Yandex School of Data Analysis machine translation systems for WMT13

Alexey Borisov, Jacob Dlougach, Irina Galinskaya
Yandex School of Data Analysis
16, Leo Tolstoy street, Moscow, Russia
{alborisov,jacob,galinskaya}@yandex-team.ru

Abstract

This paper describes the English-Russian and Russian-English statistical machine translation (SMT) systems developed at Yandex School of Data Analysis for the shared translation task of the ACL 2013 *Eighth Workshop on Statistical Machine Translation*. We adopted phrase-based SMT approach and evaluated a number of different techniques, including data filtering, spelling correction, alignment of lemmatized word forms and transliteration. Altogether they yielded +2.0 and +1.5 BLEU improvement for ru-en and en-ru language pairs. We also report on the experiments that did not have any positive effect and provide an analysis of the problems we encountered during the development of our systems.

1 Introduction

We participated in the shared translation task of the ACL 2013 Workshop on Statistical Machine Translation (WMT13) for ru-en and en-ru language pairs. We provide a detailed description of the experiments carried out for the development of our systems.

The rest of the paper is organized as follows. Section 2 describes the tools and data we used. Our Russian→English and English→Russian setups are discussed in Section 3. In Section 4 we report on the experiments that did not have any positive effect despite our expectations. We provide a thorough analysis of erroneous outputs in Section 5 and draw conclusions in Section 6.

2 Tools and data

2.1 Tools

We used an open source SMT system Moses (Koehn et al., 2007) for all our experiments excluding the one described in Section 4.1 due to its performance constraints. To overcome the limitation we employed our in-house decoder.

Language models (LM) were created with an open source IRSTLM toolkit (Federico et al., 2008). We computed 4-gram LMs with modified Kneser-Ney smoothing (Kneser and Ney, 1995).

We used an open source MGIZA++ tool (Gao and Vogel, 2008) to compute word alignment.

To obtain part of speech (POS) tags we used an open source Stanford POS tagger for English (Toutanova et al., 2003) and an open source suite of language analyzers, FreeLing 3.0 (Carreras et al., 2004; Padró and Stanilovsky, 2012), for Russian.

We utilized a closed source free for non-commercial use morphological analyzer, Mystem (Segalovich, 2003), that used a limited dictionary to obtain lemmas.

We also made use of the in-house language recognizer based on (Dunning, 1994) and a spelling corrector designed on the basis of the work of Cucerzan and Brill (2004).

We report all results in case-sensitive BLEU (Papineni et al., 2002) using mt-eval13a script from Moses distribution.

2.2 Data

Training data

We used News Commentary and News Crawl monolingual corpora provided by the organizers of the workshop.

Bilingual training data comprised English-Russian parallel corpus release by Yandex[1], News Commentary and Common Crawl corpora provided by the organizers.

We also exploited Wiki Headlines collection of three parallel corpora provided by CMU[2] as a

[1] https://translate.yandex.ru/corpus

[2] http://www.statmt.org/wmt13/wiki-titles.ru-en.tar.gz

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 99–103,
Sofia, Bulgaria, August 8-9, 2013

source of reliable data.

Development set

The newstest2012 test set (Callison-Burch et al., 2012) was divided in the ratio 2:1 into a tuning set and a test set. The latter is referred to as newstest2012-test in the rest of the paper.

3 Primary setups

3.1 Baseline

We built the baseline systems according to the instructions available at the Moses website[3].

3.2 Preprocessing

The first thing we noticed was that some sentences marked as Russian appeared to be sentences in other languages (most commonly English). We applied a language recognizer for both monolingual and bilingual corpora. Results are given in Table 1.

Corpus	Filtered out (%)
Bilingual	3.39
Monolingual (English)	0.41
Monolingual (Russian)	0.58

Table 1: Results of the language recognizer: percentage of filtered out sentences.

The next thing we came across was the presence of a lot of spelling errors in our training data, so we applied a spelling corrector. Statistics are presented in Table 2.

Corpus	Modified (%)
Bilingual (English)	0.79
Bilingual (Russian)	1.45
Monolingual (English)	0.61
Monolingual (Russian)	0.52

Table 2: Results of the spelling corrector: percentage of modified sentences.

3.3 Alignment of lemmatized word forms

Russian is a language with rich morphology. The diversity of word forms results in data sparseness that makes translation of rare words difficult. In some cases inflections do not contain any additional information and are used only to make an agreement between two words. E.g. ADJ + NOUN: красив|ая| арфа *(beautiful harp),* красив|ое| пианино *(beautiful piano),* красив|ый| рояль *(beautiful grand piano).* These inflections reflect the gender of the noun words, that has no equivalent in English.

In this particular case we can drop the inflections, but for other categories they can still be useful for translation, because the information they contain appears in function words in English. On the other hand, most of Russian morphology is useless for word alignment.

We applied a morphological analyzer Mystem (Segalovich, 2003) to the Russian text and converted each word to its dictionary form. Next we computed word alignment between the original English text and the lemmatized Russian text. All the other steps were executed according to the standard procedure with the original texts.

3.4 Phrase score adjustment

Sometimes phrases occur one or two times in the training corpus. In this case the corresponding phrase translation probability would be overestimated. We used Good-Turing technique described in (Gale, 1994) to decrease it to some more realistic value.

3.5 Decoding

Minimum Bayes-Risk (MBR)

MBR decoding (Kumar and Byrne, 2004) aims to minimize the expected loss of translation errors. As it is not possible to explore the space of all possible translations, we approximated it with the 1,000 most probable translations. A minus smoothed BLEU score (Lin and Och, 2004) was used for the loss function.

Reordering constrains

We forbade reordering over punctuation and translated quoted phrases independently.

3.6 Handling unknown words

The news texts contained a lot of proper names that did not appear in the training data. E.g. almost 25% of our translations contained unknown words. Dropping the unknown words would lead to better BLEU scores, but it might had caused bad effect on human judgement. To leave them in Cyrillic was not an option, so we exploited two approaches: incorporating reliable data from Wiki Headlines and transliteration.

[3] http://www.statmt.org/moses/?n=moses.baseline

	newstest2012-test	newstest2013
Russian→English		
Baseline	28.96	21.82
+ Preprocessing	29.59	22.28
+ Alignment of lemmatized word forms	29.97	22.61
+ Good-Turing	30.31	22.87
+ MBR	30.45	23.21
+ Reordering constraints	30.54	23.33
+ Wiki Headlines	30.68	23.46
+ Transliteration	30.93	23.73
English→Russian		
Baseline	21.96	16.24
+ Preprocessing	22.48	16.76
+ Good-Turing	22.84	17.13
+ MBR and Reordering constraints	23.27	17.45
+ Wiki Headlines and Transliteration	23.54	17.80

Table 3: Experimental results in case-sensitive BLEU for Russian→English and English→Russian tasks.

Wiki Headlines

We replaced the names occurring in the text with their translations, based on the information in "guessed-names" corpus from Wiki Headlines.

As has been mentioned in Section 3.3, Russian is a morphologically rich language. This often makes it hard to find exactly the same phrases, so we applied lemmatization of Russian language both for the input text and the Russian side of the reference corpus.

Russian→English transliteration

We gained considerable improvement from incorporating Wiki Headlines, but still 17% of translations contained Cyrillic symbols.

We applied a transliteration algorithm based on (Knight and Graehl, 1998). This technique yielded us a significant improvement, but introduced a lot of errors. E.g. Джеймс Бонд *(James Bond)* was converted to *Dzhejms Bond*.

English→Russian transliteration

In Russian, it is a common practice to leave some foreign words in Latin. E.g. the names of companies: *Apple, Google, Microsoft* look inadmissible when either translated directly or transliterated.

Taking this into account, we applied the same transliteration algorithm (Knight and Graehl, 1998), but replaced an unknown word with its transliteration only if we found a sufficient number of occurrences of its transliterated form in the monolingual corpus. We used five for such number.

3.7 Experimental results

We summarized the gains from the described techniques for Russian→English and English→Russian tasks on Table 3.

4 What did not work

4.1 Translation in two stages

Frequently machine translations contain errors that can be easily corrected by human post-editors. Since human aided machine translation is cost-efficient, we decided to address this problem to the computer.

We propose to translate sentences in two stages. At the first stage a SMT system is used to translate the input text into a preliminary form (in target language). At the next stage the preliminary form is translated again with an auxiliary SMT system trained on the translated and the target sides of the parallel corpus.

We encountered a technical challenge, when we had to build a SMT system for the second stage. A training corpus with one side generated with the first stage SMT system was not possible to be acquired with Moses due to its performance constraints. Thereupon we utilized our in-house SMT decoder and managed to translate 2M sentences in time.

We applied this technique both for ru-en and en-ru language pairs. Approximately 20% of the sen-

tences had changed, but the BLEU score remained the same.

4.2 Factored model

We tried to build a factored model for ru-en language pair with POS tags produced by Stanford POS tagger (Toutanova et al., 2003).

Unfortunately, we did not gain any improvements from it.

5 Analysis

We carefully examined the erroneous outputs of our system and compared it with the outputs of the other systems participating in ru-en and en-ru tasks, and with the commercial systems available online (Bing, Google, Yandex).

5.1 Transliteration

Russian→English

The standard transliteration procedure is not invertible. This means that a Latin word being transfered into Cyrillic and then transliterated back to Latin produces an artificial word form. E.g. Хавард Хальварсен **/** *Havard Halvarsen* was correctly transliterated by only four out of 23 systems, including ours. Twelve systems either dropped one of the words or left it in Cyrillic. We provide a list of typical mistakes in order of their frequency: *Khavard Khalvarsen, Khavard Khal'varsen, Xavard Xaljvarsen*. Another example: Мисс Уайэтт *(Miss Wyatt)* → *Miss Uayett* (all the systems failed).

The next issue is the presence of non-null inflections that most certainly would result in wrong translation by any straight-forward algorithm. E.g. Хайдельберг[а] *(Heidelberg)* → *Heidelberga*.

English→Russian

In Russian, most words of foreign origin are written phonetically. Thereby, in order to obtain the best quality we should transliterate the transcription, not the word itself. E.g. the French derived name *Elsie Monereau* ['elsi monə'rəʊ] being translated by letters would result in Элси Монереау while the transliteration of the transcription would result in the correct form Элси Монро.

5.2 Grammars

English and Russian make use of different grammars. When the difference in their sentence structure becomes fundamental the phrase-based approach might get inapplicable.

Word order

Both Russian and English are classified as subject-verb-object (SOV) languages, but Russian has rather flexible word order compared to English and might frequently appear in other forms. This often results in wrong structure of the translated sentence. A common mistake made by our system and reproduced by the major online services: не изменились и правила *(rules have not been changed either)* → *have not changed and the rules*.

Constructions

- **there is / there are** is a non-local construction that has no equivalent in Russian. In most cases it can not be produced from the Russian text. E.g. на столе стоит матрёшка *(there is a matryoshka doll on the table)* → *on the table is a matryoshka*.

- **multiple negatives** in Russian are grammatically correct ways to express negation (a single negative is sometimes incorrect) while they are undesirable in standard English. E.g. Там никто никогда не был *(nobody has ever been there)* being translated word by word would result in *there nobody never not was*.

5.3 Idioms

Idiomatic expressions are hard to discover and dangerous to translate literary. E.g. a Russian idiom была не была *(let come what may)* being translated word by word would result in *was not was*. Neither of the commercial systems we checked managed to collect sufficient statistic to translate this very popular expression.

6 Conclusion

We have described the primary systems developed by the team of Yandex School of Data Analysis for WMT13 shared translation task.

We have reported on the experiments and demonstrated considerable improvements over the respective baseline. Among the most notable techniques are data filtering, spelling correction, alignment of lemmatized word forms and transliteration. We have analyzed the drawbacks of our systems and shared the ideas for further research.

References

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT12)*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

Silviu Cucerzan and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 293–300.

Ted Dunning. 1994. Statistical identification of language. Technical report, Computing Research Lab (CRL), New Mexico State University, Las Cruces, NM, USA.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1618–1621.

William Gale. 1994. Good-Turing smoothing without tears. *Journal of Quantitative Linguistics (JQL)*, 2:217–237.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 49–57.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 181–184.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondřej Bojar. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 163–171.

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 (COLING)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul, Turkey, May.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Processings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In Hamid R. Arabnia and Elena B. Kozerenko, editors, *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications (MLMTA)*, pages 273–280, Las Vegas, NV, USA, June. CSREA Press.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 252–259.