Yandex School of Data Analysis Russian-English Machine Translation System for WMT14

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

Abstract

This paper describes the Yandex School of Data Analysis Russian-English system submitted to the ACL 2014 Ninth Workshop on Statistical Machine Translation shared translation task. We start with the system that we developed last year and investigate a few methods that were successful at the previous translation task including unpruned language model, operation sequence model and the new reparameterization of IBM Model 2. Next we propose a {simple yet practical} algorithm to transform Russian sentence into a more easily translatable form before decoding. The algorithm is based on the linguistic intuition of native Russian speakers, also fluent in English.

1 Introduction

The annual shared translation task organized within the ACL Workshop on Statistical Machine Translation (WMT) aims to evaluate the state of the art in machine translation for a variety of languages. We participate in the Russian to English translation direction.

The rest of the paper is organized as follows. Our baseline system as well as the experiments concerning the methods already discussed in literature are described in Section 2. In Section 3 we present an algorithm we use to transform the Russian sentence before translation. In Section 4 we discuss the results and conclude.

2 Initial System Development

We use all the Russian-English parallel data available in the constraint track and the Common Crawl English monolingual corpus.

2.1 Baseline

We use the phrase-based Moses statistical machine translation system (Koehn et al., 2007) with mostly default settings and a few changes (Borisov et al., 2013) made in the following steps.

Data Preprocessing includes filtering out non Russian-English sentence pairs and correction of spelling errors.

Phrase Table Smoothing uses Good-Turing scheme (Foster et al., 2006).

Consensus Decoding selects the translation with minimum Bayes risk (Kumar and Byrne, 2004).

Handling of Unknown Words comprises incorporation of proper names from Wiki Headlines parallel data provided by CMU[1] and transliteration. We improve the transliteration algorithm in Section 2.4.

Note that unlike last year we do not use word alignments computed for the lemmatized word forms.

2.2 Language Model

We use 5-gram unpruned language model with modified Kneser-Ney discount estimated with KenLM toolkit (Heafield et al., 2013).

2.3 Word alignment

Word alignments are generated using the fast_align tool (Dyer et al., 2013), which is much faster than IBM Model 4 from MGIZA++ (Gao and Vogel, 2008) and outperforms the latter in terms of BLEU. Results are given in Table 1.

2.4 Transliteration

We employ machine transliteration to generate additional translation options for out-of-vocabulary

[1] http://www.statmt.org/wmt14/wiki-titles.tgz

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 66–70,
Baltimore, Maryland USA, June 26–27, 2014.

	MGIZA++	fast_align
Run Time	22 h 14 m	2h 49 m
Perplexity		
– ru→en	97.00	90.37
– en→ru	209.36	216.71
BLEU		
– WMT13	25.27	25.49
– WMT14	31.76	31.92

Table 1: Comparison of word alignment tools: MGIZA++ vs. fast_align. fast_align runs ten times as fast and outperforms the IBM Model 4 from MGIZA++ in terms of BLEU scores.

words. The transformation model we use is a transfeme based model (Duan and Hsu, 2011), which is analogous to translation model in phrase-based machine translation. Transformation units, or transfemes, are trained with Moses using the default settings. Decoding is very similar to beam search. We build a trie from the words in English monolingual corpus, and search in it, based on the transformation model.

2.5 Operation Sequence Model

The Operation Sequence N-gram Model (OSM) (Durrani et al., 2011) integrates reordering operations and lexical translations into a heterogeneous sequence of minimal translation units (MTUs) and learns a Markov model over it. Reordering decisions influence lexical selections and vice versa thus improving the translation model. We use OSM as a feature function in phrase-based SMT. Please, refer to (Durrani et al., 2013) for implementation details.

3 Morphological Transformations

Russian is a fusional synthetic language, meaning that the relations between words are redundant and encoded inside the words. Adjectives alter their form to reflect the gender, case, number and in some cases, animacy of the nouns, resulting in dozens of different word forms matching a single English word. An example is given in Table 2. Verbs in Russian are typically constructed from the morphemes corresponding to functional words in English (to, shall, will, was, were, has, have, had, been, etc.). This Russian phenomenon leads to two problems: data sparsity and high number of one-to-many alignments, which both may result in translation quality degradation.

		Number	
		SG	PL
Case	**Gender**		
NOM	MASC	летний	летние
NOM	FEM	летняя	
NOM	NEUT	летнее	
GEN	MASC	летнего	летних
GEN	FEM	летней	
GEN	NEUT	летнего	
DAT	MASC	летнему	летним
DAT	FEM	летней	
DAT	NEUT	летнему	
ACC	MASC, AN	летнего	летним
ACC	MASC, INAN	летний	
ACC	FEM	летнюю	
ACC	NEUT	летнее	
INS	MASC	летним	летним
INS	FEM	летней	
INS	FEM	летнею	
INS	NEUT	летним	
ABL	MASC	летнем	летних
ABL	FEM	летней	
ABL	NEUT	летнем	

Table 2: Russian word forms corresponding to the English word "summer" (adj.).

Hereafter, we propose an algorithm to transform the original Russian sentence into a more easily translatable form. The algorithm is based on the linguistic intuition of native Russian speakers, also fluent in English.

3.1 Approach

Based on the output from Russian morphological analyzer we rewrite the input sentence based on the following principles:

1. the original sentence is restorable (by a Russian native speaker)
2. redundant information is omitted
3. word alignment is less ambiguous

3.2 Algorithm

The algorithm consists of two steps.

On the first step we employ in-house Russian morphological analyzer similar to Mystem (Segalovich, 2003) to convert each word (WORD) into a tuple containing its canonical form (LEMMA), part of speech tag (POS) and a set

Category	Abbr.	Values
Animacy	ANIM	AN, INAN
Aspect	ASP	IMPERF, PERF
Case	CASE	NOM, GEN, DAT, ACC, INS, ABL
Comparison Type	COMP	COMP, SURP
Gender	GEND	MASC, FEM, NEUT
Mood	MOOD	IND, IMP, COND, SBJV
Number	NUM	SG, PL
Participle Type	PART	ACT, PASS
Person	PERS	PERS1, PERS2, PERS3
Tense	TNS	PRES, NPST, PST

Table 3: Morphological Categories

of other grammemes associated with the word (GRAMMEMES). The tuple is later referred to as LPG. If the canonical form or part of speech are ambiguous, we set LEMMA to WORD; POS to "undefined"; and GRAMMEMES to ∅. Grammemes are grouped into grammatical categories listed in Table 3.

WORD ⟶ LEMMA + POS + GRAMMEMES

On the second step, the LPGs are converted into tokens that, we hope, will better match English structure. Some grammemes result in separate tokens, others stay with the lemma, and the rest get dropped. The full set of morphological transformations we use is given in Table 4.

An example of applying the algorithm to a Russian sentence is given in Figure 1.

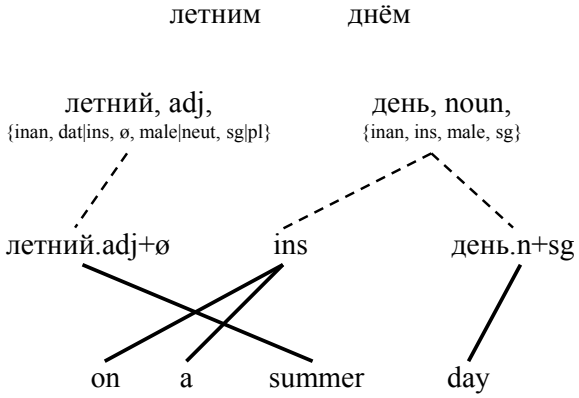

Figure 1: An illustration of the proposed algorithm to transform Russian sentence летним днём *(letnim dnem)*, meaning *on a summer day*, into a more easily translatable form. First, for each word we extract its canonical form, part of speech tag and a set of associated morphological properties (grammemes). Then we apply hand-crafted rules (Table 4) to transform them into separate tokens.

3.3 Results

The translation has been improved in several ways:

Incorrect Use of Tenses happens quite often in statistical machine translation, which is especially vexing in simple cases such as *asks* instead of *asked*, *explains* instead of *explain* along with more difficult ones e.g. *has increased* instead of *would increase*. The proposed algorithm achieves considerable improvement, since it explicitly models tenses and all its relevant properties.

Missing Articles is a common problem of most Russian-English translation systems, because there are no articles in Russian. Our model creates an auxiliary token for each noun, which reflects its case and motivates an article.

Use of Simple Vocabulary is not desirable when the source text is a vocabulary-flourished one. News are full of academic, bookish, inkhorn, and other rare words. Phrase Table smoothing methods discount the translation probabilities for rare phrase pairs, preventing them from appearing in English translation, while many of these rare phrase pairs are correct. The good thing is that the phrase pairs containing the transformed Russian words may not be rare themselves, and thereby are not discounted so heavily. A more effective use of English vocabulary has been observed on WMT13 test dataset (see Table 5).

We have demonstrated the improvements on a qualitative level. The quantitative results are summarized in Table 6 (baseline – without morphological transformations; proposed – with morphological transformations).

LPG ⇒ tokens
LEMMA, adj, {ANIM, CASE, COMP, GEND, NUM} ⇓ LEMMA.adj+COMP
LEMMA, noun, {ANIM, CASE, GEND, NUM} ⇓ CASE LEMMA.n+NUM
LEMMA, verb (ger), {ASP, TNS} ⇓ LEMMA.vg+ASP+TNS
LEMMA, verb (inf), {ASP} ⇓ LEMMA.vi+ASP
LEMMA, verb (part), {PART, ASP, TNS} ⇓ LEMMA.vp+PART+ASP+TNS
LEMMA, verb (–), {PART, ASP, MOOD, TENSE, NUM, PERS} ⇓ 1. TNS={PRES} ∣ TNS={NPST} & ASP={IMPERF} a. PERS3 ∈ PERS & SG ∈ NUM LEMMA.v+pres+MOOD+PERS+NUM b. otherwise LEMMA.v+pres+MOOD 2. TNS={PST} ASP LEMMA.v+pst+MOOD 3. TNS={NPST} & ASP={IMPERF} fut LEMMA.v+MOOD 4. if ambiguous LEMMA.v+PART+ASP+MOOD +TNS+NUM+PERS
LEMMA, OTHER, GRAMMEMES ⇓ LEMMA.POS+GRAMMEMES

Table 4: A set of rules we use to transform the LPGs (LEMMA, POS, GRAMMEMES), extracted on the first step, into individual tokens.

4 Discussion and Conclusion

We described the Yandex School of Data Analysis Russian-English system submitted to the ACL 2014 Ninth Workshop on Statistical Machine Translation shared translation task. The main contribution of this work is an algorithm to transform the Russian sentence into a more easily translatable form before decoding. Significant improvements in human satisfaction and BLEU scores have been demonstrated from applying this algorithm.

Input	**Translation**
разногласия (raznoglasiya)	(a) differences (b) disputes
пропагандистом (propagandistom)	(a) promoter (b) propagandist
преимущественно (preimuschestvenno)	(a) mainly (b) predominantly

Table 5: Morphological Transformations lead to more effective use of English vocabulary. Translations marked with *"a"* were produced using the baseline system; with *"b"* also use Morphological Transformations.

	Baseline	**Proposed**
Distinct Words	899,992	564,354
OOV Words		
– WMT13	829	590
– WMT14	884	660
Perplexity		
– ru→en	90.37	99.81
– en→ru	216.71	128.15
BLEU		
– WMT13	25.49	25.63
– WMT14	31.92	32.56

Table 6: Results of Morphological Transformations. We improved the statistical characteristics of our models by reducing the number of distinct words by 37% and managed to translate 25% of previously untranslated words. BLEU scores were improved by 0.14 and 0.64 points for WMT13 and WMT14 test sets respectively.

One limitation of the proposed algorithm is that it does not take into account the relations between words sharing the same root. E.g. the word аистиных *(aistinyh)* meaning stork (adj.) is handled independently from the word аист *(aist)* meaning stork (n.). Our system as well as the major online services (Bing, Google, Yandex) transliterated this word, but the word *aistinyh* does not make much sense to a non-Russian reader. It might be worthwhile to study this problem in more detail.

Another direction for future work is to apply the proposed algorithm in reverse direction. We suggest the following two-step procedure. English

sentence is first translated into Russian* (Russian after applying the morphological transformations), and at the next step it is translated again with an auxiliary SMT system trained on the (Russian*, Russian) parallel corpus created from the Russian monolingual corpus.

References

Alexey Borisov, Jacob Dlougach, and Irina Galinskaya. 2013. Yandex school of data analysis machine translation systems for wmt13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*, pages 97–101. Association for Computational Linguistics.

Huizhong Duan and Bo-June Paul Hsu. 2011. Online spelling correction for query completion. In *Proceedings of the 20th international conference on World Wide Web (WWW)*, pages 117–126. ACM.

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 (ACL)*, pages 1045–1054. Association for Computational Linguistics.

Nadir Durrani, Barry Haddow, Kenneth Heafield, and Philipp Koehn. 2013. Edinburgh's machine translation systems for european language pairs. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*, pages 112–119. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 644–648. Association for Computational Linguistics.

George Foster, Roland Kuhn, and John Howard Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 53–61. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 690–696, Sofia, Bulgaria, August. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.