PROMT DeepHybrid system for WMT12 shared translation task

Alexander Molchanov
PROMT LLC
16A, Dobrolubova av.
197198, St. Petersburg, Russia
Alexander.Molchanov@promt.ru

Abstract

This paper describes the PROMT submission for the WMT12 shared translation task. We participated in two language pairs: English-French and English-Spanish. The translations were made using the PROMT DeepHybrid engine, which is the first hybrid version of the PROMT system. We report on improvements over our baseline RBMT output both in terms of automatic evaluation metrics and linguistic analysis.

1 Introduction

In this paper we present the PROMT DeepHybrid submission for WMT12 shared translation task for two language pairs: English-French and English-Spanish.

A common approach to create hybrid machine translation (MT) systems on the basis of rule-based machine translation (RBMT) systems is to build a statistical phrase-based post-editing (SPE) system using state-of-the-art SMT technologies (see Simard et al. 2007). An SPE system views the output of the RBMT system as the source language, and reference human translations as the target language. SPE systems are used to correct typical mistakes of the RBMT output and to adapt RBMT systems to specific domains. (Dugast et al. 2007) report on good results both in terms of automatic evaluation metrics and human evaluation for the SPE systems based on PORTAGE (Sadat et al. 2005) and Moses (Koehn et al. 2007). However, an SMT model in fact makes translation output less predictable in comparison with RBMT output. We propose a different approach to hybrid MT technology. We developed and incorporated the SPE component into our translation system (the statistical post-editing data is controlled by the PROMT hybrid translation engine). Besides, we have an internal language model (LM) component that scores the generated translation candidates.

The remainder of the paper is organized as follows: in section 2 we provide the detailed description of our hybrid MT technology. In section 3 we evaluate the performance of the technology on two language pairs: English-French and English-Spanish. We gain improvements over the baseline RBMT system in terms of BLEU score on test sets. We also introduce the results of linguistic evaluation performed by our experts. Section 4 summarizes the key findings and outlines open issues for future work.

2 System description

The PROMT DeepHybrid system is based on our RBMT engine. The baseline system has been augmented with several modules for hybrid training and translation. The training technology is fully automated, but each step can be fulfilled and tuned separately.

2.1 Rule-based component

PROMT covers 51 language pairs for 13 different source languages. Our system is traditionally classified as a ‘rule-based’ system. PROMT uses morphosyntactic analyzers to analyze the source sentence and transfer rules to translate the sentence

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 345–348,
Montréal, Canada, June 7-8, 2012.

into the target language. The crucial component of our system is the PROMT bilingual dictionaries which contain up to 250K entries for each language pair. Each entry is supplied with various linguistic (lexical and grammatical, morphological, semantic) features. Besides the 'baseline' dictionaries the PROMT system has a large number of domain-specific dictionaries.

2.2 Parallel corpus processing

We have a specific component for processing parallel corpora before training the hybrid system. This component can process data in plain text and XML formats. We also perform substantial data filtering. All punctuation and special symbols (ligatures etc.) are normalized. The length of the words in a sentence and the length of sentences are taken into account (sentences having length above a set threshold are discarded). All duplicated sentences are discarded as well. On top of that, we remove parallel segments with different number of sentences because such segments corrupt phrase alignment. Strings containing few alphabetic symbols and untranslated sentences are filtered out from the parallel corpus.

2.3 Automated dictionary extraction

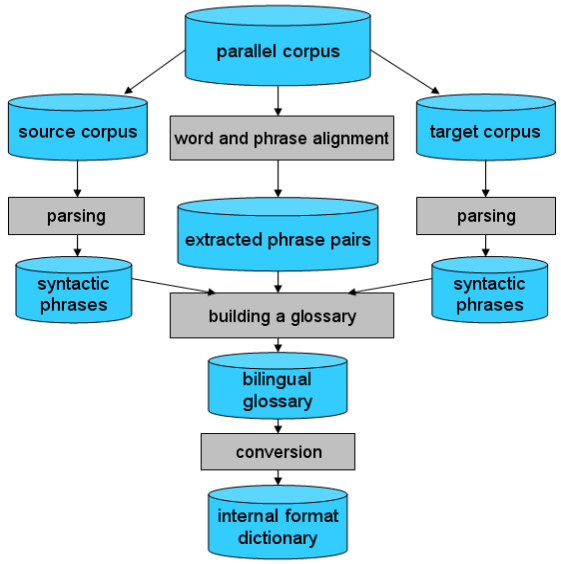

Figure 1. Dictionary extraction pipeline.

The extraction technology is shown in figure 1. The whole process can be subdivided into two separate tasks: 1) statistical alignment of a parallel corpus 2) extraction of syntactic phrases from the source and target sides of the parallel corpus. We then combine the results of these tasks to extract bilingual terminology. We use GIZA++ to perform the word alignment (Och and Ney, 2003). Then we use the common heuristics to extract parallel phrase pairs (Koehn et al. 2007). We use the PROMT parsers to extract grammatically correct phrases from source and target sides of the parallel corpora. PROMT parsers are rule-based multi-level morphosyntactic analyzers. Parsers extract noun phrases, verb phrases and adverbial phrases. The extraction is done as follows: each sentence of the corpus is parsed, a parse tree is created, the extracted syntactic phrases are stored in memory; after the whole corpus is processed, all extracted phrases are lemmatized and presented in a list. Each phrase is supplied with a set of linguistic features (part of speech, lemma, lemma frequency etc.). The next step is building a bilingual glossary using two sets of syntactic phrases extracted from the source and the target sides of the parallel corpus on the one hand and a statistically aligned set of phrase pairs on the other hand. We do not add geographic names, proper names and named entities (dates etc.) to the glossary because they are well processed by the RBMT engine.

2.4 Statistical phrase-based post-editing

The technology of obtaining data for statistical post-editing is standard. We translate the source corpus using the RBMT engine. Then we align the MT corpus and the target corpus using GIZA++ and extract parallel phrase pairs to obtain a phrase-table. Then the phrase-table is filtered. The phrase length and translation probability are taken into account. Only pairs having length of the source phrase from three to seven words are selected. This specific length range was chosen according to the detailed analysis of the resulting hybrid MT quality performed by our linguists. The selected phrase pairs are stored in the special SPE component of the hybrid engine and are used to apply post-editing to the translation candidates generated by the RBMT engine during the translation process.

2.5 Language model component

The language model (LM) component is used to score the translation candidates generated by the

engine. The RBMT engine can generate several translation candidates depending on the number of homonymic words and phrases and transfer rules variants. Statistical phrase-based post-editing is applied separately to each of the generated candidates. All of the candidates (with and without post-edition) are scored by the LM component and the candidate with the lowest perplexity one is selected.

3 Experimental setting

We used the total Europarliament (EP) and NewsCommentary (NC) corpora provided by the organizers for the English-Spanish submission. We translated both (EP and NC) corpora using the RBMT engine and then built a single phrase-table for both corpora. Then we filtered the phrase-table according to the source phrase length and translation probabilities as described in section 2.4. Only 10% of the initial phrase-table were used as statistical post-editing data. The target 5-gram language model was trained on all provided monolingual data except the LDC corpora. We did not extract the dictionary for this language pair.

As for the English-French submission, we performed bilingual training data selection from EP and United Nations (UN) corpora. We trained the source and target language models on English and French monolingual News corpora respectively. These models were used to score each sentence pair of EP and UN corpora. Then we selected sentence pairs from EP and UN corpora via the geometric mean of perplexities of the source and target sentences. About 85% of EP (35M words of the source corpus) and 35% of UN (68M words of the source corpus) were selected. Then we translated the selected EP and UN subcorpora and the whole NC corpus with the RBMT engine. A single phrase-table was built for all three corpora. The phrase-table was fitered with the same parameters as for the English-Spanish submission. Approximately 8% of the initial phrase-table were used as statistical post-editing data. The target 5-gram language model was trained on all provided monolingual data except the LDC corpora.

We also performed automated dictionary extraction for the English-French pair. Examples of the extracted entries can be found in Table 1. The details about the extracted dictionary can be found in Table 2. We only extracted verbs, nouns and noun phrases for this shared task. The translations for extracted verbs and nouns are automatically added into the existing PROMT dictionary entries using our multifunctional dictionary component. Thus we increase the number of lexical variants and generated translation candidates. The extracted noun phrases are added to the PROMT dictionary as new entries. We only extract 'informative' entries, i.e. the noun phrases which are absent in the baseline PROMT dictionary or have an incorrect or infrequent translation. It should also be mentioned that the initial size of the noun phrases glossary was over 25K entries, but we decided to raise the source phrase frequency threshold a bit. Our hypothesis was that non-frequent phrases from out-of-domain corpora (EP and UN) would not fit for translation of news texts. 20K entries are selected.

KEY	KEY_FRQ	TRANSLATION	PROB	POS
comprehensive peace agreement	2427	accord de paix global	0,803049	n
automaker	7	constructeur automobile	0,428571	n
contemplate	452	envisager	0,400443	v

Table 1. Examples of extracted dictionary entries.

Part of speech	nouns	noun phrases	verbs
Number of entries	1187	19780	215

Table 2. Number of entries in the extracted English-French dictionary.

4 Experimental results and linguistic evaluation

In this section we present the results of our experiments on *newstest2012*. BLEU scores for different system configurations are presented in Table 3. The percentage of sentences changed by statistical post-editing compared to baseline RBMT output is presented in Table 4. We also

provide details of linguistic evaluation performed for the English-French submission.

System configuration	BLEU (English-French)	BLEU (English-Spanish)
RBMT (baseline)	24.00	27.26
Hybrid (+LM)	24.09	27.26
Hybrid (+LM +dictionary)	24.25	-
Hybrid (+LM +SPE)	-	28.60
Hybrid (+LM +dictionary +SPE)	24.80	-

Table 3. Translation results in terms of BLEU score for *newstest2012*.

Language pair	Impact
English-French	43%
English-Spanish	48%

Table 4. Impact of statistical post-editing on *newstest2012* (percentage of sentences changed by statistical post-editing).

Language pair	Improv	Degrad	Equiv
English-French	54	16	30
English-Spanish	48	20	32

Table 5. Statistics on improvements, degradations and equivalents for the DeepHybrid translation compared to baseline RBMT output (*newstest2012*).

Our linguists compared 100 random RBMT and DeepHybrid (with extracted dictionary and statistical post-editing) translations for both language pairs in terms of improvements and degradations. The results presented in Table 5 show that the DeepHybrid engine outperforms the RBMT engine according to human evaluation. Most of the degradations are minor grammatical issues (wrong number, disagreement etc.).

5 Conclusions and future work

We presented the PROMT DeepHybrid system submissions for WMT12 shared translation task. We showed improvements both in terms of BLEU scores and human evaluation compared to baseline PROMT RBMT engine.

We extracted a dictionary from a corpus of over 200M words. The size of the dictionary (~20K entries) is relatively small due to our robust linguistic and statistical data filtering. However, such filtering minimizes the number of possible mistranslations and guarantees that the extracted entries are universal. We are planning to add the extracted data to our baseline English-French dictionary after manual check and perform the same experiments for other language pairs.

As for statistical post-editing, the impact on the RBMT output is quite moderate (less than 50%). This is also due to our approach which includes filtering out infrequent phrase pairs from statistical post-editing data. We assume that the RBMT output is already good enough and therefore does not require much statistical post-editing to be applied. It should be mentioned that for the present we only use perplexity to score translation candidates. Several other features will be implemented in the next version of the hybrid engine. To avoid grammatical inconsistency in the hybrid MT output, we are planning to apply linguistic filters to statistical post-editing data.

References

L. Dugast, J. Senellart, and P. Koehn. 2007. *Statistical Post-Edition on SYSTRAN Rule-Based Translation System.* In Proceedings of the Second Workshop On Statistical Machine Translation, Prague, Czech Republic.

Koehn Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. *Moses: Open source toolkit for statistical machine translation.* ACL 2007, demonstration session. Prague, Czech Republic.

Och, Franz Josef and Hermann Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics, Vol. 29(1). 19-51.

F. Sadat, H. Johnson, A. Agbago, G. Foster, R. Kuhn, J. Martin, and A. Tikuisis. 2005. *PORTAGE: A Phrase-Based Machine Translation System.* In Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 129-132, Ann Arbor, USA.

M. Simard, C. Goutte, and P. Isabelle. 2007. *Statistical Phrase-Based Post-Editing.* In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 08.515, Rochester, USA.