SYSTRAN Chinese-English and English-Chinese Hybrid Machine Translation Systems for CWMT2011

Jin Yang and Satoshi Enoue
SYSTRAN Software, Inc.
4444 Eastgate Mall, Suite 310
San Diego, CA 92121, USA
E-mail: {jyang,enoue}@systransoft.com

Jean Senellart
SYSTRAN S.A.
5 rue Feydeau
75002 Paris, France
E-mail: senellart@systran.fr

Abstract: *This report describes SYSTRAN's Chinese-English and English-Chinese machine translation systems that participated in the CWMT 2011 machine translation evaluation tasks. The base systems are SYSTRAN rule-based machine translation systems, augmented with various statistical techniques. Based on the translations of the rule-based systems, we performed statistical post-editing with the provided bilingual and monolingual training corpora. In this report, we describe the technology behind the systems, the training data, and finally the evaluation results in the CWMT 2011 evaluation. Our primary Chinese-English system was ranked first in BLEU in the translation tasks.*

Keywords: *Chinese-English Machine Translation, English-Chinese Machine Translation, Rule-Based Machine Translation System, Hybrid Approach, Statistical Post-Editing*

SYSTRAN 混合策略汉英和英汉机器翻译系 CWMT2011 技术报告

Jin Yang and Satoshi Enoue
SYSTRAN Software, Inc.
4444 Eastgate Mall, Suite 310
San Diego, CA 92121, USA
E-mail: {jyang,enoue}@systransoft.com

Jean Senellart
SYSTRAN S.A.
5 rue Feydeau
75002 Paris, France
E-mail: senellart@systran.fr

摘要：本文介绍了 SYSTRAN 参加 CWMT 2011 机器翻译评测的汉英和英汉机器翻译系统。SYSTRAN 系统的基本系统是融入了各种统计方法的基于规则的机器翻译系统。在规则系统翻译结果的基础上，我们用统计方法后编辑技术，使用提供的双语和单语语料，进行自动的后编辑。本文介绍了系统中运用的技术，训练数据和在 CWMT 2011 中的评测结果。SYSTRAN 汉英系统在评测中名列前茅。

关键字：汉英机器翻译 英汉机器翻译 基于规则的机器翻译 混合策略机器翻译 统计方法译后编辑

1 Introduction

SYSTRAN has the longest history of any machine translation (MT) developer in the world. Traditionally, SYSTRAN systems adopt the rule-based approach, using enormous and diversified linguistic resources. For the last several years, SYSTRAN has been focusing on the introduction of statistical approaches to its rule-based backbone, leading to "Hybrid Machine Translation". Our Hybrid Chinese-English systems participated in the CWMT 2008 and CWMT 2009 evaluations, ranking among the top three in BLEU4-SBP and first in NIST5 (Yang, Stephan, Senellart 2008, Yang, Enoue, Senellart, Croiset 2009).

The techniques used in the Chinese-English MT system for CWMT 2008 and CWMT 2009 include: a) Employing various statistical techniques in the development of the rule-based machine translation (RBMT) systems (Senellart 2006); b) Utilizing statistical post-editing (Simard et al. 2007, Dugast, Senellart, Koehn 2007) to automatically edit the output of the RBMT system. In the past two years, we continued improving and refining these techniques, and experimenting and expanding to more language pairs and domains. In addition, more statistical techniques were introduced in the rule-based system to help making difficult linguistic decisions. In the CWMT 2011 evaluation, we participated in the Chinese-English news machine translation (ZH-EN-NEWS) task, English-Chinese news machine translation (EN-ZH-NEWS) task and S&T machine translation (EN-ZH-SCIE) task. In this paper we describe the technology behind the two systems used, the training data, and finally the evaluation results.

2 System Description

2.1 Submissions

For each of the tasks in which we participated, we trained two systems, primary and contrast. For the ZH-EN-NEWS progress task, we submitted "ce-news-progress-systran-primary-systema" and "ce-news-progress-systran-contrast-systemb". For the EN-ZH-NEWS task, we submitted the primary and contrast system outputs for each of the current and progress tasks: "ec-news-current-systran-primary-systema", "ec-news-current-systran-contrast-systemb", "ec-news-progress-systran-primary-systema", and "ec-news-progress-systran-contrast-systemb". And for the EN-ZH-SCIE task, we submitted "ec-tech-test-systran-primary-systema" and "ec-tech-test-systran-contrast-systemb".

The hybrid approach with rule-based and statistical post-editing (SPE) was used throughout the tasks to produce the primary and contrast outputs. The main difference is that the contrast systems used additional monolingual data for language modeling. Further details will be described in the following sections.

2.2 SYSTRAN Hybrid Machine Translation Systems

The traditional SYSTRAN systems are general-purpose fully automatic machine translation systems, employing a rule-based transfer approach. A unified and highly modular architecture applies to all language-pair systems. SYSTRAN's dictionaries and parsers have evolved over a long period of time, have been tested on large amounts of text, and contain extremely detailed linguistic rules and a large terminology database covering various domains. Most importantly, SYSTRAN's success in the machine translation field is built on constant and sustainable development and modernization.

The development of the SYSTRAN Chinese-English MT system began in August 1994. Work on lexical development and linguistic analysis have been continuing over the years, with steady improvement. Recent development concentrates on incorporating statistical techniques in the various components of the system: a) corpus-based monolingual and bilingual terminology extractions; b) incorporating corpus evidence in the linguistic rules (Senellart 2006); c) introducing statistical components to help making difficult linguistic decisions. At the same time, continued lexical and linguistic development is still underway: a) addition of large amounts of Named Entities; b) adapting the system to colloquial and web genres; c) continued improvement

in the science & technology domain. The current RBMT Chinese-English system contains over 2.1 million bilingual words, expressions, and linguistic rules spanning various domains. And the size is still growing rapidly.

The SYSTRAN English-Chinese system was built based on the existing SYSTRAN English parser and dictionaries developed more than a decade ago. The English system was initially built for translating technical manuals, and it uses a multi-target dictionary structure. The initial development effort for the English-Chinese system was made by adding Chinese targets to the existing English multi-target dictionaries, and adding basic transfer and generation rules. The work on the English-Chinese system has been quite limited. The priority of the recent development is to adapt the system to user-defined technical domains. In addition, over 250,000 Named Entities from the Chinese-English dictionary were automatically reversed and incorporated into the English-Chinese dictionary. Overall, there are some big improvement areas for the English-Chinese system.

2.3 Statistical Post-Editing

Given bilingual corpus resources we can generate a Statistical Post-Editing module (SPE). A SPE is in principal a translation module by itself, but it is trained on rule-based translations and reference data. All of our systems are based on this fully integrated SPE approach. Using this two step process will implicitly keep long distance relations and other constraints decided by the rule-based system while significantly improving phrasal fluency (Dugast, Senellart & Koehn 2007, Simard et al. 2007, Ueffing et al. 2008).

Based on the success of the hybrid approach, SYSTRAN has incorporated the technology into its product – SYSTRAN Enterprise 7. Corporate users can independently train Enterprise Server 7 to specific domains or business objectives with available – even limited – monolingual or bilingual data based on previously translated material to improve translation quality. This is the first product powered by SYSTRAN's new hybrid machine translation engine which combines the predictability and language consistency of rule-based MT with the fluency and flexibility of statistical MT that meet corporate customer quality requirements.

3 Data

All bilingual training data came from the data provided by the CWMT 2011 organizer. The monolingual data (Reuters English corpus and SogouCA Chinese corpus) provided by the CWMT 2011 were also used for the Chinese and English language modeling. As out-of-list data, we used a portion of the LDC Chinese Gigaword corpus (LDC, 3RD edition, 2007, xin-1991 to xin-2006) for training language models for the EN-ZH-NEWS and EN-ZH-SCIE contrast systems.

4 Experiments

In the experiments described below, we used Moses for decoding, GIZA++ for word alignment, and SRILM tool kit for language modeling. The model tuning was done using Minimal Error Rate Training (MERT) with BLEU4-SBP using the development sets provided by the CWMT 2011 organizer.

4.1 ZH-EN-NEWS

For training translation models, we used all of the bilingual data provided for the news task. The total number of sentences after tokenization, normalization, and filtering was approximately 5.5 million sentences. We trained a bidirectional phrase alignment table and trimmed it (Johnson et al., 2007) to suppress all unique phrase pairs before calculating the probabilities for the final phrase table. The same translation model is used for both primary and contrast systems.

For training language models, we used the English side of the provided news bilingual data and the Reuters corpus, totaling 14 million sentences (288 million words) for the primary system. For the contrast system, we additionally used the English side of the bilingual data provided for CWMT 2011 EN-ZH-SCIE task (0.8 million sentences, 22.6 million words). The order is 5-gram with interpolation, modified Kneser-Ney discounting and Good-Tuning lower cutoffs. The model perplexity was optimized with the provided news development set.

The same development set was used for system tuning. The distortion limit for reordering was set to 4 for the primary and to 6 for the contrast.

4.2 EN-ZH-NEWS

For training translation models, we used the same bilingual data used for training the Chinese-English translation model (5.5 million sentences). The Chinese tokens were segmented by word (not by character) using the SYSTRAN translation engine (Yang, Senellart and Zajac 2003). We trained a trimmed bidirectional phrase alignment table used for both primary and contrast systems.

For training language models, we used the Chinese side of the bilingual data and the SogouCA corpus for the primary system. For the contrast system, we additionally used an additional 15.5 million sentences (343 million words) from the LDC Gigaword XINHUA news corpus (LDC, 3RD edition, 2007, xin-1991 to xin-2006). The order is 5-gram with interpolation, modified Kneser-Ney discounting and Good-Tuning lower cutoffs. The model perplexity was optimized with the provided news development set.

The same development set was used for system tuning. The distortion limit for reordering was set to 6 for both primary and contrast systems.

4.3 EN-ZH-SCIE

For training translation models, we used the bilingual data provided for CWMT 2011 EN-ZH-SCIE task. The total number of sentences after tokenization, normalization, and filtering was approximately 0.8 million sentences. The Chinese tokens were segmented by word (not by character) using the SYSTRAN translation engine. We trained a trimmed bidirectional phrase alignment table used for both primary and contrast systems.

For training language models, we used the Chinese side of the bilingual data provided for CWMT 2011 EN-ZH-SCIE task and the SogouCA corpus for the primary system. For the contrast system, we additionally used the Gigaword corpus. The order is 5-gram with interpolation, modified Kneser-Ney discounting and Good-Tuning lower cutoffs. The model perplexity was optimized with the provided science and technology development set.

The same development set was used for system tuning. The distortion limit for reordering was set to 6 for both primary and contrast systems.

5 Evaluation Results

The results from the automatic evaluation scores (case-sensitive) are listed in the following tables. The primary score is BLEU4-SBP for target English and BLEU5-SBP (character-based) for target Chinese.

5.1 ZH-EN-NEWS (progress)

Our primary Chinese-English system was ranked 1st among all 28 systems including the contrast systems. The system also achieved a high score by the Woodpecker metrics. Compared with the results of CWMT2009, there was 1.17 increase in BLEU4-SBP. However, there were decreases in the Woodpecker general score as well as many scores in the subcategories. This result is puzzling, and needs to be investigated.

Table 1: Automatic results of the SYSTRAN Chinese-English systems in the ZH-EN-NEWS tasks (progress)

Systems	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	Mper	ICT
2011 Primary	0.2387	0.2558	7.8973	0.7103	0.6865	0.4906	0.3319
2011 Contrast	0.2372	0.2527	7.9136	0.7052	0.6937	0.4939	0.3222
2009 Primary	0.2260	0.2348	7.9608	0.7140	0.7151	0.4908	0.3136
2009 Contrast	0.2262	0.2348	7.9218	0.7097	0.7152	0.4939	0.3089

Table 2: Woodpecker results of the SYSTRAN Chinese-English system in the ZH-EN-NEWS tasks (progress)

Systems	General score	Source words	Source phrases	Target words	Target phrases
2011 Primary	0.2927	0.4901	0.3703	0.4876	0.2572
2009 Primary	0.2981	0.5186	0.3761	0.5029	0.2614

5.2 EN-ZH-NEWS (progress)

For the EN-ZH-NEWS progress task, our primary Chinese-English system was ranked 7th among 10 primary systems. There was 1.32 BLEU5-SBP increase compared with CWMT2009.

Table 3: Automatic results of the SYSTRAN English-Chinese systems in the EN-ZH-NEWS tasks (progress)

Systems	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
2011 Primary	0.3270	0.3430	0.2783	9.6235	9.6338	0.7743	0.6428	0.3816	0.3968
2011 Contrast	0.3308	0.3482	0.2832	9.6478	9.6586	0.7825	0.6444	0.3823	0.3936
2009 Primary	0.3138	0.3275	0.2626	9.5463	9.5557	0.7779	0.6716	0.3881	0.3679
2009 Contrast	0.3166	0.3312	0.2659	9.5856	9.5956	0.7786	0.6697	0.3862	0.3712

5.3 EN-ZH-NEWS (current)

For the EN-ZH-NEWS current task, our primary Chinese-English system was ranked 7th among 10 primary systems.

Table 4: Automatic results of the SYSTRAN English-Chinese systems in the EN-ZH-NEWS tasks (current)

Systems	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
2011 Primary	0.3114	0.3226	0.2623	9.2080	9.2198	0.7533	0.6333	0.3920	0.3671
2011 Contrast	0.3122	0.3245	0.2642	9.1580	9.1717	0.7588	0.6381	0.3975	0.3603

5.4 EN-ZH-SCIE (current)

For the EN-ZH-SCIE current task, our primary Chinese-English system was ranked 9th among 13 primary systems.

Table 5: Automatic results of the SYSTRAN English-Chinese systems in the S&T translation tasks (current)

Systems	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
Primary	0.3735	0.3840	0.3193	10.179	10.201	0.8273	0.6167	0.3187	0.3678
Contrast	0.3735	0.3843	0.3197	10.189	10.210	0.8261	0.6175	0.3189	0.3707

6 Discussions and Future Improvement Areas

The Statistical Post-Editing approach has proven again to be very efficient for improving accuracy and precision of rule-based MT systems. These good results are obtained through the simplest combination scheme, bringing together linguistic knowledge and the power of corpus-driven methods. Continued lexical and linguistic developments are also contributing to the progress of the system quality.

Our goal now is to continue separating the multiple effects and to implement dedicated and specialized statistical decision modules that would achieve individual improvements for various different areas that were obtained through statistical post-editing, with limited risks of degradations. Most of these techniques exist and are operational.

7 Conclusion

For our third participation in CWMT, our Chinese-English primary system ranked first in BLEU and second in the Woodpeck general score in the progress tasks. The hybrid approach has proven again to be every effective for improving the accuracy and fluency of the rule-based MT systems.

References

Dugast, Loic, J. Senellart, P. Koehn. 2007. Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In *Proceedings of the ACL 2nd Workshop on Machine Translation.* Prague, Czech Republic.

Dugast, Loic, J. Senellart, P. Koehn. 2009. Selective addition of corpus-extracted phrasal lexical rules to a rule-based machine translation system. In *Proceedings of the twelfth Machine Translation Summit, Ottawa, Canada*

Senellart, Jean. 2006. Boosting linguistic rule-based MT systems with corpus-based approaches. *Global Autonomous Language Exploitation PI Meeting.* Boston, USA.

Simard, Michel, N. Ueffing, P. Isabelle and R. Kuhn. 2007. Rule-based Translation With Statistical Phrase-based Post-Editing. In *Proceedings of the ACL 2nd Workshop on Machine Translation.* Prague, Czech Republic.

Ueffing, Nicola, J. Stephan, E. Matusov, L. Dugast, G. Foster, R. Kuhn, J. Senellart, J. Yang. 2008.

Tighter Integration of Rule-based and Statistical MT in a Serial System Combination. In *Proceedings of the 22nd COLING*. Manchester, United Kingdom.

Yang, Jin, J. Senellart, R. Zajac. 2003. SYSTRAN's Chinese Word Segmentation. In *Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Yang, Jin, J. Stephan, J. Senellart. 2008. SYSTRAN Chinese-English Hybrid Machine Translation. In *Proceedings of the 4th China Workshop on Machine Translation, Beijing, China.*

Yang, Jin, S. Enoue, J. Senellart. 2009. SYSTRAN Chinese-English and English-Chinese Hybrid Machine Translation Systems. In *Proceedings of the 5th China Workshop on Machine Translation, Nanjing, China.*