Factored Machine Translation Systems for Russian-English

Stéphane Huet, Elena Manishina and Fabrice Lefèvre
Université d'Avignon, LIA/CERI, France
FirstName.LastName@univ-avignon.fr

Abstract

We describe the LIA machine translation systems for the Russian-English and English-Russian translation tasks. Various factored translation systems were built using MOSES to take into account the morphological complexity of Russian and we experimented with the romanization of untranslated Russian words.

1 Introduction

This paper presents the factored phrase-based Machine Translation (MT) systems (Koehn and Hoang, 2007) developed at LIA, for the Russian-English and English-Russian translation tasks at WMT'13. These systems use only data provided for the evaluation campaign along with the LDC English Gigaword corpus.

We summarize in Section 2 the resources used and the main characteristics of the systems based on the MOSES toolkit (Koehn et al., 2007). Section 3 reports experiments on the use of factored translation models. Section 4 describes the transliteration process used to improve the Russian to English task. Finally, we conclude in Section 5.

2 System Architecture

2.1 Pre-processing

The corpora available for the workshop were preprocessed using an in-house script that normalizes quotes, dashes, spaces and ligatures. Long sentences or sentences with many numeric or non-alphanumeric characters were also discarded. Since the Yandex corpus is provided as lowercased, we decided to lowercase all the other corpora. The same pipeline was applied to the LDC Gigaword; also only the documents classified as "story" were retained. Table 1 summarizes the used data and introduces designations that we follow in the remainder of this paper to refer to these corpora.

Russian is a morphologically rich language with nouns, adjectives and verbs inflected for case, number and gender. This property requires introducing morphological information inside the MT system to handle the lack of many inflectional forms inside training corpora. For this purpose, each corpus was previously tagged with Part-of-Speech (PoS) tags. The tagger TREETAGGER (Schmid, 1995) was selected for its good performance on several comparable tasks. The Russian tagger associates each word (e.g. ящика (*boxes*)) with a complex PoS including morphological information (e.g. "Ncmpnn" for "Noun Type=common Gender=masculine Number=plural Case=nominative Animate=no") and its lemma (e.g. ящик (*box*)). A description of the Russian tagset can be found in (Sharoff et al., 2008). The English tagger provides also a lemmatization and outputs PoS from the Penn Treebank tagset (Marcus et al., 1993) (e.g. "NNS" for "Noun plural").

In order to simplify the comparison of different setups, we used the tokenizer included in the TREETAGGER tool to process all the corpora.

2.2 Language Models

Kneser-Ney discounted LMs were built from monolingual corpora using the SRILM toolkit (Stolcke, 2002). 5-gram LMs were trained for words, 7-gram LMs for lemmas and PoS. A LM was built separately on each monolingual corpus: *mono-news-c* and *news-s*. Since *ldc* was too large to be processed as one file, it was split into three parts according to the original publication year of the document. These LMs were combined through linear interpolation. Weights were fixed by optimizing the perplexity on a corpus made of the WMT test sets from 2008 to 2011 for English and on the WMT 2012 test set for Russian (the

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 154–157,
Sofia, Bulgaria, August 8-9, 2013

CORPORA	DESIGNATION	SIZE (SENTENCES)
English-Russian Bilingual training		
News Commentary v8	*news-c*	146 k
Common Crawl	*crawl*	755 k
Yandex	*yandex*	978 k
English Monolingual training		
News Commentary v8	*mono-news-c*	247 k
Shuffled News Crawl corpus (from 2007 to 2012)	*news-s*	68 M
LDC Gigaword	*ldc*	190 M
Russian Monolingual training		
News Commentary v8	*mono-news-c*	182 k
Shuffled News Crawl corpus (from 2008 to 2012)	*news-s*	20 M
Development		
newstest2012	*test12*	3,003

Table 1: Used bilingual and monolingual corpora

only available at that time).

2.3 Alignment and Translation Models

All parallel corpora were aligned using MGIZA++ (Gao and Vogel, 2008). Our translation models are phrase-based models (PBMs) built with MOSES using default settings. Weights of LM, phrase table and lexicalized reordering model scores were optimized on *test12*, thanks to the MERT algorithm (Gao and Vogel, 2008). Since only one development corpus was made available for Russian, we used a 3-fold cross-validation so that MERT is repeated three times for each translation model on a 2,000-sentence subsample of *test12*.

To recase the corpora, translation models were trained using a word-to-word translation model trained on the parallel corpora aligning lowercased and cased sentences of the monolingual corpora *mono-news-c* and *news-s*.

3 Experiments with Factored Translation Models

The evaluation was performed using case-insensitive BLEU and was computed with the `mteval-v13a.pl` script provided by NIST. The BLEU scores shown in the tables below are all averaged on the test parts obtained from the 3-fold cross validation process.

In the remainder of the paper, we employ the notation proposed by Bojar et al. (2012) to refer to factored translation models. For example, tW-W:tL-L+tP-P+gLaP-W, where "t" and "g" stand for "translation" and "generation", denotes a translation system with two decoding paths:

- a first one directly translates words to words (tW-W),
- a second one is divided into three steps:
  1. translation from lemmas to lemmas (tL-L),
  2. translation from PoS to PoS (tP-P) and
  3. generation of target words from target lemmas and PoS (gLaP-W).

3.1 Baseline Phrase-Based Systems

Table 2 is populated with the results of PBMs which use words as their sole factor. When LMs are built on *mono-news-c* and *news-s*, an improvement of BLEU is observed each time a training parallel corpus is used, both for both translation directions (columns 1 and 3). We can also notice an absolute increase of 0.4 BLEU score when the English LM is additionally trained on *ldc* (column 2).

3.2 Decomposition of factors

Koehn and Hoang (2007) suggested from their experiments for English-Czech systems that "it is beneficial to carefully consider which morphological information to be used." We therefore tested various decompositions of the complex Russian PoS tagset (P) output by TREETAGGER. We considered the grammatical category alone (C), morphological information restrained to case, number

	EN → RU	+LDC	RU → EN
news-c	26.52	26.82	19.89
+crawl	29.49	29.82	21.06
+yandex	31.08	**31.49**	**22.16**

Table 2: BLEU scores measured with standard PBMs.

Tagset	#tags	Examples
C	17	Af, Vm, P, C
M1	95	fsg, -s-, fsa, —
M2	380	fsg, -s-, fsa, ЧТО (*that*)
M3	580	fsg, -s-1ife, fsa3, ЧТО (*that*)
P	604	Afpfsg, Vmif1s-a-e, P-3fsa, C

Table 3: Statistics on Russian tagsets.

and gender (M1), the fields included in M1 along with additional information (lemmas) for conjunctions, particles and adpositions (M2), and finally the information included in M2 enriched with person for pronouns and person, tense and aspect for verbs (M3). Table 3 provides the number of tags and shows examples for each used tagset.

To speed up the training of translation models, we experimented with various setups for factor decomposition from *news-c*. The results displayed on Table 4 show that factors with morphological information lead to better results than a PBM trained on word forms (line 1) but that finally the best system is achieved when the complex PoS tag output by TREETAGGER is used without any decomposition (last line).

tW-W	19.89
tW-WaC	19.81
tW-WaM1	20.04
tW-WaCaM1	19.95
tW-WaM2	19.92
tW-WaCaM2	19.91
tW-WaM3	19.98
tW-WaCaM3	19.89
tW-WaP	**20.30**

Table 4: BLEU scores for EN→RU using *news-c* as training parallel corpus.

tL-W	29.23
tW-W	31.49
tWaP-WaP	31.62
tW-W:tL-W	31.69
tW-WaP	31.80
tW-WaP:tL-WaP	**31.89**

Table 5: BLEU scores for RU→EN using the three available parallel corpora.

3.3 Experimental Results for Factored Models

The many inflections for Russian induce a hight out-of-vocabulary rate for the PBMs, which generates many untranslated Russian words for Russian to English. We experimented with the training of a PMB on lemmatized Russian corpora (Table 5, line 1) but observed a decrease in BLEU score w.r.t. a PBM trained on words (line 2). With two decoding paths — one from words, one from lemmas (line 4) — using the MOSES ability to manage multiple decoding paths for factored translation models, an absolute improvement of 0.2 BLEU score was observed.

Another interest of factored models is disambiguating translated words according to their PoS. Translating a (word, PoS) pair results in an absolute increase of 0.3 BLEU (line 5), and of 0.4 BLEU when considering two decoding paths (last line). Disambiguating source words with PoS did not seem to help the translation process (line 3).

The Russian inflections are far more problematic in the other translation direction since morphological information, including case, gender and number, has to be induced from the English words and PoS, which are restrained for that language to the grammatical category and knowledge about number (singular/plural for nouns, 3rd person singular or not for verbs). Disambiguating translated Russian words with their PoS resulted in a dramatic increase of BLEU by 1.6 points (Table 6, last line vs line 3). The model that translates independently PoS and lemmas, before generating words, albeit appealing for its potential to deal with data sparsity, turned out to be very disappointing (first line). We additionally led experiments training generation models gLaP-W on monolingual corpora instead of the less voluminous parallel corpora, but we did not observed a gain in terms of BLEU.

tL-L+tP-P+gLaP-W	17.06
tW-W	22.16
tWaP-WaP	23.34
tWaP-LaP+gLaP-W	23.48
tW-LaP+gLaP-W	23.58
tW-WaP	**23.72**

Table 6: BLEU scores for EN→RU using the three available parallel corpora.

	BEFORE	AFTER
tW-WaP	31.80	32.15
tW-WaP:tL-WaP	31.89	**32.21**

Table 7: BLEU scores for RU → EN before and after transliteration.

4 Transliteration

Words written in Cyrillic inside the English translation output were transliterated into Latin letters. We decided to restrain the use of transliteration for the English to Russian direction since we found that many words, especially proper names, are intentionally used in Latin letters in the Russian reference.

Transliteration was performed in two steps. Firstly, untranslated words in Cyrillic are looked up in the *guessed-names.ru-en* file provided for the workshop and built from Wikipedia. Secondly, the remaining words are romanized with rules of the BGN/PCGN romanization method for Russian (on Geographic Names, 1994). Transliterating words in Cyrillic resulted in an absolute improvement of 0.3 BLEU for our two best factor-based system (Table 7, last column).

The factored model with the tW-WaP:tL-WaP translation path and a transliteration post-processing step is the final submission for the Russian-English workshop translation task, while the tW-WaP is the final submission for the other translation direction.

5 Conclusion

This paper presented experiments carried out with factored phrase-based translation models for the two-way Russian-English translation tasks. A minor gain was observed after romanizing Russian words (+0.3 BLEU points for RU → EN) and higher improvements using word forms, PoS integrating morphological information and lemma as factors (+0.4 BLEU points for RU → EN and +1.6 for EN → RU w.r.t. to a phrase-based restrained to word forms). However, these improvements were observed with setups which disambiguate words according to their grammatical category or morphology, while results integrating a generation step and dealing with data sparsity were disappointing. It seems that further work should be done to fully exploit the potential of this option inside MOSES.

References

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. 2012. Probes in a taxonomy of factored phrase-based models. In *7th NAACL Workshop on Statistical Machine Translation (WMT)*, pages 253–260.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Proceedings of the ACL Workshop: Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868—-876.

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 *45th Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume*, pages 177–180.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 2:313–330.

U.S. Board on Geographic Names. 1994. Romanization systems and roman-script spelling conventions. Technical report, Defense Mapping Agency.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *ACL SIGDAT Workshop*, pages 47–50.

Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a russian tagset. In *6th International Conference on Language Resources and Evaluation (LREC)*, pages 279–285.

Andreas Stolcke. 2002. SRILM — an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing (ICSLP)*.