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.

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 (0), (0), (0) 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 (0), (0), (0). 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 assumed 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 Statistical Machine Translation setup

Statistical Machine Translation nowadays has become one of the easiest and cheapest paradigms of the MT systems. From the various available tools, we chose to experiment with Moses, an open-source implementation of phrase-based statistical translation system.

2.1 Moses

The Moses toolkit (0) 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.

To establish a baseline for further experiments, we trained translation models for direct translation from Russian to Czech (*ru→cs simple*) and Russian to English (*ru→en simple*), optimizing them on the development set.

In our second experiment, we used the so-called factored models. Factored translation is an extension of the basic translation model, where on both the source and the target side there doesn't have to be just form, but we can enrich the forms with some more information. We will touch on the specific factors later.

2.2 Data

Phrase-based SMT systems need huge amount of parallel data in order to extract dictionaries of phrases and their translations, so called phrase tables. In our work we exploited data from a parallel Czech-English-Russian corpus called UMC (UFAL Multilingual Corpus) with automatic pairwise sentence alignment. The texts were downloaded from from the Project Syndicate[1] page. The data are divided into three sets: training set(train), development set(dev) and test set. The statistics of the data are summarized in the Table 1.

	Languages	Sentences
Language Model	cs	92,233
Translation Model	ru → cs	93395
Translation Model	en → cs	92775
Dev	cs, en, ru	765
Test	cs, en, ru	2000

Table 1: Parallel corpus size.

English	Czech	Russian
jolly elephant	veselý	veselyj slon
	veselé slona	veselogo slona
	veselé slonu	veselomu slonu
	veselé slona	veselogo slona
	veselé slonu	veselom slone
	veselýonem	veselym slonom
	veselýe	

Table 2: Declension of a noun phrase.

As we can see, the number of sentences is slightly different; it is because the sentences are not always aligned one-to-one, but are often ????[2]

3 Out-of-vocabulary words

High out-of-vocabulary rate and mistakes in morphological forms are most typical of translating from and especially to the morphologically rich languages. Almost all works cited in the introduction presented a research on a MT where one language of the translation pair was morphologically rich. Slavic languages are mostly inflecting languages characterizing by free word order and rich inflectional paradigms. The table 2 shows the exposion of word forms in Slavic languages on the example of a noun phrase.

The above example of declension demonstrates the morphological complexity of Czech and Russian. This creates a problem of data sparseness that increase the number of out-of-vocabulary words(forms).

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

[2] TODO: najít přesně!!!

3.1 Statistics of OOV words for simple models

Following is the table that demonstrates the correlation of bleu score and the oov rate for different type of languages. The oov rate was calculated rather in a primitive way - we inspected the translation output for alien characters. The words that contained cyrillic alphabet letters were considered to be 'unknown' within the Czech or English text. And otherwise, the latin characters in the Russian output text signalized in the majority of cases the out-of-vocabulary word. [3]

As we can see from the table, the morphological properties of languages seems to affect the bleu score and the oov rate differently - in a rather predictable way though. In the translation into English the oov rate was minimal. Bleu score is bound to the OOV rate, the more is the bleu score, the less unknown words occur in the translated text. We also tried to see if language type has some impact on the OOV rate, and it did not. The only factor that mattered was the type of data - domain, size and quality. When trained on the corpus UMC with news thematics(100,000 sentences) the OOV rate was rather high.

4 Using morphological analyzers to improve the translation of unknown words

One of the ways to improve the out-of-vocabulary rate is using additional morphological information, the method that was successfully implemented for example by (0), bringing a decrease of a OOV words without introducing more parallel data. First we opted for taggers that are available on-line. Those taggers (Morce for Czech, TreeTagger for Russian and English) assigned each word form with a lemma and a tag. As our main task on the current stage was only to check how much words will be translated properly, we are more interested in increasing the OOV rate than a BLEU score. The latter is not supposed to be that good for the evaluating translation into morphologically rich languages that often have free word order. Still, it will serves the purpose of comparison the translation quality into the same language (we have chosen Czech as a target) under the same conditions(training data).

[3]These numbers are not very precise - words in latin within Russian text can be just terms or proper names(like linux, Java, USA etc.) that can be tolerable in Russian text

Cz: *Informace|informace|NNFP1-------A--- o|o-1|RR6------- pástáké|pástákýIS6-----1A--- jaderné|jadernýIS61A------ programu|program-1|NNIS6-------A---*
En: *The|the|DT visionaries|visionary|NNS would|would|MD have|have|VH gotten|get|VVN nowhere|nowhere|RB*

Figure 1: Facored corpus, tagsets from TreeTagger(En) and Morce(Cs)

In order to train a factored model we tagged and lemmatized the UMC corpus with the help of TreeTagger for English and Russian and Morce morphological tagger for Czech. Each word form is assigned by a lemma and a morphological tag as described below:

Our second experiment using the factored data is of a more complex structure. The word alignment is made on lemmas so that various forms of the same word were aligned, in the contrast to the simple model. We built two phrase tables: first one contained the mapping lemma $\rightarrow$ form + tag, the second one form $\rightarrow$ form +tag. Then we constructed the language model for forms and tags. The results of the experiment in terms of BLEU score and OOV rate are summarized

5 Stemming

Stemming - exploiting a stem(root) of a word is a primitive thus efficient technique to support OOV words guessing. Especially for the agglutinative languages stemming can bring some fruit because each morphemic category is related one-to-one to its surface formCzech and Russian are flective languages, so they combine the morphemes by fusion/flexion, not just putting it one after another. So for instance, if a substantive in Czech has categories number, gender and case, the morphemes presenting those categories will be represented only by one morpheme-ending. As we tried to use the maximum of baseline data, we decided to derive stems from the words without using any additional morphological information like the list of word endlings that are to be eliminated. The technique is primitive - it presents taking the first n characters

translation pair	bleu	OOV
ru → cs	11%	6%
ru → en	15%	8%

Table 3: BLEU score for simple model - baseline.

stem length	BLEU	OOV
6	12.04	1.8%
5	**12.22**	1.1%
4	11.04	0.6%
3	11.99	0.1%

Table 4: BLEU score for models on stems with different length.

lang pair	Simple model		Factored-lemma		Factored-stem	
	BLEU	OOV	BLEU	OOV	BLEU	OOV
ru → cs	11.14	6.41	11.68	2.81	12.22	1.19
en → cs	14.58	4.67	15.49	3.11	15.39	3.47

Table 5: Overall evaluation.

of a word and then selecting the optimal length of a stem that bring the better improvement of a Bleu score and OOV rate. The example of a stemmed text:
En: *the|the gaza|gaza cease|cease -|- fire|fire should|shoul be|be allowed|allow to|to facilitate|facil reconciliation|recon between|betwe fatah|fatah and|and hamas|hamas* The setup of this experiment is the same as the previous - factored, where stems are used instead of lemmas, the results are shown in Table 4.

The alignment on stems that are 3 characters long brought the lowest OOV rate, but we can not trust enough the unknown words that were guessed with this step. The optimal number of characters selected as stems for a translation into a morphologically rich language was 5, so we applied it to other language pairs. We examined the unknown words for the experimental setup stem-5, and it appeared, that it contained either rarely used named entities, less frequent spelling variants, typos, and a minimum of meaningful words.

5.1 Results

In order to see which technique was more efficient for our task we compared all the experiments - simple, factored on lemmas and stems. described above. The results are shown in Table 5.

It became evident, that our techniques to improve the translation quality help especially in the case of MT between morphologically rich languages. The score for English-Czech translation, both simple or factored, was higher than Russian-Czech, but have not gained much improvement when factored models were introduced.

6 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. These models have shown the slightest improvement in terms of BLEU score and a considerable decrease of out-of-vocabulary words especially for the morphologically rich languages. The OOV rate for the translation between Czech and Russian reduced 2 times(lemma model) and 5 times(stem model) against the baseline. The improvement in terms of OOV for English-Czech translation was not significant and the BLEU score has not changed a lot as well.

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, PA, USA, 57-60.

Popovic,M., Hermann, N.: 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

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.