PROMT Translation Systems for WMT 2016 Translation Tasks

Alexander Molchanov, Fedor Bykov
PROMT LLC
17E Uralskaya str. building 3, 199155,
St. Petersburg, Russia
firstname.lastname@promt.ru

Abstract

This paper provides an overview of the PROMT submissions for the WMT16 Shared Translation Tasks. We participated in seven language pairs with three different system configurations (rule-based, statistical and hybrid). We describe the architecture of the three configurations. We show that fast and accurate customization of the rule-based system can increase the BLEU scores significantly.

1 Introduction

This paper presents the PROMT systems submitted for the Shared Translation Task of WMT16. We participated in seven language pairs with three different types of systems: English-Russian, Russian-English, English-German (Rule-based systems); Finnish-English, Turkish-English (Statistical systems); English-Spanish, English-Portuguese (Hybrid systems). The paper is organized as follows. In Section 1, we briefly outline the three types of our systems and their features. In Section 2, we describe the experimental setups and the training data and present the results. Finally, Section 3 concludes the paper.

2 Systems Overview

2.1 RBMT System

The PROMT rule-based machine translation (RBMT) System is a mature machine translation system with huge linguistic structured databases containing morphological, lexical and syntactic features for the English, German, French, Spanish, Italian, Portuguese and Russian languages.

2.2 SMT System

Basic components

The PROMT SMT system is based on the Moses open-source toolkit (Koehn et al., 2007). We use MGIZA (Gao and Vogel, 2008) to generate word alignments. We build the phrase tables and lexical reordering tables using the Moses toolkit. The IRSTLM toolkit (Federico et al., 2008) is used to build language models, which are scored using KenLM (Heafield, 2011) in the decoding process. We use ZMERT (Zaidan, 2009) for weights optimization. We use a complex recaser combining a Moses-based recasing model and in-house rule-based algorithms based on source text information and word alignments.

Text preprocessing

We have a standard procedure for preprocessing and filtering parallel data, which includes removing too long sentences, discarding sentence pairs with significant length ratios etc. Text data is tokenized with in-house tokenizers and lowercased before generating word alignments.

Processing Named Entities

The in-house Named Entities (NEs) Recognition module allows to extract and process multiple types of entities including personal and company names, phone numbers, e-mails, dates etc. The numeric elements of NEs are replaced with placeholders in training data. We use XML markup for NEs and preserve the original values for numeric elements during decoding.

2.3 Hybrid System

The PROMT Hybrid system is based on three components: the RBMT module, the RBMT post-processor and the statistical post-editing (SPE) module. Text translation is performed as follows. First, the RBMT module translates the source text

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 339–343,
Berlin, Germany, August 11-12, 2016.

and outputs a complex structure containing the translation and its linguistic features (morphological and syntactic information, extracted named entities etc.). Second, the RBMT postprocessor generates XML based on the output of the RBMT module. Finally, the XML is fed to the SPE module which generates the output translation. The SPE module is basically a SMT system built on a parallel corpus of RBMT translations and their human references as described in (Simard et al., 2007). The SPE technique allows us to 1) handle systematic RBMT errors which are hard to deal with algorithmically; 2) fast and effectively adapt a translation system to a specific domain.

3 Experimental settings and results

In this section, we describe the experimental settings and report the results.

3.1 RBMT System

In this Section, we describe the RBMT submissions for English-Russian-English (News Task) and for English-German (News and IT Tasks).

Data

We used the News Commentary v11 and the Wiki Headlines parallel corpora to tune the system for the News Task. The batch 1 and batch2 sets from WMT16 training data were used for the English-German system for the IT Task.

RBMT system tuning

We have a semi-supervised technique for tuning the RBMT system. The technique is based on using the PROMT parsers. We use the following pipeline. We extract and build frequency lists of various types of NEs, out-of-vocabulary words (OOVs) and syntactic constructions. We analyze the most frequent units using human linguistic expertise. We modify the system by adding, removing or changing the values for the linguistic features of the system database elements. As a result, we obtain a system tuned for a specific text domain.

Results

Table 1 shows the BLEU scores (Papineni et al., 2002) for the baseline and the tuned RBMT systems for different language pairs measured on the newstest2016 test set and the batch3 test set for the IT Task. The huge difference between the baseline and the tuned configuration for the English-German system (IT Task) is explained mostly by the fact that we use specific in-domain databases for IT.

Language pair	Baseline	Tuned
en-ru	19,9	22,6
ru-en	20,29	21,21
en-de (News)	19,57	22,62
en-de (IT)	30,62	40,3

Table 1: Results for the RBMT submissions.

3.2 SMT System

In this Section, we describe the SMT submissions for Turkish-English (News Task) and Finish-English (News Task).

Turkish-English

Data We used all Opus (Tiedemann, 2012) data and company private parallel data (which consists mostly of crawled and aligned texts from different news web-sites). The Subtitles were preprocessed as follows: 1) we built a list of unique source sentences with all corresponding target sentences, for each source sentence we selected the most frequent target sentence (this helped us to get rid of most noisy data); 2) the selected data was filtered using in-house language recognition tool; 3) target data was filtered using a language model built on 2014, 2015 news texts corpora from statmt.org. Table 2 shows the statistics regarding the parallel training data (note that statistics for OPUS do not include Subtitles as they are presented separately).

Corpus	#word S (M)	#word T (M)
Opus	47,2	36,9
Subtitles	291,9	274,8
Private data	2,9	2,7
Overall	342	314,4

Table 2: Parallel data statistics for the Turkish-English system for the source (S) and the target (T) sides. #words is in millions (M).

A 3-gram language model was built on 2014, 2015 news texts corpora from statmt.org. We used randomly selected sentence pairs from Tatoeba and TED corpora (4000 sentence pairs) and the whole newsdev2016 development set for tuning.

Morphological preprocessing Turkish is a highly agglutinative language with complex morphology. A common technique to reduce data sparseness and produce better word alignments is morphological segmentation of the Turkish side of parallel data (Bisazza and Federico, 2009). We apply this technique to our training data using the Nuve[1] morphological analyzer . We split off 32 types of affixes (one of them is removed from source text as it is not expected to have English counterparts). The source vocabulary size reduced substantially (2.3 to 1.8 million units). We do not yet perform the disambiguation, so we split words in every case when we have an analysis variant which contains affixes described in our segmentation rules.

OOVs We use the Nuve built-in stemmer to process OOVs. The technique is quite simple. The SMT model uses two phrase-tables: the primary table and the back-off table used to translate OOVs. The back-off table consists of the primary table vocabulary stems with several translations selected by a certain direct probability threshold. An OOV is stemmed and retranslated during decoding.

Finnish-English

Data The 2016 system is based on the existing PROMT 2015 system. The 2015 system uses OPUS data (except IT documentation corpora and Subtitles) and company private parallel data (which consists mostly of crawled and aligned texts from different news web-sites). We added the Subtitles corpus to the training data for the 2016 system. The subtitles were preprocessed in the same way as for the Turkish-English system except that we used a higher threshold when filtering the texts with the news language model. Table 3 shows the statistics regarding the parallel training data.

Corpus	#word S (M)	#word T (M)
Opus	274,1	192
Subtitles	100,2	95,6
Private data	2,8	3,3
Overall	377,1	290,9

Table 3: Parallel data statistics for the Finnish-English system for the source (S) and the target (T) sides. #words is in millions (M).

[1]https://github.com/hrzafer/nuve

We used the language model built for the Turkish-English system. The newsdev2015 and newstest2015 sets were used for tuning.

OOVs We use the NLTK (Loper and Bird, 2002) implementation of the Snowball stemming algorithm (Porter, 1980) and the in-house splitter for compound words based on the algorithm described in (Koehn and Knight, 2003). The procedure for processing OOVs is pretty much the same as for the Turkish-English system, but with the additional step of splitting compound words which are not present in the back-off phrase-table.

Results

The BLEU scores for the Finnish-English and the Turkish-English experiments are reported in Tables 4 and 5 respectively.

System	BLEU
2015 system	19,88
2016 system	21,05
2016 system+UNK	21,21

Table 4: Results for the Finnish-English SMT submissions. UNK stands for using the unknown words processing technique.

System	BLEU
baseline	14,69
baseline+morph. segmentation	14,77
baseline+morph. segmentation+UNK	14,85

Table 5: Results for the Turkish-English SMT submissions.

We did not perform the significance tests for the scores difference between system configurations. However, the difference between the Turkish-English models with and without morphological segmentation seems to be insignificant. This may be due to the absence of a disambiguation algorithm (our splitting technique may be improving and worsening the translation at the same time). We will see to that in future.

3.3 Hybrid System

In this Section, we describe the Hybryd submissions for English-Spanish (IT Task) and English-Portuguese (IT Task).

Data

We built two systems for each language pair: the baseline (built only on WMT16 IT Task data) and

the improved system (WMT16 IT Task data with in-house IT documentation data). For the baseline system, we used the data as is. The target side of the private data for the improved system was filtered using a language model built on batch1 and batch2 development sets. 5% and 6.5% of the data were discarded for the English-Spanish and the English-Portuguese systems respectively. The discarded data is mostly some junk with residual html formatting. We also normalized the target data for English-Portuguese by converting the orthography for 50 most common words from Brazilian to Portuguese language variety. The filtered private data used for training amounts to 51,4 million tokens for Spanish and 29,7 million tokens for Portuguese. The language models for the systems were built on all target data. The batch1 and batch2 development sets were used for tuning the SPE module.

Results

The BLEU scores for both experiments are reported in Table 6. It is worthy to mention the substantial difference between the English-Spanish and English-Portuguese results when comparing the baseline and improved hybrid systems. The difference between the training data size is not drastically significant whereas the difference in BLEU scores is. This may be due to the quality of our Portuguese data. We will examine this question in future.

Language pair	System		
	rbmt	**hybrid (baseline)**	**hybrid (improved)**
en-sp	32,0	37,6	42,7
en-pt	27,2	32,0	32,7

Table 6: Results for the hybrid submissions.

4 Conclusions and future work

We have described the different approaches that we used for our participation in the WMT16 Shared Translation Task. Using different approaches to machine translation allows us to perform competitively in all language pairs. We describe the fast semi-supervised RBMT system customization technique which is effective in terms of BLEU. We plan to research the disambiguation impact on our morphological segmentation technique for Turkish and a more careful way of handling OOVs for our SMT systems.

References

Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for turkish to english statistical machine translation. In *Proceedings of IWSLT 2009*, Tokyo, Japan.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech*, Brisbane, Australia.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing, SETQA-NLP 08*, pages 49 – 57, Stroudsburg, PA, USA.

Kenneth Heafield. 2011. Kenlm : Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, number 2009, pages 187 – 197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary.

Philipp Koehn, 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. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL 07*, pages 177 – 180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics (ACL 02)*, volume 1, pages 63 – 70, Philadelphia, PA, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pages 311 – 318, Philadelphia, PA, July.

Martin Porter. 1980. An algorithm for suffix stripping. In *Program: electronic library and information systems*, number 14(3), pages 130 – 137.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proceedings of The North American Chapter of the*

Association for Computational Linguistics Conference (NAACL-07), pages 508 – 515, Rochester, NY, April.

Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Omar F. Zaidan. 2009. Z-mert: A fully configurable open source tool for minimum error rate training of machine translation systems. In *The Prague Bulletin of Mathematical Linguistics*, number 91, pages 79 – 88.