PJIIT’s systems for WMT 2017 Conference

Krzysztof Wołk

Multimedia Department

Polish-Japanese Academy of Information
Technology, Koszykowa 86,
kwolk@pja.edu.pl

Krzysztof Marasek

Multimedia Department

Polish-Japanese Academy of Information
Technology, Koszykowa 86,
kmarasek@pja.edu.pl

Abstract

In this paper, we attempt to improve Statistical Machine Translation (SMT) systems between Czech, Latvian and English in WNT’17 News translation task. We also participated in the Biomedical task and produces translation engines from English into Polish, Czech, German, Spanish, French, Hungarian, Romanian and Swedish. To accomplish this, we performed translation model training, created adaptations of training settings for each language pair, and implemented BPE (subword units) for our SMT systems. Innovative tools and data adaptation techniques were employed. Only the official parallel text corpora and monolingual models for the WMT 2017 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 for some less popular languages 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 2017 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, Latvian and English in WNT’17 News translation task. We also participated in the Biomedical task and produces translation engines from English into Polish, Czech, German, Spanish, French, Hungarian, Romanian and Swedish. This paper is structured as follows: Section 2 explains the data preparation. Section 3 presents experimental

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 416–421
Copenhagen, Denmark, September 711, 2017. ©2017 Association for Computational Linguistics

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 in-domain data adaptation approach.

2.1 Data pre-processing

The texts were 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. For example in Polish-English corpora approximately 3% of the text in the training set contained spelling errors, and approximately 2% of the text had insertion errors. A tool described in (Wolk and Marasek, 2014b) was used to correct these errors automatically. 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-msdbidirectional-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. The same goes to the biomedical task. 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 indomain 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.

For Cosine tf-idf every document D_i is represented as a vector $(w_{i1}, w_{i2},\ldots,w_{in})$ and n is the size of the vocabulary. So W$_{ij}$ is calculated as follows:

$$W_{ij} = tf_{ij} \times log\ (idf_j)$$

In which $tf_{ij}i$ is the term frequency (TF) of the j-th word in the vocabulary in the document D$_i$ and idf$_j$ is the inverse document frequency (IDF) of the j-th word calculated. The likeness between the two texts is later explained as the cosine of the angle between two vectors. This formula is applied in accordance to Lü et al. (2007) and Hildebrand et al. (2005). This approach supposes that M is the size of query set and N is the number of sentences put together from general corpus according to each and every query. Thus, the size of the cosine tf-idf based quasi in-domain sub corpus is defined as:

$$Size_{Cos-IR} = M \times N$$

Perplexity is focused on the cross-entropy (Koehn 2004) that is the average of the negative logarithm of the word probabilities. Consider

$$H(p,q) = -\sum_{i=1}^{n} p(w_i) \log q(w_i) \\ = -\frac{1}{N}\sum_{i=1}^{n} \log q(w_i)$$

where p symbolizes the empirical distribution of the sample of the test. If w_i appeared n times in the test sample of N size, then $q(w_i)$ is the probability of the w_i event approximated from the training set.

For that, perplexity (pp) can be performed simply at the base point that is presented in the system, and is often applied as a cosmetic alternative of perplexity for the data selection as:

$$pp = b^{H(p,q)}$$

where b is the based of measured cross-entropy, $H(p,q)$ is the cross-entropy as given in (Koehn 2004) (often used as substitute of the perplexity in data selection Axelrod et al. 2011; Moore and Lewis 2010).

Let $H_I(p,q)$ and $H_O(p,q)$ be the cross-entropy of $\mathrm{w_i}$ string in accordance with the language model, which is subsequently, trained by general-domain dataset and in-domain dataset. While looking at the target (tgt) dimensions and the sources (src) of training data, there are three perplexity-based variants. The first one is known as basic cross-entropy defined as:

$$H_{I-src}(p,q)$$

The second is Moore-Lewis cross-entropy difference (Moore and Lewis 2010):

$$H_{I-src}(p,q) - H_{G-src}(p,q),$$

that attempts to choose the sentences that are more identical to I one and other but different to others in G. Both the standards mentioned above, consider only the sentences in the source language. Moreover, Axelrod et al. (2011) proposed a metric that adds cross-entropy differences over both sides:

$$[H_{I-src}(p,q) - H_{G-src}(p,q)] \\ + [H_{I-tgt}(p,q) \\ - H_{G-tgt}(p,q)]$$

For instance, candidates with lower scores (Daumé III and Jagarlamudi 2011; Papineni et al. 2002; Mansour and Ney 2012) have higher relevancy to target specific domain. The size of the perplexity-based quasi in-domain subset must be equal to one another. In practice, we work with SRILM toolkit to train 5-gram LMs with interpolated modified Kneser-Ney discounting (Stolcke 2002; Chen and Goodman 1996).

In the realm of information theory and computer science, the Levenshtein distance is regarded as a string metric for the measurement of dissimilarity between two sequences. In casual terms, the Levenshtein distance between points or words is the minimum possible number of unique edits like the insertions or deletions in the data that is required to replace one word with another one.

Levenshtein distance also refers to the edit distance, only wider in its approach as it incorporates a wider area of subjects the distance metrics. It has a close association with pairwise string arrangement as well.

Mathematically, the Levenshtein distance between two strings a, b (of length $|a|$ and $|b|$ respectively) is given by $lev_{a,b}(|a|,|b|)$ where

$$lev_{a,b}(i,j) \\ = \begin{cases} \max(i,j) & if \min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise. \end{cases}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $lev_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b.

It is to be noted that the first component that is in the least correspondence of the deletion (from a to b), the second of the insertion and the third to match or mismatch, varying on whether the respective symbols are the matching.

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)

2.2 Sub-word units

Neural machine translation (NMT) models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem. In SMT vocabularies that are disproportional are similar problem. Authors (Sutskever, Vinyals and Le, 2014) introduced a simple and effective approach, making the MT models capable of handling such problems by encoding rare and unknown words as sequences of subword units. This is based on the intuition that various word classes are translatable via smaller units than words, for instance names (via character copying or transliteration), compounds (via compositional translation), and cognates and loanwords (via phonological and morphological transformations). We applied this technique to our SMT engines for Slavic languages and obtained improved results of about 1.2 points in BLEU score.

3 Experimental setup

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 lowercase, splitting, and final cleaning after splitting. Language models were developed and tuned using only the constrained 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 in 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. "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 permissible data) 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.

Task	Language and Direction	System	BLEU
News	CS->EN	BASE	21.18
News	CS->EN	EXT	22.67
News	CS->EN	BEST	23.9
News	EN->CS	BASE	14.04
News	EN->CS	EXT	15.44
News	EN->CS	BEST	16.6
News	LV->EN	BASE	10.09
News	LV->EN	EXT	12.17

News	LV->EN	BEST	12.9
News	EN->LV	BASE	8.78
News	EN->LV	EXT	9.78
News	EN->LV	BEST	10.4
Biomedical	EN->PL	BASE	12.45
Biomedical	EN->PL	EXT	18.62
Biomedical	EN->PL	BEST	18.86
Biomedical	EN->PL	BEST + BPE	18.88
Biomedical	EN->CS	BASE	14.56
Biomedical	EN->CS	EXT	18.12
Biomedical	EN->CS	BEST	19.96
Biomedical	EN->DE	BASE	21.43
Biomedical	EN->DE	EXT	24.64
Biomedical	EN->DE	BEST	25.13
Biomedical	EN->RO	BASE	19.43
Biomedical	EN->RO	EXT	23.18
Biomedical	EN->RO	BEST	24.91

Table 1: News and Biomedical Task Translation Results

4 Summary

We have improved our SMT systems for the WMT 2017 evaluation campaign using only permissible data. 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. We also successfully used BPE inside SMT for morphologically rich language (Polish). This brings promise of improvement for other slavic languages as well.

References

Amittai Axelrod, Xiaodong He, Jianfeng Gao. 2011. *Domain adaptation via pseudo in-domain data selection*. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 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

Hal Daumé III, Jagadesh Jagarlamudi. 2011. *Domain adaptation for machine translation by mining unseen words.* In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 407–412

Stanley Chen, Joshua Goodman. 1996. *An empirical study of smoothing techniques for language modeling.* In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL '96). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 310–318. doi: 10.3115/981863.981904

Nadir Durrani, et al. 2014. *Integrating an Unsupervised Transliteration Model into Statistical Machine Translation*. EACL 2014. p. 148-153.

Kenneth Heafield. 2011. *KenLM: Faster and smaller language model queries*. Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2011. p. 187-197.

Almut Silja Hildebrand et al. 2005. *Adaptation of the translation model for statistical machine translation based on information retrieval.* In: Proceedings of EAMT 10th Annual Conference, 30-31 May 2005, Budapest, Hungary. Association for Computational Linguistics, Stroudsburg, PA, pp 133–142

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

Philipp Koehn. 2004. *Pharaoh: a beam search decoder for phrase-based statistical machine translation models.* In: Proceedings of the

Antenna Measurement Techniques Association (AMTA '04). Springer, Berlin, Germany, pp 115–124

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

Yajuan Lü , Jin Huang, Qun Liu. 2007. *Improving statistical machine translation performance by training data selection and optimization*. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 343–350

Saab Mansour, Hermann Ney. 2012. *A simple and effective weighted phrase extraction for machine translation adaptation.* In: Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT '12). Springer-Verlag, Berlin, Heidelberg, pp 193–200

Robert Moore, William Lewis. 2010. *Intelligent selection of language model training data.* In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 220–224

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

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

Kishore Papineni et al. 2002. *BLEU: a method for automatic evaluation of machine translation.* In: Proceedings of the Workshop on Automatic Summarization (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 311–318. doi: 10.3115/1073083.1073135

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

Ilya Sutskever, Oriol Vinyals, & Quoc Le. 2014. *Sequence to sequence learning with neural networks.* In Advances in neural information processing systems (pp. 3104-3112).

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.