PJAIT Systems for the WMT 2016

Krzysztof Wołk
Multimedia Department
Polish-Japanese Academy of Information Technology,
Koszykowa 86,
02-008 Warsaw
kwolk@pja.edu.pl

Krzysztof Marasek
Multimedia Department
Polish-Japanese Academy of Information Technology,
Koszykowa 86,
02-008 Warsaw
kmarasek@pja.edu.pl

Abstract

In this paper, we attempt to improve Statistical Machine Translation (SMT) systems between Czech and English. To accomplish this, we performed translation model training, created adaptations of training settings for each language pair, and obtained comparable corpora for our SMT systems. Innovative tools and data adaptation techniques were employed. Only the official parallel text corpora and monolingual models for the WMT 2016 evaluation campaign were used to train language models, and to develop, tune, and test the system. We explored the use of domain adaptation techniques, symmetrized word alignment models, the unsupervised transliteration models and the KenLM language modeling tool. To evaluate the effects of different preparations on translation results, we conducted experiments and used the BLEU, NIST and TER metrics. Our results indicate that our approach produced a positive impact on SMT quality.

1 Introduction

Statistical Machine Translation (SMT) must deal with a number of problems to achieve high quality. These problems include the need to align parallel texts in language pairs and cleaning harvested parallel corpora to remove errors. This is especially true for real-world corpora developed from text harvested from the vast data available on the Internet. Out-Of-Vocabulary (OOV) words must also be handled, as they are inevitable in real-world texts (Wolk and Marasek, 2014a).

The lack of enough parallel corpora is another significant challenge for SMT. Since the approach is statistical in nature, a significant amount of quality language pair data is needed to improve translation accuracy. In addition, very general translation systems that work in a general text domain have accuracy problems in specific domains. SMT systems are more accurate on corpora from a domain that is not too wide. This exacerbates the data problem, calling for the enhancement of parallel corpora for particular text domains (Wolk and Marasek, 2014b).

This paper describes SMT research that addresses these problems, particularly domain adaptation within the limits of permissible data for the WMT 2016 campaign. To accomplish this, we performed model training, created adaptations of training settings and data for each language pair.

Innovative tools and data adaptation techniques were employed. We explored the use of domain adaptation techniques, symmetrized word alignment models, the unsupervised transliteration models, and the KenLM language modeling tool (Heafield, 2011). To evaluate the effects of different preparations on translation results, we conducted experiments and evaluated the results using standard SMT metrics (Koehn et al., 2007).

The languages translated during this research were: Czech, and English, in both directions. Czech is found in the Slavic branch of that language family. English falls in the Western group (The Technology Development Group, 2013-2014)

This paper is structured as follows: Section 2 explains the data preparation. Section 3 presents

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 411–414,
Berlin, Germany, August 11-12, 2016.

experiment setup and the results. Lastly in Section 4 we summarize the work.

2 Data Preparation

This section describes our techniques for data preparation for our SMT systems. We give particular emphasis to preparation of the language data and models and our domain adaptation approach.

2.1 Data pre-processing

Two languages were involved in this research: Czech and English. The text was encoded in UTF-8 format, separated into sentences, and provided in pairs of languages.

Pre-processing, both automatic and manual, of this training data was required. There were a variety of errors found in this data, including spelling errors, unusual nesting of text, text duplication, and parallel text issues. Approximately 2% of the text in the training set contained spelling errors, and approximately 4% of the text had insertion errors. A tool described in (Wolk and Marasek, 2014b) was used to correct these errors. Previous studies have found that such cleaning increases the BLEU score for SMT by a factor of 1.5–2 (Wolk and Marasek, 2014a).

SyMGiza++, a tool that supports the creation of symmetric word alignment models, was used to extract parallel phrases from the data. This tool enables alignment models that support many-to-one and one-to-many alignments in both directions between two language pairs. SyMGiza++ is also designed to leverage the power of multiple processors through advanced threading management, making it very fast. Its alignment process uses four different models during training to progressively refine alignment results. This approach has yielded impressive results in Junczys-Dowmunt and Szał (2012).

Out-Of-Vocabulary (OOV) words pose another significant challenge to SMT systems. If not addressed, unknown words appear, untranslated, in the output, lowering the translation quality. To address OOV words, we used implemented in the Moses toolkit Unsupervised Transliteration Model (UTM). UTM is an unsupervised, language-independent approach for learning OOV words (Moses statistical machine translation, 2015). We used the post-decoding transliteration option with this tool. UTM uses a transliteration phrase translation table to evaluate and score multiple possible transliterations (Durrani et al., 2014).

The KenLM tool was applied to the language model to train and binarize it. This library enables highly efficient queries to language models, saving both memory and computation time. The lexical values of phrases are used to condition the reordering probabilities of phrases. We used KenLM with lexical reordering set to hier-msd-bidirectional-fe. This setting uses a hierarchical model that considers three orientation types based on both source and target phrases: monotone (M), swap (S), and discontinuous (D). Probabilities of possible phrase orders are examined by the bidirectional reordering model (Costa-Jussa and Fonollosa, 2010; Moses statistical machine translation, 2013).

2.2 Domain Adaptation

The news data sets have a rather a wide domain, but rather not as wide-ranging in topic as the variety of WMT permissible texts. Since SMT systems work best in a defined domain, this presents another considerable challenge. If not addressed, this would lead to lower translation accuracy.

The quality of domain adaptation depends heavily on training data used to optimize the language and translation models in an SMT system. Selection and extraction of domain-specific training data from a large, general corpus addresses this issue (Axelrod, He and Gao, 2011). This process uses a parallel, general domain corpus and a general domain monolingual corpus in the target language. The result is a pseudo in-domain sub-corpus.

As described by Wang et al. in (2014), there are generally three processing stages in data selection for domain adaptation. First, sentence pairs from the parallel, general domain corpus are scored for relevance to the target domain. Second, resampling is performed to select the best-scoring sentence pairs to retain in the pseudo in-domain sub-corpus. Those two steps can also be applied to the general domain monolingual corpus to select sentences for use in a language model. After collecting a substantial amount of sentence pairs (for the translation model) or sentences (for the language model), those models are trained on the sub-corpus that represents the target domain (Wang et al., 2014).

Similarity measurement is required to select sentences for the pseudo in-domain sub-corpus. There are three state-of-the-art approaches for similarity measurement. The cosine tf-idf criterion looks for word overlap in determining similarity. This technique is specifically helpful in reducing the number of OOV words, but it is

sensitive to noise in the data. A perplexity-based criterion considers the n-gram word order in addition to collocation. Lastly, edit distance simultaneously considers word order, position, and overlap. It is the strictest of the three approaches. In their study (Wang et al., 2014), Wang et al. found that a combination of these approaches provided the best performance in domain adaptation for Chinese-English corpora (Wang et al., 2014)

In accordance with Wang et al. (2014)'s approach, we use a combination of the criteria at both the corpora and language models. The three similarity metrics are used to select different pseudo in-domain sub-corpora. The sub-corpora are then joined during resampling based on a combination of the three metrics. Similarly, the three metrics are combined for domain adaptation during translation. We empirically found acceptance rates that allowed us only to harvest 20% of most domain-similar data (Wang et al., 2014).

3 Experimental Results

Various versions of our SMT systems were evaluated via experimentation. In preparation for experiments, we processed the corpora. This involved tokenization, cleaning, factorization, conversion to lower case, splitting, and final cleaning after splitting. Language models were developed and tuned using the training data.

The Experiment Management System (Koehn et al., 2007) from the open source Moses SMT toolkit was used to conduct the experiments. Training of a 6-gram language model was accomplished our resulting systems using the KenLM Modeling Toolkit instead of 5-gram SRILM (Stolcke, 2002) with an interpolated version of Kneser-Key discounting (interpolate –unk –kndiscount) that was used in our baseline systems. Word and phrase alignment was performed using SyMGIZA++ (Junczys-Dowmunt and Szał, 2012) instead of GIZA++. KenLM was also used, as described earlier, to binarize the language models. The OOV's were handled by using Unsupervised Transliteration Model (Durrani, 2014).

The results are shown in Table 1. Each language pair was translated in both directions. "BASE" in the tables represents the baseline SMT system. "EXT" indicates results for the baseline system, using the baseline settings but extended with additional permissible data (limited to parallel Europarl v7, Common Crawl, News Commentary, CzEng and monolingual News Crawl 07-15) with data adaptation. "BEST" indicates the results when the new SMT settings were applied and using all permissible data after data adaptation.

Three well-known metrics were used for scoring the results: Bilingual Evaluation Understudy (BLEU), the US National Institute of Standards and Technology (NIST) metric and Translation Error Rate (TER).

The results show that the systems performed well on all data sets in comparison to the baseline SMT systems. Application of the new settings and use of all permissible data improved performance even more.

LANG	SYSTEM	BLEU	NIST	TER
CS-EN	BASE	25.99	5.51	64.35
	EXT	27.92	6.04	62.58
	BEST	29.31	6.97	60.45
EN-CS	BASE	22.20	5.36	67.60
	EXT	24.62	5.57	64.25
	BEST	26.14	5.74	62.02

Table 1: Progressive Results, 2014 Test Data

4 Summary

We have improved SMT for CS-EN in 2 directions in News Translation task, using only data permissible for the WMT 2016 evaluation campaign. We cleaned, prepared, and tokenized the training data. Symmetric word alignment models were used to align the corpora. UTM was used to handle OOV words. A language model was created, binarized, and tuned. We performed domain adaptation of language data using a combination of similarity metrics.

The results show a positive impact of our approach on SMT quality across the choose language pair.

Reference

Amittai Axelrod, Xiaodong He and Jianfeng Gao.. 2011. Domain adaptation via pseudo in-domain data selection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11), p. 355–362

Marta R. Costa-Jussa and Jose R. Fonollosa. 2010. Using linear interpolation and weighted reordering hypotheses in the Moses system, Barcelona, Spain

Nadir Durrani, et al. 2014. Integrating an unsupervised transliteration model into statistical machine translation. In: EACL 2014, p. 148.

Marcin Junczys-Dowmunt and Arkadiusz Szał. 2012. SyMGiza++: symmetrized word alignment models for statistical machine translation. In: Security and Intelligent Information Systems. Springer Berlin Heidelberg, p. 379-390.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries, In: Proc. of Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics

Philipp Koehn et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, pp. 177–180

Moses statistical machine translation, "OOVs." Last revised February 13, 2015. Retrieved September 27, 2015 from: http://www.statmt.org/moses/?n=Advanced.OOVs#ntoc2

Moses statistical machine translation, "Build reordering model." Last revised July 28, 2013. Retrieved October 10, 2015 from: http://www.statmt.org/moses/?n=FactoredTraining.BuildReorderingModel

The Technology Development Group. 2014. Czech. Last revised March 21, 2014. Retrieved September 27, 2015 from: aboutworldlanguages.com/Czech

The Technology Development Group. 2014. English. Last revised October 14, 2013. Retrieved Septeber 27, 2015 from: aboutworldlanguages.com/english

The Technology Development Group. 2014. French. Last revised March 21, 2014. Retrieved September 27, 2015 from: aboutworldlanguages.com/french

The Technology Development Group. 2014. German. Last revised March 21, 2014. Retrieved September 27, 2015 from: aboutworldlanguages.com/german

The Technology Development Group. 2014. Vietnamese. Last revised March 21, 2014. Retrieved September 27, 2015 from: aboutworldlanguages.com/vietnamese

Andreas Stolcke. 2002. SRILM - An Extensible Laguage Modeling Toolkit., INTERSPEECH, 2002.

Longyue Wang, Derek F. Wong, Lidia S. Chao, Yi Lu, and Junwen Xing. 2014. A Systematic Coparison of Data Selection Criteria for SMT Domain Adaptation., The Scientific World Journal, vol. 2014, doi:10.1155/2014/745485

Krzysztof Wołk, Krzysztof Marasek. 2014a. Polish - English Speech Statistical Machine Translation Systems for the IWSLT 2014, In: Proceedings of International Workshop on Spoken Language Translation, Lake Tahoe, California, USA, pp. 143-148.

Krzysztof Wołk, Krzysztof Marasek. 2014b. A Sentence Meaning Based Alignment Method for Parallel Text Corpora Preparation. In: New Perspectives in Information Systems and Technologies, Volume 1. Springer International Publishing, 2014. p. 229-237.

Krzysztof Wołk and Krzysztof Marasek. 2015. Tuned and GPU-accelerated Parallel Data Mining from Comparable Corpora, In: Lecture Notes in Artificial Intelligence, p. 32 – 40