Tighter Integration of Rule-based and Statistical MT in Serial System Combination

Nicola Ueffing[1], Jens Stephan[2], Evgeny Matusov[3], Loïc Dugast[2], George Foster[1], Roland Kuhn[1], Jean Senellart[2], and Jin Yang[2]
[1]Institute for Information Technology
National Research Council of Canada (NRC)
Gatineau, Québec, Canada

[2]SYSTRAN SA
Paris, France

[3]RWTH Aachen University
Aachen, Germany

Abstract

Recent papers have described machine translation (MT) based on an *automatic post-editing* or *serial combination* strategy whereby the input language is first translated into the target language by a rule-based MT (RBMT) system, then the target language output is automatically post-edited by a phrase-based statistical machine translation (SMT) system. This approach has been shown to improve MT quality over RBMT or SMT alone. In this previous work, there was a very loose coupling between the two systems: the SMT system only had access to the final 1-best translations from RBMT. Furthermore, the previous work involved European language pairs and relatively small training corpora. In this paper, we describe a more tightly integrated serial combination for the Chinese-to-English MT task. We will present experimental evaluation results on the 2008 NIST constrained data track where a significant gain in terms of both automatic and subjective metrics is achieved through the tighter coupling of the two systems.

1 Introduction

MT systems tend to make the same errors over and over again. A recent paper (Simard *et al.*, 2007a) proposed that a phrase-based SMT system be trained on the *manually* post-edited output of an RBMT system to become an *automatic* post-editor for the output of that system. Thus, the SMT system learns how to "translate" the output of the RBMT system to another version of these same translations that will more closely resemble translations produced by qualified humans. In the above-cited paper, this serial combination approach was shown to produce large improvements in translation performance for the English-French language pair (both directions) compared not only to the RBMT system but also compared to the SMT system trained to translate directly from the source language. However, these experiments were carried out in an unusual context, where a large corpus of manually corrected RBMT output was available.

Simard *et al.* (2007b) examines the impact of an automatic post-editor trained on more commonly available data. The source-language part of a bilingual parallel corpus is input to an RBMT system, thus creating a corpus where each manually produced target-language sentence is aligned with an *independently* generated automatic translation of the same source sentence. Training an automatic post-editor on this kind of corpus is quite different from training one on a corpus containing corrected MT output: a manually post-edited translation will have similar word choices and word order to the original translation, while two independently generated translations of the same text may differ radically.

The experiments in (Simard *et al.* 2007b) are carried out on English-French data (both

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 913–920
Manchester, August 2008

directions) from the 2007 Workshop on Statistical Machine Translation[1]. SMT post-editing significantly improves the BLEU scores of the RBMT system for both translation directions and for both domains tested, the News Commentary domain (50K training sentence pairs) and the Europarl domain (1.3M training sentence pairs). In terms of BLEU score, the hybrid system and the pure SMT system are essentially tied in the Europarl domain (for both language directions), while in the News Commentary domain, the hybrid system outperforms the pure SMT system by about 3 BLEU points for English-to-French and by about 1.5 points for French-to-English.

These results suggest that if one has little training data, serial RBMT-SMT combination is the best strategy, while with more training data, serial combination or the use of a pure SMT system yields equivalent results. However, the BLEU score may underestimate the advantages of serial system combination. **Table 1** shows some data from the 2007 Workshop on SMT (Callison-Burch *et al.* 2007) involving an RBMT system from the company Systran (Yang *et al.* 2003), and the SMT system PORTAGE from National Research Council Canada (Ueffing *et al.* 2007). The BLEU score is shown in italics along with human rankings of system output (for Europarl). Although the BLEU scores of the pure SMT system and the hybrid system are nearly identical, human evaluators ranked the hybrid system first of eight English-to-French systems and second of seven French-to-English systems, as compared to inferior rankings for SMT or RBMT alone: the serial combination makes a more favorable impression on human evaluators than BLEU would predict.

System	En→Fr: *(BLEU)* **Rank/8**	Fr→En: (BLEU) **Rank/7**
RBMT (Systran)	*(23.3)* **6**	*(21.1)* **6**
SMT (PORTAGE)	*(29.4)* **5**	*(31.2)* **5**
RBMT→SMT	*(30.1)* **1**	*(31.3)* **2**

Table 1: BLEU score and human rankings for En→Fr and Fr→En (SMT 2007 Europarl data)

Dugast *et al.* (2007) describes similar experiments for the other four language pairs in the WMT 2007 evaluation: the Systran RBMT system is serially combined with the Moses SMT system. This paper also presents a qualitative analysis of the changes made by the SMT post-editor on the RBMT output.

The experiments described above involve European language pairs and relatively small amounts of training data. We decided to carry out serial combination experiments on the Chinese-to-English translation task, for which large amounts of training data are available. As explained above, the serial combination is built by first having the RBMT system translate all Chinese training data into English. The SMT component is then trained on sentence pairs where each RBMT English output is aligned with the English reference sentence for the original Chinese sentence. Thus, the SMT system is capable of "translating" RBMT output into even better English.

The Chinese-English experiments were carried out in the context of both the GALE project and the 2008 NIST Chinese-to-English MT evaluation. The integration of the two components of the serial combination was deepened by having the RBMT component break its output into chunks, with some of these chunks being annotated with a confidence level; chunks with low confidence are more likely to be changed by the SMT component.

2 Description of RBMT System

The RBMT system in these experiments is the Systran Chinese-English system. This system is described extensively in (Yang *et al.* 2003). In particular, it uses a rule-based word-boundary module and a bilingual lexicon with about 1.2M entries, containing words, expressions and rules. This system is specifically geared to translation in the domain of science and technology. Recent evolution of this system has focused on developing structured markup making it possible to monitor rule application and to interact with the rule engine (Attnäs *et al* 2005).

3 Description of SMT System

The SMT system in these experiments is the PORTAGE system from the National Research Council of Canada (Ueffing *et al.* 2007). PORTAGE is a standard phrase-based system that carries out beam search decoding using a loglinear model. Major features in this

[1] http://www.statmt.org/wmt07/

loglinear model include phrase tables derived from symmetrized IBM2 and HMM word alignments (with a phrase length limit of seven words), and a static 4-gram language model (LM) trained on the Gigaword corpus using the SRILM toolkit (Stolcke 2002). An important feature is an adapted 4-gram LM derived from the parallel corpus using the technique of (Foster and Kuhn 2007); it is a linear mixture of N-gram models trained on parallel subcorpora (Hong Kong Hansard, FBIS, UN, *etc.*). Other features are word count and phrase-displacement distortion. Parameter tuning of this loglinear combination is performed using Och's max-BLEU algorithm (Och 2003) with a closest-match brevity penalty. Decoding uses the cube-pruning algorithm of (Huang and Chiang 2007) with a seven-word distortion limit.

4 Annotating RBMT Output

The output of the RBMT system is broken into annotated "chunks"; some of these chunks have confidence values assigned to them. The annotations come from several steps of the RBMT translation process. Currently, the following five types of chunk are annotated:

- Named entities, dates, numbers, *etc.* (one chunk type) – a confidence value is assigned based partly on how well the hypothesized entity matches source-language entity patterns and partly on the extent to which the entity matches entries in the bilingual lexicon (*e.g.*, a hypothesized compound proper name is assigned a higher confidence if it is made up of individual proper names).
- Unknown words or unlikely sequence of short words (two different chunk types) – unknown words identify words that are not in the lexicon or that have no meaning. For Chinese, unknown words are rare since any expression can be decomposed into single characters, each of which is almost always a possible word. However, a sequence of single-character words is unlikely in Chinese and such a sequence is also detected and annotated.
- 'Strong' rules output – two different chunk types identifying *e.g.*, rules based on a long distance syntactic relationship, or a long multiword expression. These chunks are very reliable.

This information is passed on to the SMT system in two different ways. In both methods, each chunk output by the RBMT consists of a translation into the target language and a probability for this translation. The SMT decoder hypothesizes both the translation provided in the RBMT output and translations provided by the phrase tables. The translation yielding the highest overall sentence score will be the one chosen. The two different ways of using the markup are:

1. The information in the chunk is inserted into the phrase table with the given probability, which replaces any other probability for that phrase pair.
2. As in **1**, but in addition, each chunk type is defined as a decoder feature outputting a confidence value. The weight assigned to this feature is

RBMT output	<rule><entity type="HUMANS" confidence="1"><expression type="1">US Secretary of State Powell</expression></entity></rule> the other day the visit which carried on to <entity type="GPE" confidence="1">Russia</entity> is still a public opinion widely attention focal point.
Input to baseline SMT:	us secretary of state powell the other day the visit which carried on to russia is still a public opinion widely attention focal point .
Output of baseline:	us secretary of state colin powell said the other day visit to russia is still a wide public opinion focus .
Input to SMT (markup strategy 1B):	<rule target="us secretary of state powell" prob="1">us secretary of state powell</rule> the other day the visit which carried on to <rule target="russia" prob="1">russia</rule> is still a public opinion widely attention focal point .
Output from SMT (markup strategy 1B):	us secretary of state powell 's visit to russia is still a public opinion generally focus of attention .
Reference:	u.s. secretary of state powell 's recent visit to russia remains a focus of widespread public opinion .

Table 2: Example sentence from the test set showing annotated RBMT output and its use in SMT.

found by the optimization algorithm (see Sec.3).

Orthogonal to these two approaches, we must decide whether to use

A. all five types of chunks

B. or only those chunks that have confidence values associated with them.

In strategy B, each chunk's confidence value is used directly; in strategy A, missing confidence values are assigned manually.

5 Experiments

5.1 GALE system combination

Recent work has focused on computation of a consensus translation from the outputs of multiple MT systems (Matusov *et al*. 2006, Rosti *et al*. 2007). This approach to system combination was applied to Chinese-English data available within the DARPA-sponsored GALE project by researchers at RWTH. One of the systems in the RWTH system combination is itself a combination: our hybrid RBMT→SMT (Systran→PORTAGE) system. This version of the hybrid system had loosely coupled components (confidence information isn't output by the RBMT). The other systems in this "parallel" combination were all pure SMT systems. In earlier experiments, an attempt had been made to include the RBMT system (Systran) by itself, but it always received a negligible weight in the parallel system combination.

Although the RBMT→SMT hybrid system is not the best among the systems combined, as measured either by BLEU score or TER score, it is always assigned a heavy weight by the system combination algorithm (usually the heaviest weight). Since the heavy weights assigned to the RBMT→SMT system could be dismissed as an artifact, experiments were also done (using 554 newswire sentences as test data) in which one system at a time was dropped from the five-way system combination. The results are shown in **Table 3** for case-sensitive BLEU and TER (high BLEU and low TER are desirable). Note that in addition to RBMT→SMT we also trained a "direct" Chinese-English system using the same SMT software (PORTAGE). When RBMT→SMT is dropped, the system combination performs worse by 0.9 BLEU and 2.6 TER. This performance drop is worse than that caused by dropping any other single system. One may infer that despite its unremarkable BLEU and TER scores taken in isolation, the RBMT→SMT serial combination provides information that complements that of the other systems (which were all SMT systems).

Description	BLEU	TER
All (system comb).	**17.0**	**65.4**
drop PORTAGE	17.0	64.9
drop system 1	17.2	65.0
drop system 2	17.2	65.6
drop system 3	16.6	65.4
drop Systran→PORTAGE	**16.1**	**68.0**

Table 3: "drop-one" in GALE system combination.

5.2 Markup experiments

We experimentally evaluated the methods for using markup proposed in Section 4 above on the so-called constrained data track for Chinese-English of the 2008 NIST MT evaluation. The evaluation was carried out on the test sets from 2004 and 2008. **Table 2** shows an example sentence from the test set. The first row presents the RBMT output in which different chunks are annotated as described in Section 4. This annotation can be nested as the first chunk shows. In this example, both chunks are assigned confidence 1. Note that other chunks in the RBMT output have lower confidence, and many of them are not annotated with confidence at all.

The RBMT system has recognized that "US Secretary of State Powell" is a human with confidence 1, and also a multiword expression. This markup in the RBMT output is stripped off before being input to the baseline SMT system. The example also shows the input to the SMT system under markup strategy 1B described in Section 4, and the output from that strategy. Finally, a reference translation is shown. Note that the baseline system inserted two spurious words, "colin" and "said" into the translation. By contrast, when markup strategy 1B was applied, the SMT system decided to retain the expression "us secretary of state powell" passed to it in the RBMT output (and this had the side-effect of removing the spurious word "said" and generating a better translation for the rest of the sentence).

Table 4 presents the results of the four different ways of using the markup of the RBMT output described in Section 4, along with baseline results without markup. The results show that the hybrid system is improved through the

use of markup. There is no clear preference for one of the two approaches to using the markup (phrase table insertion only *vs.* decoder feature). However, it is clear that using only the chunks with assigned confidence performs better than the use of all chunk types. This might be due to the fact that the chunks with assigned confidence include many named entities which are particularly hard to translate for MT systems. Thus, having confidence values expressing the reliability of the translations improves the hybrid system significantly.

Description	NIST04	NIST08
Baseline (no markup)	34.0	25.0
Phrase table insertion of markup (approach 1, p.3)		
All markup (A)	34.6	24.7
Only markup with confidence (B)	**35.0**	25.8
Markup as decoder feature (approach 2, p.3)		
All markup (A)	34.4	25.3
Only markup with confidence (B)	34.6	**26.0**

Table 4 Translation quality in terms of BLEU score on NIST data

In addition to the automatic evaluation, we had an English native speaker manually assess the translations of the first 100 sentences of the 2004 test set for the baseline and "strategy 1B" output. Outputs were shuffled for each sentence to get reliable blind evaluation of the two systems. Fluency and adequacy were judged on a scale of 0-5 (0=worst, 5=best). The results are shown in **Table 5**, and they support the outcome of the automatic evaluation: the hybrid system which uses the markup is significantly preferred to the one which does not use the markup. The evaluation scores for fluency and adequacy increase by 10% and 13% relative, respectively. The table also shows which of the two translations receives a higher evaluation score, and this clearly indicates the superiority of the system with markup: in a third of all cases, this system was judged better than the baseline serial combination system.

	Fluency	Adequacy
Baseline (no markup)	1.9	2.2
Strategy 1B	2.1 (+10%)	2.5 (+13%)
Comparison no markup vs. strategy 1B: 1B worse/equal/better	10/58/32	6/64/30

Table 5 NIST04 – Qualitative human evaluation

5.3 Participation in NIST MT 2008

Both the pure statistical system PORTAGE and the serial combination Systran→PORTAGE (with markup strategy 1B) participated in the 2008 NIST Chinese-English MT evaluation. Both systems were trained on data from the constrained data track (thus, to create the Systran→PORTAGE system all the Chinese constrained-track training data was translated by Systran, and the English output aligned with the English references was then used to train PORTAGE). Both systems placed roughly in the middle of competing systems according to BLEU score, with PORTAGE obtaining a score of 24.58 and Systran→PORTAGE obtaining a score of 25.23. (Internal tests of Systran alone on the same data give 18.21). However, Systran→PORTAGE excelled according to an important human evaluation metric. In cases where human evaluators had ranked a translated sentence sufficiently highly, they were asked an additional question as to whether the translation captured all the needed information. Systran→PORTAGE obtained the highest number of "yes" answers (35.82% of all sentences) among all 14 Chinese-English systems evaluated.

6 Discussion and Future Work

In this paper, we built on earlier work showing the advantages of coupling RBMT and SMT in a serial combination RBMT→SMT where the SMT system "translates" output from the RBMT system into the target language. The conclusions to be drawn about the performance of our serial system combination Systran→PORTAGE depend on whether one consults automatic metrics like BLEU or scores assigned by human evaluators. According to BLEU, Systran→PORTAGE is typically superior to Systran alone and roughly

tied with PORTAGE (if enough data are available to train the latter). According to human evaluators, however, Systran→PORTAGE is clearly superior to either of its "parents". This can be seen from the results in (Callison-Burch *et al.* 2007) and from the fact that Systran→PORTAGE placed first among all systems according to the "additional qualitative question" metric in the NIST 2008 Chinese-English MT evaluation (PORTAGE alone did not do well according to this metric; Systran alone was not evaluated, but has a BLEU score low enough to ensure it would have done poorly). Another result of this paper (in Section 5.1) is that the hybrid system appears to contribute information complementary to that provided by pure SMT systems, as judged by its important contribution to a "parallel combination" system.

Why does Systran→PORTAGE yield superior translations to those produced by either Systran or PORTAGE on its own? We have not yet conducted a thorough study to answer this question. However, our perception is that Systran's syntactic rules preserve structure and allow long-distance movement of constituents. On the other hand, this system has a weak model of the target language (English), yielding disfluencies. Post-editing by PORTAGE preserves syntactically motivated rearrangements while correcting disfluencies in the output and making more appropriate lexical choices. Although we have not systematiccally studied this behaviour, we have observed it in several examples, such as the one shown in **Table 6**.

Description	**Sentence**
Chinese	欧盟 官员 现在 最 担心 的 是 即将 从 非洲 飞回 的 候鸟。
Systran	What the EU officials most were worried now is soon the migratory bird which flies back from Africa.
PORTAGE	EU officials are most concerned about is coming from Africa flew back to the migratory birds.
Systran→ PORTAGE	The EU officials are most worried about the migratory birds that fly back from Africa.

Table 6: three translations of a Chinese sentence

The main purpose of the current paper is to show that tighter integration of the two systems further improves the hybrid system, both in terms of automatic and human evaluation. This was achieved by having the RBMT system (Systran) mark up its output in such a way that the SMT system (PORTAGE) could distinguish between more and less reliable chunks in the output.

The integration of the two systems could be further deepened, *e.g.*, by passing on more complex annotation between the two systems, or by exploring the fact that source and target language for the SMT system are the same. This provides an interesting alternative to training data gigantism: rather than using ever more data, we make better use of existing data.

Moreover, the two combined MT paradigms can benefit from each other. For instance, the RBMT confidence estimation could be data-driven, and the SMT post-editing could be constrained by basic linguistic rules.

Acknowledgement

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

M. Attnäs, P. Senellart and J. Senellart. Integration of SYSTRAN MT Systems in an Open Workflow. *Proc. MT Summit X*, 2005.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. (Meta-) Evaluation of Machine Translation. *Proc. 2nd ACL Workshop on Statistical Machine Translation*, pp. 136-158, June 2007.

L. Dugast, J. Senellart and P. Koehn. Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. *Proc. 2nd ACL Workshop on Statistical Machine Translation*, pp. 220-223, June 2007.

G. Foster and R. Kuhn. Mixture-Model Adaptation for SMT. *Proc. 2nd ACL Workshop on Statistical Machine Translation*, pp. 128-135, June 2007.

L. Huang and D. Chiang. Forest Rescoring: Faster Decoding with Integrated Language Models. *Proc. ACL 2007*, pp. 144-151, June 2007.

E. Matusov, N. Ueffing and H. Ney. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. *Proc.EACL 2006*, pp. 33-40, April 2006.

F. Och. Minimum Error Rate Training for Statistical Machine Translation. *Proc. ACL 2003*, July 2003.

A.-V. Rosti, S. Matsoukas, and R. Schwartz. Improved Word-Level System Combination for Machine Translation. *Proc. ACL 2007*, June 2007.

M. Simard, C. Goutte, and P. Isabelle. Statistical Phrase-based Post-Editing. *Proc. HLT-NAACL*, pp. 508-515, April 2007a.

M. Simard, N. Ueffing, P. Isabelle and R. Kuhn. Rule-Based Translation with Statistical Phrase-based Post-editing. *Proc. 2nd ACL Workshop on Statistical Machine Translation*, pp. 203-206, June 2007b.

A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. *Proc. ICSLP*, pp. 901-904, Septermber 2002.

N. Ueffing, M. Simard, S. Larkin, and J. H. Johnson. NRC's PORTAGE System for WMT 2007. *Proc. 2nd ACL Workshop on Statistical Machine Translation*, pp. 185-188, June 2007.

J. Yang, J. Senellart and R. Zajac. SYSTRAN's Chinese Word Segmentation. *Proc. Second SIGHAN Workshop on Chinese Language Processing*, July 2003.

This page is intentionally blank.