QCRI-MES Submission at WMT13: Using Transliteration Mining to Improve Statistical Machine Translation

Hassan Sajjad[1], **Svetlana Smekalova**[2], **Nadir Durrani**[3], **Alexander Fraser**[4], **Helmut Schmid**[4]

[1]Qatar Computing Research Institute – hsajjad@qf.org.qa

[2]University of Stuttgart – smekalsa@ims.uni-stuttgart.de

[3]University of Edinburgh – dnadir@inf.ed.ac.uk

[4]Ludwig-Maximilians University Munich – (fraser|schmid)@cis.uni-muenchen.de

Abstract

This paper describes QCRI-MES's submission on the English-Russian dataset to the Eighth Workshop on Statistical Machine Translation. We generate improved word alignment of the training data by incorporating an unsupervised transliteration mining module to GIZA++ and build a phrase-based machine translation system. For tuning, we use a variation of PRO which provides better weights by optimizing BLEU+1 at corpus-level. We transliterate out-of-vocabulary words in a post-processing step by using a transliteration system built on the transliteration pairs extracted using an unsupervised transliteration mining system. For the Russian to English translation direction, we apply linguistically motivated pre-processing on the Russian side of the data.

1 Introduction

We describe the QCRI-Munich-Edinburgh-Stuttgart (QCRI-MES) English to Russian and Russian to English systems submitted to the Eighth Workshop on Statistical Machine Translation. We experimented using the standard Phrase-based Statistical Machine Translation System (PSMT) as implemented in the Moses toolkit (Koehn et al., 2007). The typical pipeline for translation involves word alignment using GIZA++ (Och and Ney, 2003), phrase extraction, tuning and phrase-based decoding. Our system is different from standard PSMT in three ways:

- We integrate an unsupervised transliteration mining system (Sajjad et al., 2012) into the GIZA++ word aligner (Sajjad et al., 2011). So, the selection of a word pair as a correct alignment is decided using both translation probabilities and transliteration probabilities.
- The MT system fails when translating out-of-vocabulary (OOV) words. We build a statistical transliteration system on the transliteration pairs mined by the unsupervised transliteration mining system and transliterate them in a post-processing step.
- We use a variation of Pairwise Ranking Optimization (*PRO*) for tuning. It optimizes BLEU at corpus-level and provides better feature weights that leads to an improvement in translation quality (Nakov et al., 2012).

We participate in English to Russian and Russian to English translation tasks. For the Russian/English system, we present experiments with two variations of the parallel corpus. One set of experiments are conducted using the standard parallel corpus provided by the workshop. In the second set of experiments, we morphologically reduce Russian words based on their fine-grained POS tags and map them to their root form. We do this on the Russian side of the parallel corpus, tuning set, development set and test set. This improves word alignment and learns better translation probabilities by reducing the vocabulary size.

The paper is organized as follows. Section 2 talks about unsupervised transliteration mining and its incorporation to the GIZA++ word aligner. In Section 3, we describe the transliteration system. Section 4 describes the extension of PRO that optimizes BLEU+1 at corpus level. Section 5 and Section 6 present English/Russian and Russian/English machine translation experiments respectively. Section 7 concludes.

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 219–224,
Sofia, Bulgaria, August 8-9, 2013

2 Transliteration Mining

Consider a list of word pairs that consists of either transliteration pairs or non-transliteration pairs. A non-transliteration pair is defined as a word pair where words are not transliteration of each other. They can be translation, misalignment, etc. Transliteration mining extracts transliteration pairs from the list of word pairs. Sajjad et al. (2012) presented an unsupervised transliteration mining system that trains on the list of word pairs and filters transliteration pairs from that. It models the training data as the combination of a transliteration sub-model and a non-transliteration sub-model. The transliteration model is a joint source channel model. The non-transliteration model assumes no correlation between source and target word characters, and independently generates a source and a target word using two fixed unigram character models. The transliteration mining model is defined as an interpolation of the transliteration model and the non-transliteration model.

We apply transliteration mining to the list of word pairs extracted from English/Russian parallel corpus and mine transliteration pairs. We use the mined pairs for the training of the transliteration system.

2.1 Transliteration Augmented-GIZA++

GIZA++ aligns parallel sentences at word level. It applies the IBM models (Brown et al., 1993) and the HMM model (Vogel et al., 1996) in both directions i.e. source to target and target to source. It generates a list of translation pairs with translation probabilities, which is called the t-table. Sajjad et al. (2011) used a heuristic-based transliteration mining system and integrated it into the GIZA++ word aligner. We follow a similar procedure but use the unsupervised transliteration mining system of Sajjad et al. (2012).

We define a transliteration sub-model and train it on the transliteration pairs mined by the unsupervised transliteration mining system. We integrate it into the GIZA++ word aligner. The probability of a word pair is calculated as an interpolation of the transliteration probability and the translation probability stored in the t-table of the different alignment models used by the GIZA++ aligner. This interpolation is done for all iterations of all alignment models.

2.1.1 Estimating Transliteration Probabilities

We use the algorithm for the estimation of transliteration probabilities of Sajjad et al. (2011). We modify it to improve efficiency. In step 6 of Algorithm 1 instead of taking all f that coocur with e, we take only those that have a word length ratio in range of 0.8-1.2.[1] This reduces $cooc(e)$ by more than half and speeds up step 9 of Algorithm 1. The word pairs that are filtered out from $cooc(e)$ won't have transliteration probability $p_{ti}(f|e)$. We do not interpolate in these cases and use the translation probability as it is.

Algorithm 1 Estimation of transliteration probabilities, **e**-to-**f** direction

1: unfiltered data ← list of word pairs
2: filtered data ←transliteration pairs extracted using unsupervised transliteration mining system
3: Train a transliteration system on the filtered data
4: **for all** e **do**
5: $nbestTI(e)$ ← 10 best transliterations for e according to the transliteration system
6: $cooc(e)$ ← set of all f that cooccur with e in a parallel sentence with a word length in ratio of 0.8-1.2
7: $candidateTI(e) \leftarrow cooc(e) \cup nbestTI(e)$
8: **for all** f **do**
9: $p_{moses}(f, e)$ ← joint transliteration probability of e and f according to the transliterator
10: Calculate conditional transliteration probability $p_{ti}(f|e) \leftarrow \frac{p_{moses}(f,e)}{\sum_{f' \in CandidateTI(e)} p_{moses}(f',e)}$

2.1.2 Modified EM Training

Sajjad et al. (2011) modified the EM training of the word alignment models. They combined the translation probabilities of the IBM models and the HMM model with the transliteration probabilities. Consider $p_{ta}(f|e) = f_{ta}(f, e)/f_{ta}(e)$ is the translation probability of the word alignment models. The interpolated probability is calculated by adding the smoothed alignment frequency $f_{ta}(f, e)$ to the transliteration probability weight by the factor λ. The modified translation probabilities is given by:

$$\hat{p}(f|e) = \frac{f_{ta}(f, e) + \lambda p_{ti}(f|e)}{f_{ta}(e) + \lambda} \quad (1)$$

where $f_{ta}(f, e) = p_{ta}(f|e) f_{ta}(e)$. $p_{ta}(f|e)$ is obtained from the original t-table of the alignment model. $f_{ta}(e)$ is the total corpus frequency of e. λ is the transliteration weight which is defined as the number of counts the transliteration model gets versus the translation model. The model is not

[1] We assume that the words with very different character counts are less likely to be transliterations.

very sensitive to the value of λ. We use $\lambda = 50$ for our experiments. The procedure we described of estimation of transliteration probabilities and modification of EM is also followed in the opposite direction **f**-to-**e**.

3 Transliteration System

The unsupervised transliteration mining system (as described in Section 2) outputs a list of transliteration pairs. We consider transliteration word pairs as parallel sentences by putting a space after every character of the words and train a PSMT system for transliteration. We apply the transliteration system to OOVs in a post-processing step on the output of the machine translation system.

Russian is a morphologically rich language. Different cases of a word are generally represented by adding suffixes to the root form. For OOVs that are named entities, transliterating the inflected forms generates wrong English transliterations as inflectional suffixes get transliterated too. To handle this, first we need to identify OOV named entities (as there can be other OOVs that are not named entities) and then transliterate them correctly. We tackle the first issue as follows: If an OOV word is starting with an upper case letter, we identify it as a named entity. To correctly transliterate it to English, we stem the named entity based on a list of suffixes (а, ом, ы, е, ой, у) and transliterate the stemmed form. For morphologically reduced Russian (see Section 6.1), we follow the same procedure as OOVs are unknown to the POS tagger too and are (incorrectly) not reduced to their root forms. For OOVs that are not identified as named entities, we transliterate them without any pre-processing.

4 PRO: Corpus-level BLEU

Pairwise Ranking Optimization (PRO) (Hopkins and May, 2011) is an extension of MERT (Och, 2003) that can scale to thousands of parameters. It optimizes sentence-level BLEU+1 which is an add-one smoothed version of BLEU (Lin and Och, 2004). The sentence-level BLEU+1 has a bias towards producing short translations as add-one smoothing improves precision but does not change the brevity penalty. Nakov et al. (2012) fixed this by using several heuristics on brevity penalty, reference length and grounding the precision length. In our experiments, we use the improved version of PRO as provided by Nakov et al. (2012). We call it *PROv1* later on.

5 English/Russian Experiments

5.1 Dataset

The amount of bitext used for the estimation of the translation model is $\approx$ 2M parallel sentences. We use newstest2012a for tuning and newstest2012b (tst2012) as development set.

The language model is estimated using large monolingual corpus of Russian $\approx$ 21.7M sentences. We follow the approach of Schwenk and Koehn (2008) by training domain-specific language models separately and then linearly interpolate them using SRILM with weights optimized on the held-out development set. We divide the tuning set newstest2012a into two halves and use the first half for tuning and second for test in order to obtain stable weights (Koehn and Haddow, 2012).

5.2 Baseline Settings

We word-aligned the parallel corpus using GIZA++ (Och and Ney, 2003) with 5 iterations of Model1, 4 iterations of HMM and 4 iterations of Model4, and symmetrized the alignments using the grow-diag-final-and heuristic (Koehn et al., 2003). We built a phrase-based machine translation system using the Moses toolkit. *Minimum error rate training (MERT)*, *margin infused relaxed algorithm (MIRA)* and *PRO* are used to optimize the parameters.

5.3 Main System Settings

Our main system involves a pre-processing step – unsupervised transliteration mining, and a post-processing step – transliteration of OOVs. For the training of the unsupervised transliteration mining system, we take the word alignments from our baseline settings and extract all word pairs which occur as 1-to-1 alignments (like Sajjad et al. (2011)) and later refer to them as a *list of word pairs*. The unsupervised transliteration mining system trains on the list of word pairs and mines transliteration pairs. We use the mined pairs to build a transliteration system using the Moses toolkit. The transliteration system is used in Algorithm 1 to generate transliteration probabilities of candidate word pairs and is also used in the post-processing step to transliterate OOVs.

We run GIZA++ with identical settings as described in Section 5.2. We interpolate for ev-

	GIZA++	TA-GIZA++	OOV-TI
MERT	23.41	23.51	23.60
MIRA	23.60	23.73	23.85
PRO	23.57	23.68	23.70
PROv1	23.65	23.76	23.87

Table 1: BLEU scores of English to Russian machine translation system evaluated on tst2012 using baseline GIZA++ alignment and transliteration augmented-GIZA++. OOV-TI presents the score of the system trained using TA-GIZA++ after transliterating OOVs

ery iteration of the IBM Model1 and the HMM model. We had problem in applying smoothing for Model4 and did not interpolate transliteration probabilities for Model4. The alignments are refined using the grow-diag-final-and heuristic. We build a phrase-based system on the aligned pairs and tune the parameters using *PROv1*. OOVs are transliterated in the post-processing step.

5.4 Results

Table 1 summarizes English/Russian results on tst2012. Improved word alignment gives up to 0.13 BLEU points improvement. PROv1 improves translation quality and shows 0.08 BLEU point increase in BLEU in comparison to the parameters tuned using PRO. The transliteration of OOVs consistently improve translation quality by at least 0.1 BLEU point for all systems.[2] This adds to a cumulative gain of up to 0.2 BLEU points.

We summarize results of our systems trained on GIZA++ and transliteration augmented-GIZA++ (TA-GIZA++) and tested on tst2012 and tst2013 in Table 2. Both systems use PROv1 for tuning and transliteration of OOVs in the post-processing step. The system trained on TA-GIZA++ performed better than the system trained on the baseline aligner GIZA++.

6 Russian/English Experiments

In this section, we present translation experiments in Russian to English direction. We morphologically reduce the Russian side of the parallel data in a pre-processing step and train the translation system on that. We compare its result with the Russian to English system trained on the un-processed parallel data.

[2]We see similar gain in BLEU when using operation sequence model (Durrani et al., 2011) for decoding and transliterating OOVs in a post-processing step (Durrani et al., 2013).

SYS	tst2012	tst2013
GIZA++	23.76	18.4
TA-GIZA++	23.87	18.5*

Table 2: BLEU scores of English to Russian machine translation system evaluated on tst2012 and tst2013 using baseline GIZA++ alignment and transliteration augmented-GIZA++ alignment and post-processed the output by transliterating OOVs. Human evaluation in WMT13 is performed on TA-GIZA++ tested on tst2013 (marked with *)

6.1 Morphological Processing

The linguistic processing of Russian involves POS tagging and morphological reduction. We first tag the Russian data using a fine grained tagset. The tagger identifies lemmas and the set of morphological attributes attached to each word. We reduce the number of these attributes by deleting some of them, that are not relevant for English (for example, gender agreement of verbs). This generates a morphologically reduced Russian which is used in parallel with English for the training of the machine translation system. Further details on the morphological processing of Russian are described in Weller et al. (2013).

6.1.1 POS Tagging

We use RFTagger (Schmid and Laws, 2008) for POS tagging. Despite the good quality of tagging provided by RFTagger, some errors seem to be unavoidable due to the ambiguity of certain grammatical forms in Russian. A good example of this is neuter nouns that have the same form in all cases, or feminine nouns, which have identical forms in singular genitive and plural nominative (Sharoff et al., 2008). Since Russian sentences have free word order, and the case of nouns cannot be determined on that basis, this imperfection can not be corrected during tagging or by post-processing the tagger output.

6.1.2 Morphological Reduction

English in comparison to Slavic group of languages is morphologically poor. For example, English has no morphological attributes for nouns and adjectives to express gender or case; verbs in English have no gender either. Russian, on the contrary, has rich morphology. It suffices to say that the Russian has 6 cases and 3 grammatical genders, which manifest themselves in different

suffixes for nouns, pronouns, adjectives and some verb forms.

When translating from Russian into English, a lot of these attributes become meaningless and excessive. It makes sense to reduce the number of morphological attributes before the text is supplied for the training of the MT system. We apply morphological reduction to nouns, pronouns, verbs, adjectives, prepositions and conjunctions. The rest of the POS (adverbs, particles, interjections and abbreviations) have no morphological attributes and are left unchanged.

We apply morphological reduction to train, tune, development and test data. We refer to this data set as *morph-reduced* later on.

6.2 Dataset

We use two variations of the parallel corpus to build and test the Russian to English system. One system is built on the data provided by the workshop. For the second system, we preprocess the Russian side of the data as described in Section 6.1. Both the provided parallel corpus and the morph-reduced parallel corpus consist of 2M parallel sentences each. We use them for the estimation of the translation model. We use large training data for the estimation of monolingual language model – en $\approx$ 287.3M sentences. We follow the identical procedure of interpolated language model as described in Section 5.1. We use newstest2012a for tuning and newstest2012b (tst2012) for development.

6.3 System Settings

We use identical system settings to those described in Section 5.3. We trained the systems separately on GIZA++ and transliteration augmented-GIZA++ to compare their results. All systems are tuned using PROv1. The translation output is post-processed to transliterate OOVs.

6.4 Results

Table 3 summarizes results of Russian to English machine translation systems trained on the original parallel corpus and on the morph-reduced corpus and using GIZA++ and transliteration augmented-GIZA++ for word alignment. The system using TA-GIZA++ for alignment shows the best results for both tst2012 and tst2013. The improved alignment gives a BLEU improvement of up to 0.4 points.

Original corpus		
SYS	**tst2012**	**tst2013**
GIZA++	32.51	25.5
TA-GIZA++	33.40	25.9*
Morph-reduced		
SYS	**tst2012**	**tst2013**
GIZA++	31.22	24.30
TA-GIZA++	31.40	24.45

Table 3: Russian to English machine translation system evaluated on tst2012 and tst2013. Human evaluation in WMT13 is performed on the system trained using the original corpus with TA-GIZA++ for alignment (marked with *)

The system built on the morph-reduced data shows degradation in results by 1.29 BLEU points. However, the percentage of OOVs reduces for both test sets when using the morph-reduced data set compared to the original parallel corpus. We analyze the output of the system and find that the morph-reduced system makes mistakes in choosing the right tense of the verb. This might be one reason for poor performance. This implies that the morphological reduction is slightly damaging the data, perhaps for specific parts of speech. In the future, we would like to investigate this issue in detail.

7 Conclusion

In this paper, we described the QCRI-Munich-Edinburgh-Stuttgart machine translation systems submitted to the Eighth Workshop on Statistical Machine Translation. We aligned the parallel corpus using transliteration augmented-GIZA++ to improve the word alignments. We built a phrase-based system using the Moses toolkit. For tuning the feature weights, we used an improvement of PRO that optimizes for corpus-level BLEU. We post-processed the output of the machine translation system to transliterate OOV words.

For the Russian to English system, we morphologically reduced the Russian data in a preprocessing step. This reduced the vocabulary size and helped to generate better word alignments. However, the performance of the SMT system dropped by 1.29 BLEU points in decoding. We will investigate this issue further in the future.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful feedback and suggestions. We would like to thank Philipp Koehn and Barry Haddow for providing data and alignments. Nadir Durrani was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n ° 287658. Alexander Fraser was funded by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation. Helmut Schmid was supported by Deutsche Forschungsgemeinschaft grant SFB 732. This publication only reflects the authors views.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2).

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, USA.

Nadir Durrani, Helmut Schmid, Alexander Fraser, Hassan Sajjad, and Richárd Farkas. 2013. Munich-Edinburgh-Stuttgart submissions of OSM systems at WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, United Kingdom.

Philipp Koehn and Barry Haddow. 2012. Towards effective use of training data in statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal, Canada.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, Edmonton, Canada.

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 Association for Computational Linguistics, Demonstration Program*, Prague, Czech Republic.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, Geneva, Switzerland.

Preslav Nakov, Francisco Guzmán, and Stephan Vogel. 2012. Optimizing for sentence-level BLEU+1 yields short translations. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2011. An algorithm for unsupervised transliteration mining with an application to word alignment. In *Proceedings of the 49th Annual Conference of the Association for Computational Linguistics*, Portland, USA.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2012. A statistical model for unsupervised and semi-supervised transliteration mining. In *Proceedings of the 50th Annual Conference of the Association for Computational Linguistics*, Jeju, Korea.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, Manchester, United Kingdom.

Holger Schwenk and Philipp Koehn. 2008. Large and Diverse Language Models for Statistical Machine Translation. In *International Joint Conference on Natural Language Processing*, Hyderabad, India.

Serge Sharoff, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a russian tagset. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *16th International Conference on Computational Linguistics*, Copenhagen, Denmark.

Marion Weller, Max Kisselew, Svetlana Smekalova, Alexander Fraser, Helmut Schmid, Nadir Durrani, Hassan Sajjad, and Richárd Farkas. 2013. Munich-Edinburgh-Stuttgart submissions at WMT13: Morphological and syntactic processing for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.