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

Jin Yang, Satoshi Enoue

SYSTRAN Software, Inc.
9333 Genesee Ave. Suite PL1
San Diego, CA 92121, USA

{jyang, enoue}@systransoft.com

Jean Senellart, Tristan Croiset

SYSTRAN SA
La Grande Arche, 1, Parvis de la Défense
92044 Paris La Défense Cedex，France

{senellart,croiset}@systran.fr

Abstract: This report describes both of SYSTRAN’s Chinese-English and English-Chinese machine translation systems that participated in the CWMT2009 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 perform 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 CWMT2009 evaluation. Our primary systems were top-ranked in the evaluation 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统

Jin Yang, Satoshi Enoue

SYSTRAN Software, Inc.
9333 Genesee Ave. Suite PL1
San Diego, CA 92121, USA

{jyang, enoue}@systransoft.com

Jean Senellart, Tristan Croiset

SYSTRAN SA
La Grande Arche, 1, Parvis de la Défense
92044 Paris La Défense Cedex，France

{senellart,croiset}@systran.fr

摘要： 本文介绍了 SYSTRAN 参加 CWMT2009 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 CWMT2009 中的评测结果。SYSTRAN 汉英和英汉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 five 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 participated in the CWMT2008 evaluation, ranking third in BLEU and first in NIST (Yang, Stephan, Senellart 2008).

The techniques used in the Chinese-English system for CWMT2008 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 year, we continued improving and refining these techniques, and experimenting and expanding to more language pairs and domains. In the CWMT2009 evaluation, we participated in the Chinese-English news translation single system tasks (ZH-EN-NEWS-SINGL), English-Chinese news machine translation (EN-ZH-NEWS-MT) and S&T machine translation (EN-ZH-SCIE-MT) tasks. 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

We have two submissions for the CWMT2009 Chinese-English single system tasks ZH-EN-NEWS-SINGL: “ce-news-systran-primary-systema” and “ce-news-systran-contrast-systemb.” The same approach was used to produce the two different outputs. The difference is that the primary system utilizes all the provided bilingual training data (approximately 3.4 million sentences), whereas the contrast system utilizes a portion of the provided bilingual data (approximately 2 million sentences).

We have two submissions for the English-Chinese news machine translation tasks EN-ZH-NEWS-MT: “ec-news-systran-primary-systema” and “ec-news-systran-contrast-systemb.”; and two for the S&T machine translation tasks EN-ZH-SCIE-MT: “ec-tech-systran-primary-systema” and “ec-tech-systran-contrast-systemb”. Like the Chinese-English systems, we used the same approach with slightly different monolingual training data.

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 development and modernization.

The development of the SYSTRAN Chinese-English MT system began in August 1994. Work on linguistic analysis has been continuing over the years, with steady improvement. Recent development concentrates on incorporating statistical techniques in the various components of the system: corpus-based monolingual and bilingual terminology extractions, incorporating corpus evidence in the linguistic rules (Senellart 2006), producing translation lattices etc. The current RBMT Chinese-English system contains over 1.2 million bilingual words, expressions, and linguistic rules.

The SYSTRAN English-Chinese system was built based on the existing SYSTRAN English parser and dictionaries. 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 Chinese generation needs more work.

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).

3. Data

All bilingual training data came from the data provided by the CWMT2009 organizer. The monolingual data (i.e. Reuters English corpus and SogouCA Chinese corpus) provided by the CWMT2009 were also used in the Chinese and English language models. No other data was used for the Chinese-English systems. We used a portion of the LDC Chinese Gigaword corpus (Xinhua news portion) in the training of the English-Chinese contrast systems only. A detailed list is given below:

3.1 ZH-EN-NEWS-SINGL tasks

3.1.1 Bilingual texts

- The provided CWMT 2009 bilingual data was tokenized, normalized and filtered
- Chinese tokens were segmented by word (not by character) using the SYSTRAN translation engine (Yang, Senellart and Zajac 2003)
- Approximately 3.4 million sentences (70.8 million English words) for the primary system
- A portion of the bilingual data (i.e. 40w_CHN-ENG, 60w-CHN-ENG, CLDC200306, DATUM and HIT-IR, about 2 million sentences) was used for the contrast system.

3.1.2 Monolingual English texts

- The English texts in the CWMT 2009 bilingual data were tokenized, normalized, and filtered
- The Reuters English corpus was sentence segmented, tokenized, normalized and filtered
- True-casing and no entity replaced (as-is)
- Approximately 11.7 million sentences (244 million words) in total (3.4 million sentences from the English part of the bilingual training data; 8.2 million sentences from the Reuters English corpus).

3.2 EN-ZH-NEWS-MT & EN-ZH-SCIE-MT tasks

3.2.1 Bilingual Corpora

- The provided CWMT 2009 data was tokenized, normalized and filtered
- Chinese tokens were segmented by word using SYSTRAN translation engine
- Approximately 3.4 million sentences (70.8 million English words) in total

3.2.2 Monolingual Chinese Corpora

- The Chinese texts in the CWMT 2009 bilingual data were tokenized, normalized, and filtered
- The SougouCA corpus was sentence segmented, tokenized, normalized and filtered
- The Chinese tokens were segmented by word using the SYSTRAN translation engine
- For the primary systems, approximately 11.2 million sentences (240 million words)
- For the contrast systems, an additional 13.6 million sentences from the LDC Gigaword XINHUA news corpus (LDC, 3RD edition, 2007, xin-1991 to xin-2006) were also used for building the language modals. The total number of sentences is 24.8 million sentences (556 million words).

4. Experiments and Evaluation

4.1 Experiments

4.1.1 ZH-EN-NEWS-SINGL tasks

For each of the Chinese-English systems, we trained an SPE module consisting of a language model and a translation model. The language model was trained using all of the CWMT 2009 data (11.7 million sentences, 244 million words) with the maximum order of 5-gram. The model was interpolated and smoothed with Kneser-Ney discounting and Good-Tuning lower cutoffs, and its perplexity was optimized on the CWMT 2008 evaluation text. Both of the primary and contrast systems used the same language model.

For the translation models, we used GIZA++ for training bidirectional phrase alignment tables (phrase table and re-ordering table) with the maximum order of 5-gram. The phrase table was trimmed (Johnson et al., 2007) to suppress all unique phrase pairs before calculating the probabilities for the final phrase table. The re-ordering table was trained with the distortion value of 4. The primary system used the translation model trained with all available bilingual data (3.4 million sentences). The contrast system used the translation model trained with a subset of 2 million sentences. The tuning was done on the CWMT 2008 evaluation set using Moses minimal error rate tuning.

4.1.2 EN-ZH-NEWS-MT tasks

For the English-Chinese primary system, we trained a language model using the CWMT 2009 Chinese monolingual data (11.2 million sentences, 240 million words). For the contrast system, we trained a language model using an additional 13.6 million sentences from the LDC Gigaword XINHUA news corpus (LDC, 3RD edition, 2007, xin-1991 to xin-2006), totaling 24.8 million sentences (556 million words). Both models were trained with the maximum order of 5-gram, and interpolated and smoothed with Kneser-Ney discounting and Good-Tuning lower cutoffs. Their perplexity was optimized on the CWMT 2008 news evaluation text.

For the translation model, we used the CWMT 2009 bilingual corpus (3.4 million sentences), and both of the primary and contrast systems used the same translation model. The bidirectional phrase alignment tables were trained with the maximum order of 5-gram with no trimming. The re-ordering table was trained with the distortion value of 6. The tuning was done on the CWMT 2008 news evaluation set using Moses minimal error rate tuning.

4.1.3 EN-ZH-SCIE-MT tasks

The language models were trained in the same way as the English-Chinese news systems but their perplexity was optimized on the CWMT 2008 science and technology evaluation text instead of the news set. The primary system used a language model trained with the CWMT 2009 data only and the contrast system used a model trained with additional 13.6 million sentences from the LDC Gigaword XINHUA news corpus.

For the translation model, we used the same model used for the English-Chinese news translation but the tuning was done on the CWMT 2008 science and technology evaluation set using Moses minimal error rate tuning.

4.2 Automatic Evaluation Results

The evaluation results from the automatic evaluation metrics (case-sensitive) are listed in the following tables.

4.2.1 ZH-EN-NEWS-SINGL tasks

Systems	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT
Primary	0.2260	0.2348	7.9608	0.7140	0.7151	0.4908	0.3136
Contrast	0.2262	0.2348	7.9218	0.7097	0.7152	0.4939	0.3089

Table 1 Automatic evaluation results of the SYSTRAN Chinese-English systems

As mentioned above, the only difference between the Chinese-English primary and contrast systems output is that the primary systems used the complete bilingual training data (3.4 million sentences) whereas the contrast system used only a portion of the bilingual data (2.0 million sentences). The results from the both system are quite similar.

Systems	General score	Source words	Source phrases	Target words	Target phrases
Primary	0.2981	0.5186	0.3761	0.5029	0.2614

Table 2 Woodpecker evaluation results of the SYSTRAN Chinese-English primary system

Our primary Chinese-English system ranked first in the General Score of the Woodpecker metrics, and also ranked first or high in various subcategories. This is consistent with our experience: the SPE output has a better control over linguistic structures of the sentences. A pure ngram-based automatic evaluation metrics (e.g. BLEU/NIST) under-evaluates the quality of the translation. It is also our observation that a high BLEU score doesn’t necessarily mean high translation adequacy (Zhao et al. 2009).

4.2.2 EN-ZH-NEWS-MT tasks

Systems	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER
Primary	0.3138	0.3275	0.2626	9.5463	9.5557	0.7779	0.6716	0.3881
Contrast	0.3166	0.3312	0.2659	9.5856	9.5956	0.7786	0.6697	0.3862

Table 3 Automatic evaluation results of the SYSTRAN English-Chinese systems in the news translation tasks

The contrast system has slightly better results since additional monolingual texts were used in the language model.

4.2.3 EN-ZH-SCIE-MT tasks

systems	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
primary	0.4828	0.4876	0.4204	11.3342	11.3579	0.8814	0.4902	0.2884	0.5145
contrast	0.4826	0.4875	0.4201	11.3393	11.3631	0.8807	0.4900	0.2885	0.5155

Table 4 Automatic evaluation result of the SYSTRAN English-Chinese systems in the S&T translation tasks

The results of the two systems are quite similar. The additional monolingual Chinese data used in the language model (e.g. Xinhua news) doesn't bring much improvement or impact to the translation. This was as expected. We tried to show that additional in-domain data in the language model improves translation quality, but not out-of-domain data.

4.3 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. This proves that the potential behind this combination is huge.

Qualitative analysis of the results shows that the Statistical Post-Editing is contributing to many different areas, including better meaning selection, improved local word re-ordering, preposition choice and determiner selection (for English target). At the same time, many new types of side-effect are also observed, including word deletion, or inversed meanings and translation of an expression by a totally unrelated expression.

Our goal now is to separate 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. The main challenge is to have them work together with the rule-based engine. This requires: a) dynamically assigning weights to the different rules in the system; b) exploring multiple hypotheses in parallel; and c) implementing a more generic support of exception handling in the rule description.

5. Conclusion

For our second participation to CWMT, our Chinese-English primary system ranked third in BLEU, first in NIST and GTM scores, and first in the Woodpeck general score in the single system tasks. Our English-Chinese primary system ranked second in terms of BLEU score in the technical machine translation tasks. This shows that Hybrid Technology is particularly effective in that specific context, further consolidating what we already demonstrated in CWMT2008 and in other MT evaluation campaigns, and providing real quality for the corporate user.

For future machine translation evaluations, we would welcome the introduction of domain-specific Science and Technology translation tasks for both Chinese-English and English-Chinese machine translation tasks. Although there is already an EN-ZH-SCIE-MT task, we think that the domain is still quite general and broad. A more limited domain is desirable. This would more accurately match the needs for MT from corporate users. Also we noticed again that the BLEU scores of the English S&T tasks are much higher than the ones in the news tasks. This is consistent with our observation that the S&T texts are less complicated than the newswire texts, thus a higher BLEU score can be obtained from MT systems. We would like to see how much higher Chinese-English and English-Chinese MT systems can achieve in a relatively limited domain, as this is one of the application areas for MT users. And lastly, we support exploring, experimenting and utilizing other evaluation metrics which are different with the simple n-gram based metrics, especially human evaluation. Although human evaluation is costly, time-consuming and somewhat subjective, humans are ultimately the user of machine translation output.

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.

Zhao, Hongmei, J. Xie, Q. Liu, Y. Lü, D. Zhang, M. Li. 2009. Introduction to China's CWMT2008 Machine Translation Evaluation. In *Proceedings of the twelfth Machine Translation Summit, Ottawa, Canada.*