Morphological Neural Pre- and Post-Processing for Slavic Languages

Giorgio Bernardinello
STAR Group
Wiesholz 35, CH-8262 Ramsen
Switzerland
giorgio.bernardinello@star-group.net

Abstract

While developing NMT systems for our customers involving Slavic languages, we encountered certain issues that do not affect Latin or Germanic languages. The most striking of these is the morphological complexity inherent in a remarkable number of unique synthetic forms. For each language combination, the aim is always to find the best balance between the size of the vocabulary, the quality of the translation and the performance of the MT model (both training time and translation time). When working with Slavic idioms, the variety of cases and genders makes this challenge even more difficult and engaging. For Slavic source languages, our solution is to add an extra pre-processing step before the actual translation, in which the inflected word is reduced to its components; naturally, in the opposite direction this requires a symmetrical post-processing technique. Tests have proven high-quality results for Slavic languages, either source or target, confirming this as an effective approach.

1 Challenge

Slavic languages are characterised by an articulated inflectional structure; i.e. cases (synthetic form) are generally used instead of prepositions (analytic form) to express complements.[1] As an example, the Czech table of a regular adjective inflection is made up of 56 cells: 7 cases, 4 genders, 2 numbers. Luckily, because many of them are the same, there are "only" 11 unique variants.

These forms are not as frequent in a corpus: some of them may be used ten times less than others, and this can obviously cause the engine to inconsistently translate what appears to be the same word.

As you can see in Table 1, there are many more Czech forms than English ones, and our engine must be able to handle all of them. What makes this task even more difficult is that the customer's training material is often extremely repetitive, with similar forms repeated many times and others just a few.

To je *pěkná* **kniha**.	This is a *nice* **book**.
To jsou *pěkné* **knihy**.	These are *nice* **books**.
Viděl jsem tě s *pěknou* **knihou**.	I have seen you with a *nice* **book**.

Table 1. Sample of Czech inflections of adjectives and substantives.

2 Aim

When working with standard tokenization, the initial basic conditions required to achieve good MT translations are quality and the amount of training data. There are two typical scenarios:

- Huge, well-formed corpora that need more extensive technical resources for training (GPU, memory, RAM, etc.)
- Smaller data sets, from which it is often not easy to obtain high-quality results

In both cases, we can improve the process by tweaking the tokenization in a way that allows for intelligent handling of inflections. This can lead to better structuring of the engine's vocabulary, resulting in a win-win situation: instead of filling it with many variants of the same word, it can be made smaller and more efficient without sacrificing quality, or it may

[1] Modern Bulgarian and Macedonian are an exception to this rule; noun declension in these languages is actually disappearing.

contain more terms from different contexts without increasing its size. In simple terms, we could obtain substantial benefits if we could separate stems from affixes.

Another important consideration is that our scenario involves final users with little or no knowledge of one of the languages; in this context, reducing Out Of Vocabulary (OOV) words would be a significant goal.

3 A solution between standard tokenization and BPE

With this in mind, we need a tokenizer that works not only on word boundaries, but also in terms of the morphological construction of the token. In this respect, the BPE (Byte-Pair-Encoding) algorithm (Sennrich et al., 2016) may be a valid option, but it is based on the most common sequences of characters and thus it cannot always split words in the way a human would. It is certainly practical in the absence of further grammatical information, but it has already been proven (Ataman et al., 2017) that considering morphological aspects while tokenizing results in higher translation quality.

While observing the inflections in languages such as Czech or Polish, we noticed that the ending may vary depending on the final part of the stem, which means it would be too difficult to manually split the text using a complete list of endings. In addition to this, some of them would be too rare to be learned well by the engine. We therefore supposed that, since a native speaker can implicitly distinguish stems and inflections, a neural model (from now on referred to as a Morpho Model) could be trained to do the same; that is, identify the sequence of letters that can influence the ending and split the word into stem and affix before sending anything to the translation engine. The output tokens from this pre-processing model are the ones that the final translation engine will learn.

This approach differs from pure character-based neural machine translation in that the Morpho Model only needs to parse single complete words rather than translate whole sentences.

Of course, this model is only the core of this pre-processing technology, and can only produce high-quality results as part of a series of steps that guarantee clean input and output data. For example we noticed in the very first phase of tests that irregular forms had to be recognized and handled separately; in fact they represent a relatively small amount of widely used lemmas, with inflections which are hard to be learnt in a general abstract way.

The attempt to find a valid solution that was different from BPE came from the need to have a sort of control over the translation. With the integration of the Morpho Model, as described in the following chapters, we can minimize the risk of unexpected phaenomena, like sub-sets of words considered sequences to be inflected. For our user case it is extremely important to have an output that fulfils the customer's needs regarding not only the general quality of the translation, but also the usage/avoidance of certain forms: therefore we chose to invest resources in a system we can control under almost any aspect.

4 Description of the method

In order to successfully implement this process, it is essential to have a map with a sufficient number of examples and a good description of many morphological categories (for example, it would not be enough to know only the gender of a noun, without its case, number, etc.).

The databases we used to create the maps are free online resources. To have an idea of how big the maps are that we used, we can say that our Russian map has more than two million entries, while the Polish one has more than five million. A reduction of the map's size may be possible by comparing words in the training material for the final NMT engine with the contents of the map. Nevertheless, even words which are not contained in the customer's dictionary may help build a more consistent Morpho Model; in fact it should be trained to build up inflections with their letters, regardless of their meaning or occurrences.

Since we are working with Slavic languages as either the source or the target, the Morpho Model is used in both directions; that is to say, from an inflected word to its corresponding morphological information as well as in the opposite direction. To obtain the expected benefits for the engine's vocabulary, we need to train it using a corpus where all inflections have been reduced. However, we also want to be able to parse the engine's output back to a human-readable language, so the reduction needs to be mapped towards a real word.

Although the two directions have the same logic (from opposite perspectives), they may present distinct challenges during the translation process, once the engine has been trained.

4.1 Slavic source language

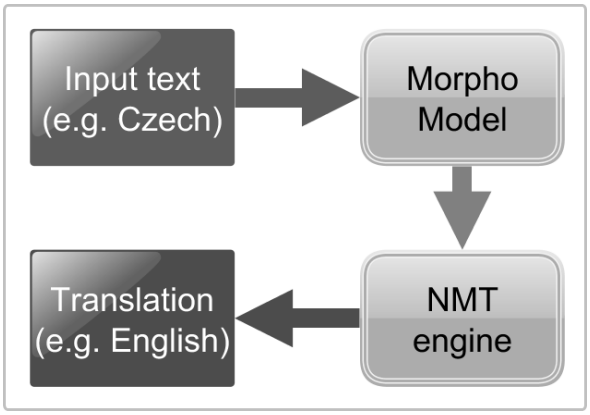

Figure 1. Overview of the process when the Slavic language is the source.

When translating from a Slavic language, the Morpho Model must parse an inflected word into its grammatical information so that the engine has everything it needs to translate properly. Figure 1 shows how the whole workflow should operate; when moving from one step to the next, further text handling may be needed, such as tokenizing or checking the format.

Since the same inflection can be mapped with many definitions (see Table 2), we must ensure that the Morpho Model produces output that can be used by the engine to guarantee a high-quality result; an even more difficult example is that of terms which can belong to two or more different parts of speech, like substantives and verbs or adjectives and verbs. In any case, we should remember that all languages of our experience have ambiguous words which can be understood only with the help of the context and it is one of the NMT engine's tasks to find the correct translation for each of them.

Inflection	Definition
pěkná	*pěkný* nom. f. s.
pěkná	*pěkný* voc. f. s.
pěkná	*pěkný* nom. n. pl.
pěkná	*pěkný* acc. n. pl.
pěkná	*pěkný* voc. n. pl.
pěkné	*pěkný* nom. f. pl.
pěkné	*pěkný* nom. n. s.
pěkné	*pěkný* gen. f. s.
pěkné	*pěkný* dat. f. s.
pěknou	*pěkný* instr. f. s.

Table 2. Sample of Czech adjective mapping - Extract.

4.2 Slavic target language

When translating into a Slavic language, the Morpho Model is employed from the definition to the inflection. In this case, the engine plays a dominant role. In fact, its translation constitutes the input for the Morpho Model, and it must be extremely reliable in order to correctly build the final word. Consequently, particular care is required when selecting the tokens to be sent to the Morpho Model (it works at word level, so it needs **one** stem and several properties to generate **one** inflected form).

There is a risk of creating incorrect or even artificial words at the end of the process, but our tests show that this risk is minimal.

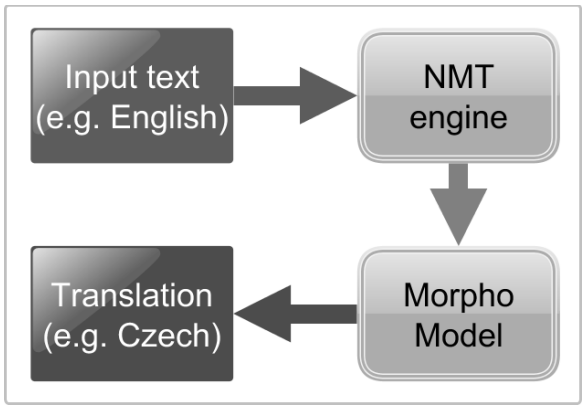

Figure 2. Overview of the process when the Slavic language is target.

4.3 Results

Test results[2] involving only the Morpho Model show that when the Slavic language is the source language, the percentage of perfect matches[3] is around 80%. This value is perfectly respectable, considering that the remaining non-perfect matches may fall into one of three categories:

- Alternative definition
- Correct stem with a mistake in the morphological properties
- Mistake in the stem

While the first two cases may cause a degree of confusion and lower the final BLEU evaluation, only the third one actually represents a disturbing factor when used as input for the incoming translation engine.

In any case, we can observe quite astonishing results in the opposite direction (i.e. Slavic as the

[2] The test set was made up of 10,000 non-trained words.

[3] We consider a perfect match only when the Morpho Model's output corresponds exactly to the definition of the test inflection (i.e. stem and all grammatical classes)

target language), where the perfect match rate is over 90% for Russian, and even 97% for Polish. The difference up to 100% represents cases in which the user may receive a spurious word that does not really exist, but such an outcome can be avoided or at least strongly reduced with a simple spellchecker, for example.

As regards the evaluation of the whole translation process, results appear not so easy to evaluate. If we take Polish as an example (but the other languages had similar behaviour) we see that pure BLEU values with Morpho Model are in both directions lower than the BPE.[4] Since the number of translations with BLEU below 0.2 was much bigger in the Morpho case than in the BPE, we took a selection of 150 of them and let them be analysed by translators who did not know about our study. We expected to find that recurrent phaenomena showed some kind of inconsistency in one or more steps of our process, but we were told that actually the translation with the Morpho Model often had a better level of comprehensibility. As a final test we let the translators make manual comparisons of BPE and Morpho translations in our web application, with particular focus on the correctness of inflected forms. After this confirmation we decided to use this new technology in production; in fact, we usually proceed only after the approval of a translator or at least a native speaker, especially for such cases when the automatic evaluation doesn't show a significant advantage for a particular case.

5 Possible drawbacks

Some reservations have been expressed concerning the time spent on a single translation, as each word has to be handled by the Morpho Model in addition to the time required by the normal NMT engine. In this respect, it is important to note that the Morpho Model is much faster than a conventional engine due to the consistency of material and the low settings required for its training (word vector and RNN far below 100).

Another criticism may be the risk of having less control over the translation, since we are using two neural models instead of one. However, thanks to other pre-/post-processing steps, we can reduce the possibility of unexpected results, as a last resort leaving the source word unchanged to prevent the model from creating spurious words.

In any case, as a company, we need to consider any MT solution in a practical way: the worst possible output for our average user is an OOV. Thus, reduction of OOVs, coupled with more consistent quality when translating the same lemma, is a major objective. In most cases, a translation containing an OOV is completely incomprehensible, while one containing the correct stem and an incorrect ending is sufficient to justify continuing with the work.

Furthermore, an error rate of 3%, as the one we had for Polish, is probably not far from the human one, especially considering that not everyone among our target users has high linguistic skills.

You might assume that a technique based on morphology requires a deep knowledge of the languages involved. To some extent that is true, in that some linguistic knowledge can be useful (detecting mistakes, faster development, problem solving). However, the grammatical aspects under consideration are not so specialised as to require an expert; at least no more than those involved in conventional training.

6 Conclusions

The accuracy of the result is strictly dependent on the quality of the map used to train the Morpho Model. Since a good amount of well-formed linguistic data is required to create the map, it is important to handle this correctly. For example, knowing that the customer generally avoids the use of certain verb forms can lead to a reduction in the size of the map, resulting in a simpler task for both the model and the engine. Moreover, the size of the map is a factor that can influence quality and performance. For customers with a small variety of subjects, the map can be reduced based on the words the engine can translate.

7 Further challenges

A potential next step for this logic could be to use it in a scenario where both the source and target languages are Slavic. The result could be a greater reduction in vocabulary; however since Slavic languages are quite a homogeneous family, the difference may not be appreciable compared to conventional training.

Another interesting field of application might be for languages with non-concatenative morphology, such as Arabic, where words are in-

[4] de-pl BPE: 0.580, de-pl Morpho: 0.571. pl-de BPE: 0.587, pl-de Morpho: 0.569.

flected with transfixes rather than prefixes or suffixes. The incentive in this case relates not only to the technical challenge, but also to the potential future business opportunities offered by the Middle East and North Africa.

References

Duygu Ataman, Matteo Negri, Marco Turchi, Marcello Federico. 2017. *Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English*, The Prague Bulletin of Mathematical Linguistics No. 108, 2017:331-342.

Rico Sennrich, Barry Haddow, Alexandra Birch. 2016. *Neural Machine Translation of Rare Words with Subword Units*, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, August 7-12:1715-1725.