第七届机器翻译研讨会 NTT 技术报告

吴先超　须藤克仁　Kevin Duh　塚田元　永田昌明
NTT Communication 科学基础研究所 京都 619-0237 日本
E-mail: {wu.xianchao, sudoh.katsuhito, kevin.duh, tsukada.hajime, nagata.masaaki}@lab.ntt.co.jp

摘 要：本文主要介绍 NTT Communication 科学基础研究所协创情报研究部言语智能研究组参加 2011 年第七届全国机器翻译研讨会(CWMT2011)评测的情况。本单位参加了全部测评项目中的两个子项：英汉科技领域机器翻译和日汉新闻领域机器翻译。本文首先简要地介绍本单位的统计机器翻译系统的实现框架，其次比较各个系统在评测数据上的性能表现，最后对翻译评测的结果加以简略分析。

关键字：人工智能、自然语言处理、统计机器翻译、系统融合

Technical Report of NTT for the 7th China Workshop on Machine Translation

Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Masaaki Nagata
NTT Communication Science Laboratories, NTT Corporation. Kyoto, 619-0237, Japan
E-mail: {wu.xianchao, sudoh.katsuhito, kevin.duh, tsukada.hajime, nagata.masaaki}@lab.ntt.co.jp

Abstract: *This paper briefly introduces the statistical machine translation (SMT) systems and the evaluation results of NTT Communication Science Laboratories (NTT CS Lab.) for attending the 7th China workshop on machine translation (CWMT 2011). This year, NTT CS Lab. participated in two evaluation tasks: English-to-Chinese translation for scientific and technological text, and Japanese-to-Chinese translation for news. This paper first briefly describes the implement framework of our SMT systems, and then reports and analyzes the translation accuracies of these systems on the test sets.*

Keywords: *artificial intelligence, natural language processing, statistical machine translation, system combination*

1 Introduction

This paper reports the techniques and performances of the statistical machine translation (SMT) systems in NTT Communication Science Laboratories (NTT CS Lab.) for attending the 7th China workshop on machine translation (CWMT 2011). This year, NTT CS Lab. participated in two evaluation tasks: English-to-Chinese (EC) translation for scientific and technological text, and Japanese-to-Chinese (JC) translation for news.

For EC translation, we used a syntactic tree based pre-ordering method alike that described in (Wu et al. 2011). For this task, we only made use of the data sets provided by CWMT 2011.

For JC translation, we made use of a dependency tree based pre-ordering method (Wu et al. 2011) and a head-finalization Chinese (HFC) based post-reordering method (Sudoh et al. 2011). For this task, we used both the data sets provided by CWMT 2011 and external data sets which are available in NTT.

In addition, for both EC and JC translation, we took minimum Bayes risk (MBR) as the decision rule for doing system combination.

2 Single SMT System Description

2.1 Dependency Tree based pre-ordering method

As described in (Wu et al. 2011) in detail, we propose to use source Japanese chunk-based dependency trees to automatically extract pre-ordering rules for Japanese-to-Chinese translation.

In order to obtain a fine-grained classification of the reordering phenomena, Japanese function words and punctuation marks were included in the pre-ordering rules as additional lexical clues.

A pre-ordering rule includes a source tree fragment and the target-language-style relative positions of the leaf nodes of this source tree fragment. Pre-ordering rules are extracted from word-aligned dependency-tree-to-string tuples. Making use of the original parallel sentences and GIZA++ (Och and Ney 2003), word alignments were trained beforehand to guide pre-ordering rule extraction. After pre-ordering of the source Japanese sentences, we run GIZA++ again to obtain better word alignments.

During pre-ordering rule extraction, we first transfer dependency trees into constituency trees. The consideration behind this is to extract and apply pre-ordering rules in an ordered way, e.g., bottom-up traversal. Through a topological scan of the chunks in a dependency tree, we create a part-of-speech (POS) node for each chunk. The POS label of the newly created node takes the POS of the head (or dominant) phrase/word in the chunk. For higher level non-terminal nodes, we take a generalized label "X" as their POS tags. For similarity, we limit to extract pre-ordering rules whose tree fragment contains no more than two layers. That is, we only record the relative positions of the direct child nodes of a non-terminal node.

During pre-ordering rule application for the development and test sets, we perform the following steps. First, we parse the input Japanese sentences and change the dependency tree into a constituency tree. Then, we retrieve available pre-ordering rules through a bottom-up traversal of the constituency tree and keep a k-best list for each non-terminal node. Finally, we pick one reordered sentence from the k-best list of the root node of the constituency tree according to the n-gram language model (LM) score of the candidate pre-ordered sentences. The n-gram LM is trained using SRILM (Stolcke 2002) on the pre-ordered Japanese sentences in the training data.

We adopted the similar approach for pre-ordering English sentences into Chinese style sentences for EC translation.

2.2 Head-finalization based post-ordering method

As described in (Sudoh et al. 2011) in detail, we propose a novel post-ordering approach for translating between languages with distinct word ordering. This approach is the inverse problem of the pre-ordering approach and can be solved by two simplified translation steps: source-to-"source-ordered target" and "source-ordered target"-to-target translations.

For JC translation, we first translate Japanese into head-finalized Chinese (HFC) with no or short-distance reordering. Then, we translate HFC into Chinese with long-distance reordering and a small number of edits on Chinese words. The usage of HFC is inspired by Head-Final English (HFE, Isozaki et al. 2010). We argue post-ordering is a promising way of translating those language pairs where good pre-ordering methods have been developed beforehand in the opposite direction.

3 System Combination Approach

We take Minimum Bayes Risk (MBR) as the decision rule for SMT system combination. MBR arose in Bayes decision theory (Duda et al. 2000) and has since then been applied to speech recognition (Goel and Byrne 2000) and machine translation (Kumar and Byrne 2004). The idea is to choose hypotheses that minimize Bayes Risk as oppose to those that maximize posterior probability. This enables the use of task-specific loss functions, such as BLEU score.

4 Methods and Toolkits for Data Processing

4.1 Word segmentation

We used the Stanford Chinese word segmenter (Chang et al. 2008) with Chinese Penn Treebank standard for Chinese word segmentation. For Japanese word segmentation, we ran Mecab v0.98. We tokenized and lowercased English sentences for EC translation.

4.2 Word alignment

We ran GIZA++ (Och and Ney 2003) to obtain bidirectional word alignments. The heuristic strategy of "*grow-diag-final-and*" (Koehn et al. 2007) was adopted to combine the word

alignments of source-to-target and target-to-source directions. The combined word alignments were used to extracting phrase translation table and training distortion model.

4.3 Language models

SRI Language Modeling Toolkit (Stolcke 2002) with modified Kneser-Ney smoothing (Chen and Goodman 1998) was used to train 5-gram target language models. Besides using the target language sentences of the training data for language model training, we also used the SogouCA Webpage news (http://www.sogou.com/labs/dl/ca.html) for generating a large-scale language model. We again ran the Stanford Chinese word segmenter with Chinese Penn Treebank standard for Chinese word segmentation. In the pre-processed Sogou data, there are 35,889,682 sentences with 569,015,193 words.

4.4 Syntactic and dependency parsers

For English parsing, we chose a state-of-the-art Head-driven Phrase Structure Grammar (HPSG) parser, Enju v2.4.2 (Miyao and Tsujii 2008). For Japanese dependency parsing, we used the Cabocha parser v0.53 (Kudo and Matsumoto 2002) which generate chunk-level dependency trees of Japanese sentences. For Chinese dependency parsing, we selected the Stanford parser (Chang et al. 2009). We extracted head information from the Chinese dependency parsing trees.

4.5 Evaluation metrics

We report the translation accuracies basing on character level BLEU score (Pepineni et al. 2002), BLEU-SBP (Chiang et al. 2008), and the metrics used by the organizers of CWMT 2011. Note that instead of using BLEU-SBP for Minimum Error Rate Training (MERT, Och 2003), we used the traditional word level BLEU4 as the metrics for MERT and system combination.

5 Experiments

5.1 English-to-Chinese Science Domain

We only use the data supplied by CWMT 2011 for the EC translation task. In the test set, we found there are numerous sentences that are extremely long and can split into several sentences to be translated separately. By using sentence separators such as "." and ";", we split the original 2,497 English sentences into 3,479 sentences. These relatively short sentences were taken as the input of our decoders. During evaluation, we again combine the output sentences into 2,497 sentences for scoring.

Table 1 shows the statistics of the experiment data for the official test set of EC translation. Note that even the final results were evaluated based on Chinese characters, we still used Chinese words as the basic translation units for training, tuning, and testing. We used Enju v2.4.2 to parse the English sentences for pre-ordering and the parse success rate ranges from 95.5% to 99.2%.

Table 2 lists the automatically evaluated translation accuracies of the official test set of EC translation through on-line testing (http://mtgroup.ict.ac.cn/demo/cwmt/index.php). Here, we take Moses as our baseline decoder. For our pre-ordering methods, we also used Moses as the testing system that took pre-ordered source sentences as experimental data.

We used the target sentences in the training data and SogouCA news data to train two 5-gram LMs, named trainLM and SogouLM. Decoded using Moses on the test set, we found that using trainLM achieved a better translation accuracy of 3.19 BLEU5-SBP points than that achieved by using SogouLM. Note that similar tendency appears when scoring the development set: using trainLM achieved a BLEU4 score of 0.3374, significantly better ($p < 0.01$) then the score 0.2866 achieved by using SogouLM. According to these comparisons, we used trainLM as the only LM for training our pre-ordering/postordering systems.

Table 1. Statistics of the experiment data for EC translation.

	Train	Dev	Test
# of sent.	911,527	1,116	3,479
# En words	25,289,129	24,332	99,900
# Ch words	25,005,181	93,482/4	NA
Parse success rate	97.4%	99.2%	95.5%

Table 2. Translation accuracies for EC translation.

	Moses +SogouLM	Moses +trainLM	Preorder +trainLM	Preorder-variables +trainLM	MBR-combination
Dev-oracle-Char-BLEU4	0.4756	**0.5290**	0.5025	0.4957	-
Dev-oracle-Word-BLEU4	0.2866	**0.3374**	0.3123	0.3090	-
BLEU5-SBP	0.3428	0.3747	**0.3816**	0.3634	**0.3882**
BLEU5	0.3531	0.3928	**0.4051**	0.3836	**0.4100**
BLEU6	0.2910	0.3274	**0.3374**	0.3165	**0.3437**
NIST6	9.1015	9.9581	**10.261**	10.036	10.251
NIST7	9.1179	9.9802	**10.283**	10.056	10.275
GTM	0.7864	0.8129	**0.8257**	0.8149	0.8219
mWER	0.6166	0.6140	0.6265	0.6345	**0.6076**
mPER	0.3513	0.3236	**0.3115**	0.3194	0.3147
ICT	0.4142	**0.4143**	0.4081	0.4023	0.4117

By using trainLM, our pre-ordering approach achieved an improvement of 0.69 BLEU5-SBP points. Also, under our pre-ordering framework, using additional variables right after predicates and their arguments as suggested by Isozaki et al. (2010) achieved an improvement of 1.82 BLEU5-SBP points than without using variables. Finally, the combination of "Moses+trainLM", "Preorder+trainLM", and "Preorder+trainLM-variables" achieved the best BLEU5-SBP of 0.3882, as compared with other single systems.

5.2 Japanese-to-Chinese News Domain

Besides using the primary data supplied by CWMT 2011, we also used the Xinhua Japanese-Chinese news data as the external data for training our systems. The statistics of the primary data and Xinhua data are listed in Table 3. Note that even the final results were evaluated based on Chinese characters, we still used Japanese words and Chinese words as the basic translation units for training, tuning, and testing.

Similar to EC translation, we also trained two 5-gram LMs by using the target sentences in the training data and Sogou news data, named trainLM and SogouLM. Based on the comparison of word level BLEU4 on the development set decoded by Moses, we used SogouLM for training our pre-ordering/postordering systems with primary data (Table 4) and used trainLM for training our pre-ordering/postordering systems with primary + Xinhua data (Table 5).

Table 4 lists the translation accuracies for the official test set of JC translation by using the primary data. The oracle character/word level BLEU4 scores of the development set are listed as well. In terms of BLEU related scores, the baseline "Moses+SogouLM" is the best among the single systems. However, when evaluated by metrics such as mWER, "Postorder+SogouLM" performed better than "Moses+SogouLM". Since Japanese is a subject-object-verb (SOV) language and different from Chinese which is a SVO language, we argue word ordering play an important role during translating and edit-distance based mWER is more sensitive to word ordering difference than n-gram matching based BLEU scores.

Using MBR-based combination of these four systems, we achieved a (Chinese character level) BLEU5-SBP of 0.3964, which is the best compared with other single systems. Due to limited experimental time, we optimized system combination based on word level BLEU4 instead of character level BLEU5-SBP. It will be valuable to investigate the system combination results by directly optimizing BLEU5-SBP.

Table 3. Statistics of the experiment data for JC translation.

	Primary-train	Xinhua-train	Primary-Dev	Primary-Test
# of sent.	282,483	593,176	500	3,500
# Ja words	3,237,775	19,680,033	17,829	119,985
# Ch words	2,475,374	14,066,920	51,472/4	NA

Table 4. Translation accuracies for JC translation by using the primary data.

	Moses +SogouLM	Moses +trainLM	Preorder +SogouLM	Postorder +SogouLM	MBR-combination
Dev-oracle-Char-BLEU4	0.4800	0.4810	0.4578	**0.4874**	-
Dev-oracle-Word-BLEU4	**0.3226**	0.3149	0.2946	0.3203	-
BLEU5-SBP	**0.3801**	0.3676	0.3738	0.3729	**0.3964**
BLEU5	**0.3941**	0.3815	0.3857	0.3930	**0.4071**
BLEU6	**0.3318**	0.3173	0.3238	0.3277	**0.3450**
NIST6	9.9059	9.8978	9.8212	**10.0079**	**10.1025**
NIST7	9.9239	9.9121	9.8382	**10.0248**	**10.1208**
GTM	0.8129	0.8156	0.8121	**0.8166**	**0.8267**
mWER	0.5847	0.5383	0.5674	**0.5241**	0.5562
mPER	0.3239	0.3217	0.3298	**0.3183**	0.3139
ICT	0.4216	0.4254	0.4080	**0.4531**	0.4266

Table 5. Translation accuracies for JC translation by using the primary + Xinhua data.

	Moses +SogouLM	Moses +trainLM	Preorder +trainLM	Postorder +trainLM	MBR-combination
Dev-oracle-Char-BLEU4	0.5136	**0.5286**	0.4938	0.5095	-
Dev-oracle-Word-BLEU4	0.3453	**0.3602**	0.3259	0.3437	-
BLEU5-SBP	**0.4339**	0.4284	0.4106	0.4191	0.4260
BLEU5	**0.4597**	0.4548	0.4315	0.4439	0.4510
BLEU6	**0.3974**	0.3908	0.3686	0.3792	0.3866
NIST6	**10.4056**	10.3787	10.2585	10.3924	**10.4398**
NIST7	**10.4300**	10.4033	10.2786	10.4146	**10.4638**
GTM	0.8263	0.8292	0.8255	**0.8329**	0.8328
mWER	0.5297	**0.5091**	0.5540	0.5223	0.5206
mPER	0.3100	0.3088	0.3141	**0.3008**	**0.3008**
ICT	0.4698	0.4787	0.4546	**0.4851**	**0.4861**

In addition, note that in terms of mWER and mPER, the combined results are no better than "Postorder + SogouLM". This reflects that the optimization of word level BLEU4 not necessarily ensure the optimization of mWER and mPER, which are essentially influenced by the word order of the translation results. Again, it will be valuable to investigate the system combination results by directly optimizing mWER or mPER and check the consequent influence to BLEU related scores.

Table 5 lists the translation accuracies for JC translation by using the primary + Xinhua data. Using MBR-based combination of these four systems, we achieved a BLEU5-SBP of 0.4260, which is not the best compared with other single systems. We argue one major reason is that due to limited experimental time, we optimized system combination on the development set based on word level BLEU4 instead of character level BLEU5-SBP.

6 Conclusion

We have reported the major techniques and the performances of the SMT systems in NTT CS Lab. for attending CWMT 2011. This year, NTT CS Lab. participated in two evaluation tasks: English-to-Chinese translation for scientific and technological text and Japanese-to-Chinese translation for news. Compared with the top ranking systems, there is still a big room to improve the performance of our systems. Through a deeper and more wildly communication with other teams, we hope to continue to push the development of SMT in the future.

References

Pi-Chuan Chang, Michel Galley and Chris Manning. 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of ACL Third Workshop on Statistical Machine Translation.*

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation.*

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University Center for Research in Computing Technology.

David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008. Decomposability of Translation Metrics for Improved Evaluation and Efficient Algorithms. *In Proceedings of EMNLP 2008*, pages 610-619.

Richard Duda, Peter Hart, and David Stork. 2000. *Pattern Classification.* Wiley-Interscience, 2nd edition.

Vaibhava Goel and William Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech and Language*, 14(2):115–135.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010. A simple reordering rule for sov languages. In *Proceedings of WMT-MetricsMATR.*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ond˘rej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of CoNLL-2002*, pages 63–69. Taipei, Taiwan.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of HLT-NAACL.*

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic hpsg parsing. *Computational Linguistics*, 34(1):35–80.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, pages 901–904.

Katsuhito Sudoh, Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada and Masaaki Nagata. 2011. Post-ordering in Statistical Machine Translation. In *Proceedings of MT Summit XIII.*

Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada and Masaaki Nagata. 2011. Extracting Pre-ordering Rules from Chunk-based Dependency Trees for Japanese-to-English Translation. In *Proceedings of MT Summit XIII.*