Unknown words in Statistical Machine Translation between morphologically rich and poor languages

Abstract

In this paper we address the problem of unknown words in Statistical Machine Translation (SMT) with respect to the morphological complexity of languages. We trained the Statistical Machine Translation system Moses for Russian-to-English - translating from the morphologically rich to the morphologically poor language - and Russian-to-Czech - the translation between two morphologically rich related languages. After the analysis of out-of-vocabulary word types, we show the ways to reduce the rate of out-of-vocabulary words (OOV), exploiting morphological analyzers and stemming techniques, and discuss the relation of OOV and other metrics. We also provide a manual evaluation of the oov words, namely how each of the suggested methods helped and show possible directions of future research of this problem.

1 Introduction

The most frequent SMT errors lowering the translation quality are untranslated words, called out-of-vocabulary words (or OOV) in this paper. Other errors (wrong morphological form of a word, syntax errors) make the translated text inconvenient to read, but still somehow understandable. But the unknown words are just kept as-is in the target language, thus giving the reader no information at all. Therefore, it is crucial to return at least some translation, even in a wrong word form.

Why can a word be OOV? In SMT, the word is kept untranslated if its form has not been seen in the training data. It might be the case of a completely out-of-domain word, but it might also be the case of another morphemic form of a word, already present in the training data.

The latter presents a challenge for morphologically richer languages, such as the whole family of Slavic languages, where one word can have tens of morphemic forms.

Researchers have been improving the OOV rate, both disregarding and taking morphological richness of languages into account.

Some authors (Habash 2008), (Turchi et al. 2011), (Bojar and Tamchyna 2011) address the problem of how to reduce the OOV rate suggesting various techniques, as, for example, introducing morphological information or additional dictionary resources.

Exploiting the surface form of a word - division into morphemes, stemming - brought positive results in terms of increasing the percentage of translated words especially when building a translation model from and to morphologically rich languages (Popovic and Ney 2004), (Oflazer 2008), (Gispert et al. 2005). Our approach mainly follows the line of research described above - making use of morphological resources and exploiting simple stemming techniques.

This paper also discusses the question of relation between language similarity and translation quality.

In the past, when the statistical models were not prevalent in machine translation and the main trend were rule-based systems, it was assumed that translation between related, but morphologically rich languages will be easier than between less related, but morphologically poorer one; for example, it was as-

English	Czech	Russian
jolly elephant	veselý slon	veselyj slon
	veselého slona	veselogo slona
	veselému slonu	veselomu slonu
	veselému slonu	veselom slone
	veselým slonem	veselym slonom
	veselý slone	-

Table 1: Declension of a noun phrase.

sumed that the translation between Czech and Russian will be easier, than between Czech and English.

Czech and Russian are both Slavic languages. They share a very similar morphological and syntactic structure (declension types, word order) and the surface form of morphemes. These properties might have been useful for the rule-based machine translation.

However, as we found out, this similarity surprisingly plays no role in the SMT, and, as we will show further, the translation between Czech and Russian demonstrates lower quality output than between English and any of the two languages.

2 Characterstics of Slavic languages

Slavic languages are mostly inflecting languages characterized by free word order and rich inflectional paradigms. The Table 1 shows the exposion of word forms in Slavic languages on We can argue that this the example of a noun phrase. (Russian has been transliterated.)

The above example of declension demonstrates the morphological complexity of Czech and Russian.

3 Out-of-vocabulary words

In statistical machine translation from and especially to the morphologically rich languages, high out-of-vocabulary rate and mistakes in morphological forms are typical.

As demonstrated in the previous section, Slavic languages are very morphologically rich. This creates a problem of data sparseness that increase the number of out-of-vocabulary words.

3.1 Calculating OOV rate

In some cases, it's hard to specifically count the OOV rate.

However, we specifically chose those language pairs, when one language is written in cyrillic, while the other is written in latin alphabet. This allows us to count OOV rate in a very efficient way, since cyrillic characters in non-cyrillic text almost surely mean untranslated cyrillic word and vice versa.

(We can sometimes argue that latin alphabet in Russian text is understandable as, for example, a named entity; even though it is most time just an OOV word, we are taking the oposite way in our experiment, as there is almost zero chance of cyrillic letters appearing in the Czech or English data.)

3.2 Relation between BLEU and OOV rates

We must always be reminded that decreasing OOV rate is not a goal in itself; we want to decrease it only to increase the translation quality.

For a trivial example - if we translate all words of the source language to one specific word in target language, we would have low OOV rate but also low translation quality.

For this, we have to measure both the OOV rate and the translation quality, in our case by counting BLEU score.

4 Decreasing the OOV rate with factors

The basic idea for our experiments was: we want to eliminate the cases, where we don't see a given word in training data in the source language, because it was there in a different form.

Therefore, we took all the morphological forms of the given word on the source side and transform them to the same wordform (how exactly is described further), but we let the words on the target side be as-is.

Because of that, we don't have to do any word generation (which is a positive), but on the other hand, the machine translation will never return a word, which don't appear on the target side in the training data (which is negative). We decided for this factored model because the generation models for Czech are not that well developed.

We then did two different experiments - first, we used just phrase tables created as described above,

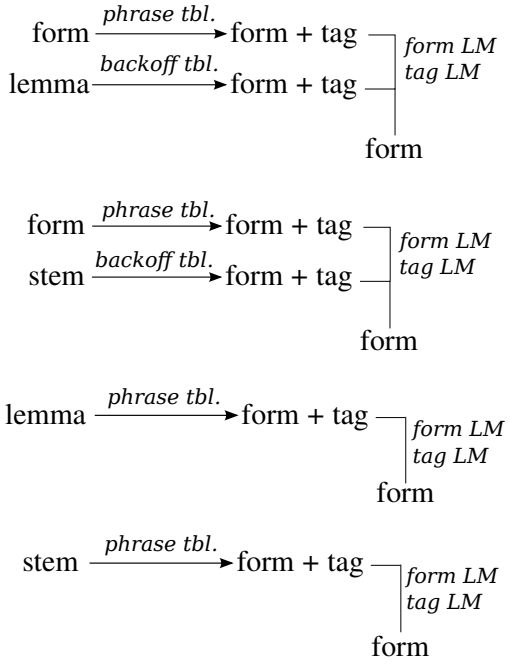

Figure 1: Factored set-up (various versions)

and then we used them only as a so-called backoff model [1].

For more grammatical sentences, we also added tags to the target side and used them to build a language model.

We used two ways of transformation of the various forms of the same word to the same wordform - the lemmatization and stemming. We will describe them in the following subsections.

The whole factored set-up can be seen in the Figure 1.

4.1 Lemmatization, tagging

We are exploiting the morphological features in two ways. First, we use lemmatization of source language (in this case, Russian) for making OOV rates smaller. Also, we use tagging of the target languages for better language models.

For Czech and English, we used Morče[2] - morphological perceptron-based tagger, developed at Charles' University in Prague. For Russian, we used TreeTagger[3] - modular tagger, developed at University of Stuttgart, with the Russian parameter file developed by Serge Sharoff.

Because the Russian is always the source language in this case, the tagset is not relevant. For Czech, Morče uses Czech positional tagset[4]. For English, Morče uses tagset derived from Penn Treebank.

4.2 "Brute force" stemming

In general, stemming is deriving a stem (root) for a word form in languages, where it makes sense.

In Slavic languages the inflection is done on the end of the word, so the stem is usually some beginning of the word form, unchanged.

There are some stemmers for all three languages that are trying to do this "properly". However, in our case, what we mean by stemming is just taking first n letters of each word, with varying n from 3 to 6 for different experiments.

We can argue that this stemming is really just "brute force" and should not give any good results. Surprisingly, the results are actually better than with the lemmatization.

5 Statistical Machine Translation setup

5.1 Moses

From the various available tools, we chose to experiment with Moses, an open-source implementation of phrase-based statistical translation system.

The Moses toolkit (Koehn et al. 2007) is a complex system which includes many components for data preprocessing and MT evaluation, for example GIZA++ involved in finding word alignment, the SRI Language Modeling Toolkit and the built-in implementation of model optimization (Minimum Error Rate Training, MERT) on a given development set of sentences.

Factors have been described in the section above.

5.2 Data

In our experiment we exploited data from a parallel Czech-English-Russian corpus called UMC (UFAL

[1] see the Moses documentation - http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc18

[2] http://ufal.mff.cuni.cz/morce/

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4] See http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html

	Languages	Sentences
Train	ru → cs	93395
Train	ru → en	92775
Dev	cs, en, ru	765
Test	cs, en, ru	2000

Table 2: Parallel corpus size.

Multilingual Corpus) with automatic pairwise sentence alignment. The texts were downloaded from the Project Syndicate[5] page. The data are divided into three sets: training set (train), development set (dev) for MERT and test set for testing BLEU and OOV rates. The statistics of the data are summarized in the Table 2.

As we can see, the number of sentences is slightly different for different language pairs; it is because the sentences are not always aligned one-to-one, but often one-to-many, many-to-many or zero-to-many. We decided to take all of the sentences, where the alignment is non-zero on one end.

6 Results

We use the following convention in this section: the baseline result is just called ”baseline”. Baseline doesn't use any tagging, stemming or lemmatizing; it is ”only” run with the basic Moses setting.

The versions with only stem or lemma on the source side are marked with ”1-” at the beginning and the versions with both form and lemma on the source side are marked with ”2-” (as in ”1 phrase table” or ”2 phrase tables”).

The various version are then named according to the technique of decreasing OOV - that is, ”-stem3” with stem, ”-lemma” with lemmatization.

So, for example, ”2-lemma” means the version with lemmatization and two phrase tables.

6.1 Russian to Czech

The results for Russian to Czech translation are in the Figure 2 and in Table 3 (with the 95% confidence intervals with BLEU). As we can see, 1-stem3 decreased OOV to nearly zero, but the BLEU went down, too. We can also see that the adding of phrase table from form as the main table makes the results better overall.

[5]http://www.project-syndicate.org/

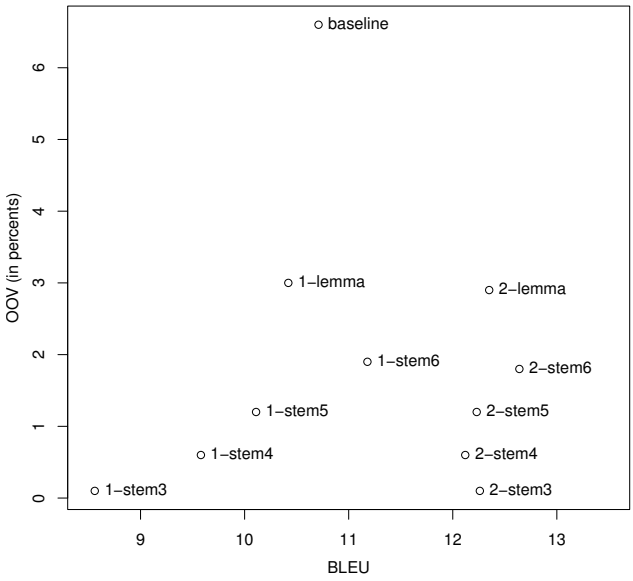

Figure 2: Russian to Czech results

Type	OOV	BLEU
baseline	6.65	10.71 (± 0.56)
1-lemma	3.03	10.42(± 0.57)
1-stem3	0.16	8.56(± 0.53)
1-stem4	0.67	9.58(± 0.50)
1-stem5	1.21	10.11(± 0.57)
1-stem6	1.91	11.18(± 0.55)
2-lemma	2.92	12.35(± 0.58)
2-stem3	0.17	12.26(± 0.64)
2-stem4	0.67	12.12(± 0.58)
2-stem5	1.22	12.23(± 0.59)
2-stem6	1.88	12.64(± 0.60)

Table 3: Results for Russian-to-Czech

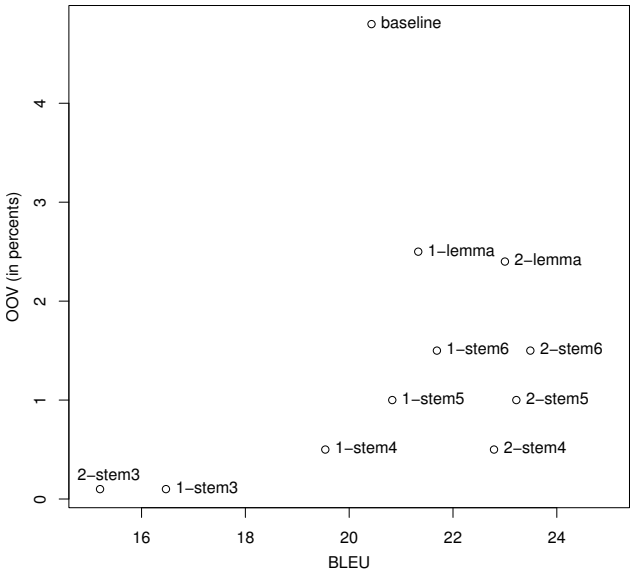

Figure 3: Russian to English results

Type	OOV	BLEU
baseline	4.81	20.43(± 0.71)
1-lemma	2.56	21.33(± 0.73)
1-stem3	0.11	16.47(± 0.60)
1-stem4	0.51	19.54(± 0.69)
1-stem5	1	20.83(± 0.74)
1-stem6	1.53	21.69(± 0.71)
2-lemma	2.45	23(± 0.79)
2-stem3	0.12	15.2(± 0.76)
2-stem4	0.5	22.79(± 0.77)
2-stem5	1	23.22(± 0.79)
2-stem6	1.52	23.49(± 0.79)

Table 4: Results for Russian-to-English

What was surprising to us is that "brute force" stemming can actually achieve better OOV rate **and** BLEU score (although the differences in BLEU score are inside the confidence interval).

We can see that we can both decrease the OOV and increase the BLEU score.

6.2 Russian to English

The results for Russian to English translation are in the Figure 3 and in Table 4 (with the 95% confidence intervals with BLEU).

We can again see the same general trends as in previous example - but it's not that surprising, taking into consideration the fact that the source text is the same. Again, brute-force stemmers work better than lemmatizers. (There is something unusual in 1-stem3 and 2-stem3, where 2-stem3 BLEU is unexpectedly lower, than 1-stem3. Given that the translations from 3 lettered stem are not very truststworthy, we consider this just a strange anomaly.)

What is, however, surprising is, that the "deeper" relationship between Russian and Czech **doesn't help the translation at all**. The BLEU score is higher in Russian to English, the OOV rate too. Again, as we touched upon this issue before, this is probably mainly the fault of the morphological richness of Slavic languages.

Even though, the results were still a little unexpected to us.

A small note: with experience from our past experiments, we think, that if we chose to do the experiments in the other direction - from Russian to English, the results would be less promising, as it is generally translation **into** a morphologically rich language that causes problems.

7 Conclusion

In this paper we have shown two ways to improve the translation quality and lower out-of-vocabulary rate: with the help of lemmatizing and stemming.

We have shown that surprisingly, "brute force" stemming works better for both BLEU and OOV rates, than lemmatization, and that ideal length is six letters.

We have also shown that morphological closeness of two rich languages is not helping us in SMT, if we have the same amount of parallel data.

References

Habash, N.: Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Stroudsburg 2008, PA, USA, 57-60.

Popovic,M., Ney, H.: Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In: Proceedings of 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, May 2004

Oflazer, K.:Statistical Machine Translation into a Morphologically Complex Language. In: Proceedings of CICLing 2008, Haifa, Israel, February 17-23, 2008.

Gispert,A., Marino,J, Crego, J: Improving statistical machine translation by classifying and generalizing inflected verb forms. In: Proceedings of 9th European Conference on Speech Communication and Technology. 2005.

Turchi, M., Ehrmann, M.: Knowledge Expansion of a Statistical Machine Translation System using Morphological Resources. In: Polibits, (43), 37-43, 2011.

Bojar,O., Tamchyna, A.: Forms Wanted: Training SMT on Monolingual Data. In: Proceedings Research Workshop of the Israel Science Foundation University of Haifa, Israel. 2011.

Koehn, P., H. Hoang, A. Birch et al: Moses: Open source toolkit for statistical machine translation. In: Proceeding ACL '07 Proceedings of the 45th Annual Meeting of the ACL, pp. 177-180, ACL, 2007.