MT08 Chinese-to-English test set, we totally detected and translated 30 datelines. Note that the whole dateline detection and translation module is built based on previous NIST MT test data and training data; and this dateline processing module is only applied to the six MSR/MSRA systems. The MT hypotheses from NRC and SRI systems are used in the combination framework as is.

4.3 Significant Data Pre/Post-Processing

In training, we dropped parallel sentences that were too long (more than 80 words on either side), or for which the word count ratio was too

large (>8.5) or too small (<0.118). At post-processing, we removed any consecutive duplicated words that were longer than two letters. However, our post-eval investigation showed that this had almost no effect on the BLEU score.

4.4 Other Data Used (Outside the Prescribed LDC Training Data)

No outside data were used.

5 KEY DIFFERENCE IN CONTRASTIVE SYSTEMS

5.1 Contrastive system 1

MSR-NRC-SRI_chinese_constrained_contrast1

Compared to the primary submission, the only difference of this contrastive system is that it uses a different dev set for system combination model training. The dev set contains 501 newswire sentences generated in a similar way as that of the primary submission. Beside these, it also contains the 483 sentences of newsgroup data from NIST MT06 test set. This is motivated by the *MT08 plan* saying there would be both newswire and web data included in the MT08 test set.

This submission achieved a case sensitive BLEU-4 score of 0.3080 on the *current* test set, according to the NIST preliminary results sheet.

5.2 Contrastive system 2

MSR-NRC-SRI _chinese_constrained_contrast2

This submission is the same as the first contrastive submission except that no hypothesis length adaptation is applied. It gave a case sensitive BLEU-4 score of 0.3048 on the *current* test set, according to the NIST preliminary results sheet.

Acknowledgments The authors are grateful to Galen Andrew for providing his word segmentation component, and to Anthony Aue for providing the Powell's search optimization tools.

REFERENCES

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, NAACL-HLT

Arul Menezes and Chris Quirk. (2007). Using Dependency Order Templates to Improve Generality in Translation. In Proc 2nd WMT at ACL, Prague, Czech Republic

Chao Wang, Michael Collins, and Philipp Koehn. (2007). Chinese Syntactic Reordering for Statistical Machine Translation. In proceedings of EMNLP-CoNLL 2007.

Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, Yi Guan, (2007). A Probabilistic Approach to Syntax-based Reordering for Statistical Machine Translation. ACL

David Chiang. (2005). A hierarchical phrase-based model for statistical machine translation. In Proceedings of ACL.

Deyi Xiong, Qun Liu and Shouxun Lin, (2006). Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. ACL

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. (2007). Generating Complex Morphology for Machine Translation. ACL.

Galen Andrew, (2006). A hybrid Markov/semi-Markov conditional random field for sequence segmentation. In Proceedings of EMNLP 2006, Sydney, Australia

George Foster and Roland Kuhn. (2007). Mixture-Model Adaptation for SMT. In Proc 2nd WMT at ACL Prague, Czech Republic.

Jianfeng Gao, Joshua Goodman, and Jiangbo Miao (2001). The use of clustering techniques for language modeling - application to Asian languages. In Computational Linguistics and Chinese Language Processing, vol 6., No. 1, pp 27-60.

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).

Liang Huang and David Chiang. (2007). Forest Rescoring: Faster Decoding with Integrated Language Models. Proc ACL.

Patrick Nguyen, Jianfeng Gao and Milind Mahajan (2007). MSRLM: a scalable language modeling

toolkit. Microsoft Research Technical Report MSR-TR-2007-144.

Robert Moore and Chris Quirk. (2007). Faster Beam-Search Decoding for Phrasal Statistical Machine Translation. MT Summit XI, Copenhagen, Denmark

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? Published in Proceedings of ACL 2003

Simon Corston-Oliver, Anthony Aue, Kevin Duh, amd Eric Ringger, (2006). Multilingual Dependency Parsing using Bayes Point Machines, Proc. of NAACL-HLT, New York, New York

Wen Wang, Andreas Stolcke, Jing Zheng (2007). Reranking Machine Translation Hypotheses With Structured and Web-based Language Models. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, Kyoto.

Xiaodong He, (2007). Using Word-Dependent Transition Models in HMM based Word Alignment for Statistical Machine Translation. In Proc 2nd WMT at ACL Prague, Czech Republic

Xiaodong He, Mei Yang, Jianfeng Gao, and Patrick Nguyen, (2008). A Study on Combining Multiple Statistical Machine Translation Systems. Microsoft Research Technical Report