# Two-Step MT: Predicting Target Morphology

*Franck Burlot, Elena Knyazeva, Thomas Lavergne, François Yvon*

LIMSI, CNRS, Univ. Paris-Sud, Université Paris Saclay, 91 403 Orsay, France

firstname.lastname@limsi.fr

## Abstract

This paper describes a two-step machine translation system that addresses the issue of translating into a morphologically rich language (English to Czech), by performing separately the translation and the generation of target morphology. The first step consists in translating from English into a normalized version of Czech, where some morphological information has been removed. The second step retrieves this information and re-inflects the normalized output, turning it into fully inflected Czech. We introduce different setups for the second step and evaluate the quality of their predictions over different MT systems trained on different amounts of parallel and monolingual data and report ways to adapt to different data sizes, which improves the translation in low-resource conditions, as well as when large training data is available.

## 1. Introduction

When translating into a morphologically rich language, statistical machine translation (SMT) systems generally perform poorly, generating several incorrect word forms that show errors in agreement within a noun phrase, that encode the wrong grammatical function of the word in the sentence, or that simply convey the wrong meaning from the source. Such errors come from an important assumption upon which current SMT systems rely: translation is based on a source-target mapping of one or several word forms that are memorized in the model disregarding any broader context.

This assumption can be problematic when both languages involved lack symmetry in their linguistic systems. A more analytical language such as English encodes grammatical information using distinct words (prepositions, negative particles, auxiliary verbs), while a language showing synthetical tendencies such as Czech encodes the same information in word inflection. Moreover, inflection encodes more explicitly the grammatical function of a word in the sentence, while such information is in English encoded in its position in the sentence (e.g. the object is on the right side of the verb).

At the word level, extracting one-to-one word mappings from the parallel data based on alignment links has two main consequences:

- One source word can translate into several target words, leaving difficult choices to be made by the system. This is typically the case of the English adjective that is invariable and translates into potentially forty-two Czech word forms (varying according to seven cases, three genders and two numbers), not considering lexical ambiguities. When the information is extracted, the context accountable for an inflection is lost.

- Such a variety of word forms that may need to be generated on the target side face important sparsity issues. At test time, one may need to produce on output a word form that has not been seen in the training data, which is challenging and may require extra processing or enriched representations (e.g. factored models). It may also occur that the word form is contained in the model, but its probability is not well estimated because of its low frequency in the training data. Mapping source to target phrases prevents enough generalization in such a situation.

This paper focuses on word representations in SMT for translation from English to Czech. In order to minimize sparsity on the target side, we proceed to a normalization of the target language by removing grammatical information, such as case, gender and number. This results in a representation of Czech that makes it more symmetric to English. We expect from such a representation an improvement in the translation quality, since the wide variety of choices the translation system has to make is minimized. In this configuration, the SMT system translates from English into a normalized version of Czech. A second translation step is thus necessary and consists in re-inflecting the previously obtained normalized language, in order to output fully inflected Czech. Re-inflexion is therefore performed independently of the translation process. It can take advantage of the full context in the output sentence and is also less dependent on the training data, since it may generate word forms that have not been seen in the parallel corpus.

After a description of related work, we will present our system that is built on a translation system that translates from English into a normalized version of Czech (Section 3.1). This output is re-inflected in a second step to turn normalized Czech into a fully inflected language (Section 3.2). Having described our experimental setup (Section 4), we then show that improvement gets lower as the quantity of training data grows (Section 5). Finally, we show a way to better leverage high quantity of data (Section 6).

## 2. Related work

This paper fits into previous work on two-step machine translation addressing morphology as a post-processing step. Minkov [1] and Toutanova et al. [2] translate from English into Russian (and Arabic) stems, which are used to generate full paradigms, then disambiguated using a classifier. In a comparable way, Chahuneau et al. [3] augment the translation model with synthetic phrases obtained by re-inflecting target stems.

Bojar et al. [4, 5, 6] use two SMT systems: the first one translates from English into Czech lemmas decorated with source-side information and the second one performs a monotone translation into fully inflected Czech. Jawaid and Bojar [7] use in the first step a hierarchical system that outputs a lattice presenting different word orders. The second system then selects the word order that allows for the best morphological predictions.

Fraser et al. [8] represent German words as lemmas followed by a sequence of tags and introduce a linguistically motivated selection of these in order to translate from English. The second step consists in predicting the tags that have been previously removed, using a different CRF for each morphological attribute to predict. Finally, word forms are produced via a look-up in a morphological dictionary. El Kholy and Habash [9, 10] propose a similar approach for Arabic. Weller et al. [11] introduce verbal subcategorization frames enabling the CRFs to make better predictions, and Weller-Di Marco et al. [12] handle the prediction of both prepositions and morphological features by building synthetic phrase tables.

The present work is close to the original idea of Fraser et al. [8] and follows unsuccessful attemps to model target morphology. Marie et al. [13] proposed a similar normalization scheme for translating from English to Russian. Allauzen et al. [14] introduced a hidden CRF model for English into Russian and Romanian aimed at directly predicting the word form, after having generated the full paradigm of the word translated at the previous step.

## 3. Morphological re-inflection

Our initial assumption is that translation could be easier if the MT model was relieved from having to make hard decisions about morphology. Two-step MT is a way to process morphology apart from the translation process. The first step of the proposed scenario consists in translating from English into normalized Czech. For this purpose, the target side of the parallel and monolingual data have to be pre-processed.

### 3.1. Normalization of the Czech data

Popovic and Ney [15], Goldwater and McClosky [16] and Durgar El-Kahlout and Yvon [17] show the benefits of normalizing the morphologically rich language (here Czech or German) on the source side when translating into English. Such a normalization consists in grouping different word forms sharing the same lemma into a common class, by removing one or many attributes (e.g. gender, number, case) that are considered as redundant with respect to English. This pre-processing has the effect of reducing the source vocabulary, making both languages more symmetric, and has a positive impact on the translation quality.

When translating in the reverse direction, these ideas hold, but one needs in addition to make sure that the attribute that was removed at normalization step is recoverable from the monolingual context in the SMT output. Indeed, the models we propose for re-inflexion do not have access to source side information (see Section 3.2). Therefore, whenever an attribute is redundant with respect to English but is needed for the prediction of other attributes in surrounding words, it needs to be kept.

In our pre-processing, a word is represented as a lemma and a tag sequence, which we obtained using Morphodita [18]. Normalizing such a word simply means removing one or many tags from the sequence. We propose a deterministic schema for each part of speech. The following attributes are preserved:

- **Nouns**: *lemma, PoS, gender and number*. Number is an attribute that is common to English, and gender is an intrinsic part of Czech nouns, meaning that it may serve to disambiguate two identical lemmas that have a different lexical meaning. Moreover, as head of a noun phrase, the word propagates gender to its dependents. Case is systematically removed and we consider that it should be predictable from the monolingual context[1].

- **Adjectives**: *lemma, PoS, negation, degree of comparison*. Since the adjective is invariable in English, we remove gender and number, but keep both negation, which has a lexical value, and the degree of comparison, which is also marked in English.

- **Numerals**: *lemma, PoS*. English numbers only have one form.

- **Pronouns**: *lemma, POS, subPoS, person, gender, number, number[psor], gender[psor]*. Only case is removed from pronouns. Gender and number of both possessor (*[psor]*) and possessed are hard to predict and are generally not well handled in SMT. We leave these attributes and are aware that their prediction would require a special attention that is beyond the scope of this paper [19]. Person is also kept and we expect it to be a useful predictor of nominative case when a pronoun agrees with a verb in the context.

[1] Some contexts make case prediction hard and this attribute should probably sometimes be conveyed from the source, as in the normalized output *jím ruka+Plur* (*eat hand+Plur*). Instrumental case needs to be predicted for the noun, in order to obtain *jím rukama* (I eat with my hands). If the case tag is lost in this output, the classifier used for re-inflexion may ignore the semantic aspect of the clause and consider the noun as a direct object, generating the semantically less likely sentence with accusative case *jím ruce* (I eat hands).

- **Prepositions**: *word form, POS, case*. Here, we keep the word form, since some prepositions have different forms depending on the right side context, e.g. ***s*** *tebou* (with you) - ***se*** *mnou* (with me). The SMT system handles well this phenomenon. Case is kept, since some prepositions can be followed by different cases and we expect this attribute to propagate through the entire preposition phrase in the output. This choice implies that verb constructions are expected to be handled by the SMT system that is considered to be able to distinguish *jít v + Accusative* (go to) and *být v + Locative* (be in).

- **Verb**: The lemma and the whole tag sequence are kept. Verbs are not normalized, and we follow the same principle as Fraser et al. [8] that this PoS be considered an anchorage point of the output. The full tag sequence should help distinguish the object from the subject with which it should agree in person, gender, and number.

- **Adverb, interjection, conjunction, particle**: Word forms are kept, since they are all invariable.

In this setup, only three attributes can be removed: gender, number and case. This constraint makes the tag prediction task easier, since only sequences of three tags need to be predicted (as opposed to sequences of fifteen tags according to the Morphodita tagset[2]). Finally, it allows us to train one different classifier for each attribute (see next section).

### 3.2. Output re-inflection

The machine translation system outputs a text in a normalized language that needs to be re-inflected. At this step, we have lemmas with a fixed sequence of attributes, some of them having missing values (gender, number and/or case). The task is therefore similar to any sequence labeling problem where the goal is to predict the right value for each empty attribute. When the full tag sequence has been predicted, a dictionary is used to recover the word form corresponding to the predictions. We report experiments with three different ways to re-inflect the normalized Czech output.

In order to train the two supervised models we used data from the Universal Dependencies Treebank project[3]. We used the Czech and Czech-CAC corpora covering general domain and transcripts of spoken language for a total of $2M$ words, from which $170k$ where held out for development.

#### *3.2.1. Language Model (LM)*

Each word of the normalized Czech output is re-inflected using an n-gram language model. First, the normalized word is used to query the Morphodita word generator [18]. As a result, we obtain one or several inflected Czech word forms. For instance, if nouns have been normalized by removing the case attribute, the morphological generator will output the forms corresponding to each of the seven Czech cases. We end up with a sentence full of ambiguities at different positions. This new sentence is represented as a lattice that is rescored with a 4-gram language model trained on fully inflected Czech sentences.

#### *3.2.2. Cascade of Conditional Random Fields (CRF)*

The first supervised model we considered is a CRF [20] that predicts three morphological attributes using the Wapiti toolkit [21]. A joint prediction of all these attributes allows us to better account for the dependencies between them, but such a model can be challenging to train due to the potentially high number of attribute combinations to consider.

A total of 180 different combinations of attributes are observed in our corpus, which are reachable for a CRF model but would require more training data than available to obtain reasonable performance. To overcome this problem, we train a cascade of CRF models, in which the first three models predict a single morphological attribute. That output is used to feed the final joint classifier. The final joint model is therefore only responsible for discovering the dependencies between the attributes and for correcting the predictions made by the previous models.

All four models are trained using 1- to 3-gram word features in an 11-word window as well as 1- to 4-gram features concerning the known morphological information in the same window. Additionally, 1- to 4-gram features on the output of each previous models are used. The models are trained in a specific order: gender, number and case are successively trained, then the joint model is learnt. The same order is followed for decoding.

To extract the features based on previous models, a full decoding of the training data by these models is necessary. To get unbiased predictions, a 10-fold cross-validation is done for the training of the first three models.

The three morphological attributes should be predicted only in words for which they have been removed during the normalization process. Gender, for example, has to be predicted for adjectives but not for nouns. The models are trained to predict the attributes for every relevant words even if they are already known, but during inference the Viterbi decoder is forced to only consider paths that go through the already known attributes. Such a forced decoding allows the model to use this knowledge to make its predictions.

#### *3.2.3. Greedy sequence labeller (Greedy)*

As an alternative to the CRF cascade model, a greedy model for sequence labeling was used. The predictions of each attribute (gender, number, case) are performed separately, one after the other, using an SVM multi-class classifier from LIBLINEAR library by Fan et al. [22]. During both the training and the decoding process, gender is predicted first, then

[2] https://ufal.mff.cuni.cz/pdt2.0

[3] http://universaldependencies.org

number, then case, in a left-to-right order for each attribute. The feature set is the same as for the CRF model except that it has the possibility of using the 1- to 4-gram features on morphological information predicted for the same attribute.

Another difference with the CRF model is that training examples are extracted only where a prediction should be made. This reduces the number of training examples and helps the model to focus on learning the real task. As for any greedy model, the error propagation problem is crucial here. To deal with it, we apply the SEARN strategy of Daumé et al. [23]. Several iterations of training are performed to alleviate the impact of previously made errors.

More precisely, the search space is generated directly during the learning/decoding process. The states are source-side lemmas and morphological information (given by the MT system), as well as all previously predicted morphological tags. During the first learning iteration, these are extracted from the reference as though decoding has been performed with no mistakes so far.

During the following iterations, we gradually add past decoding errors: for the $k$-th iteration, the probability of using a previous prediction (possibly erroneous) in the future is equal to $k/10$, otherwise the reference tag is used ($1 \leq k \leq 10$). Thus for the last iteration the search space is as close as possible to the one of the decoder. The action set is composed of all possible combinations of morphological tags. Then, all couples (state, action) produced in this way are used to train the classifier.

#### *3.2.4. Final disambiguation*

The latter two systems predict tags that are used to query the Morphodita word generator. At this step, ambiguities often remain, for which we have to make a final decision. Indeed, the Czech language may have different inflections that express mainly stylistic variations. For instance, *děkuji* is more formal than *děkuju* (both meaning *thank you*). These remaining ambiguities are solved using a unigram model, which simply selects the form that has the highest frequency in the training data.[4] We assume that the stylistic level present in the data can be captured in this simple way.

## 4. Experimental setup

The SMT systems introduced in the following sections are trained with Moses [24] and Ncode [25], and optimized with Mira. 4-gram language models are trained with removed singletons using KenLM [26].

For this task, we used the data provided at both WMT 2016[5] and IWSLT 2016.[6] All systems are optimized on a concatenation of English-to-Czech TED test sets 2010 and 2011, and tested over a concatenation of TED test sets 2012 and 2013. All Czech data is tokenized and truecased using scripts from the Moses toolkit. The English-side tokenization and truecasing relies on in-house text processing tools [27].

Our previous attempts at re-inflexion for machine translation suggest that improvement is expected mostly when low amount of either parallel or monolingual data is available. This is the case for under-resourced languages but also to some extent for domain specific translation. For this reason we choose to test our systems on TED talks (transcribed talks) to create a situation close to low resources, since less parallel data is available. On the other hand, a large quantity of monolingual out-of-domain data is at our disposal to study the effect of corpus size in this context.

We consider Czech as representative of morphologically rich languages, with a complex nominal and adjectival inflection. Many such languages are not provided with a lot of parallel data, such as Bulgarian or Ukrainian. The system described in this paper should improve the translation into these languages as well. Using Czech for our experiments allows us to actually explore the impact of data size on the re-inflection quality, which would not be possible with genuine low-resourced languages.

## 5. Impact of data size

In this section, we will explore the impact of re-inflection on translation quality in setups involving different amounts of parallel and monolingual data.

### 5.1. Parallel data

We first explore the parallel data size dimension as this is generally the main limitation in the training of SMT systems for low-resourced languages. We have trained translation systems with increasing amounts of parallel data starting with a small one containing only the first 10k sentences of the TED corpus. The next system corresponds to that same full corpus (117k sentences) which is next increased to 242k sentences by adding the QED corpus,[7] then to 885k sentences after adding Europarl. The final larger system is obtained by appending the news-commentary corpus (1M sentences).

The results of these different systems are shown in Table 1. We observe that Ncode systems have significantly higher results for two-step setups. The CRF and Greedy models provide significant improvements over the baselines (direct English to Czech translation). While there is improvement for all corpus sizes, we notice that as the amount of parallel data grows, the effectiveness of the re-inflexion decreases. This is expected as more parallel data means that the baseline systems have knowledge of more word forms and better statistics on them. With enough parallel data, it is expected that a direct translation system will reach the same performance as the two-step system.

Conversely, the word form selection using language models always deteriorates the baselines. These models are

[4] Our attempt to solve these ambiguities using a 4-gram language model did not give any improvement over the simple unigram model.

[5] http://www.statmt.org/wmt16

[6] http://workshop2016.iwslt.org

[7] http://alt.qcri.org/resources/qedcorpus/

Table 1: BLEU scores for Moses and Ncode systems over direct translations (en2cs) and two-step translations (en2cx2cs). Language models are trained over the target side of the parallel data. As the amount of parallel data grows, the effect of re-inflection gets lower.

| Data | Moses | | | | Ncode | | | |
|---|---|---|---|---|---|---|---|---|
| | en2cs | LM | CRF | Greedy | en2cs | LM | CRF | Greedy |
| 10k | 10.06 | 9.96 (-0.10) | 11.60 (+1.54) | **11.64** (+1.58) | 10.62 | 10.44 (-0.18) | 12.13 (+1.51) | **12.28** (+1.56) |
| 117k | 15.70 | 15.20 (-0.50) | 16.70 (+1.00) | **16.78** (+1.08) | 15.77 | 15.52 (-0.25) | 17.17 (+1.40) | **17.32** (+1.55) |
| 242k | 15.96 | 15.32 (-0.64) | 16.72 (+0.76) | **16.90** (+0.94) | 16.06 | 15.68 (-0.38) | 17.17 (+1.11) | **17.32** (+1.26) |
| 885k | 16.75 | 16.45 (-0.30) | 17.74 (+0.99) | **17.94** (+1.19) | 16.94 | 16.67 (-0.27) | 18.04 (+1.10) | **18.25** (+1.29) |
| 1M | 17.14 | 16.51 (-0.63) | 17.64 (+0.50) | **17.88** (+0.74) | 17.15 | 16.64 (-0.51) | 17.99 (+0.84) | **18.13** (+0.98) |

Table 2: BLEU scores for Moses and Ncode systems over direct translations (en2cs) and two-step translations (en2cx2cs). The parallel data used adds up to 885k sentences. As the amount of monolingual data grows, the effect of re-inflection gets lower.

| Data | Moses | | | | Ncode | | | |
|---|---|---|---|---|---|---|---|---|
| | en2cs | LM | CRF | Greedy | en2cs | LM | CRF | Greedy |
| 5M | 18.01 | 18.05 (+0.04) | 18.73 (+0.72) | **18.84** (+0.83) | 17.91 | 17.82 (-0.09) | 18.69 (+0.78) | **18.87** (+0.96) |
| 10M | 18.58 | 18.42 (-0.16) | 18.87 (+0.29) | **19.05** (+0.47) | 18.38 | 18.34 (-0.04) | 18.88 (+0.50) | **19.11** (+0.72) |
| 50M | 18.97 | 19.19 (+0.22) | 19.02 (+0.05) | **19.22** (+0.25) | 18.96 | 19.45 (+0.49) | 19.26 (+0.30) | **19.53** (+0.57) |
| 90M | 19.34 | 19.40 (+0.06) | 19.26 (-0.08) | **19.51** (+0.17) | 19.59 | 19.54 (+0.05) | 19.52 (-0.07) | **19.79** (+0.20) |
| 200M | 20.71 | **20.81** (+0.10) | 19.75 (-0.96) | 20.02 (-0.69) | 21.13 | **21.45** (+0.32) | 20.62 (-0.51) | 20.91 (-0.22) |

trained on the target side of the parallel data, which has inevitable consequences. A priori, they are quite small (trained on up to one million sentences) and one can not expect them to make proper estimates over such a large vocabulary created by the rich morphology of Czech. Furthermore, their vocabulary is strictly the same as in the translation model. Therefore, the language model systematically favors words that were seen in the parallel data, making the generation of the paradigm of a normalized Czech word (second step) almost pointless.[8] The CRF and Greedy models take advantage of the fact that they are not restricted to a closed vocabulary.

In a low resource context, re-inflection helps when few parallel data is available. On the other hand, monolingual data is easier and cheaper to obtain and can therefore be used in large amounts.

### 5.2. Monolingual data

In this section, we will observe how growing monolingual data impact the improvement given by our models. All the following systems use the set of bilingual text of the 885k sentences from the previous section.

On the monolingual side, we use increasing corpus size from 5M to 200M sentences. These corpora include the target side of the parallel data as well as news data, subtitles, and a filtered part of the common-crawl corpus. [9]

[8]The CRF re-inflection of the 117k system generates 1109 types (1503 tokens) that were not seen in the parallel data, while the LM re-inflection generates only 817 such types (1173 tokens) that were treated by the model as OOVs.

[9]This last corpus was filtered by applying the Moore-Lewis method with XenC [28].

Results are shown in Table 2. We again notice that the improvement given by the classifiers is higher when the translation is trained with Ncode. For small monolingual data size, the system with re-inflexion improves over the baseline as in the previous section but, as the data grow, the improvement vanishes and even starts to be detrimental for the biggest system. As opposed to the bilingual study, there is enough monolingual data to reach the point where it is possible to build a language model big enough to capture the richness of the fully inflected language efficiently. Such a model is able to make better predictions than the CRF as it can capture 4-grams dependencies between the inflected words where the CRF can only capture 2-gram dependencies. The greedy model, which also capture 4-gram dependencies, shows better results than the CRF but is still outperformed by the baseline for the larger MT system, probably due to its limited amount of training data.

The language models for re-inflexion, that performed poorly on the small data setup, start to be efficient when enough monolingual data is available. With 50M sentences, it improves over the baseline and finally outperforms our best supervised system on the biggest data size. With such amounts of data, it is interesting to note that a two-step system, where the translation is done on normalized Czech with the LM used for re-inflexion, performs better than the baseline with the same LM as it can output words never seen in the parallel data. The 200M system with LM re-inflection now generates 1138 types and 1485 tokens not seen in parallel data, which makes it more similar to the CRF re-inflection that outputs 1148 types (1493 tokens), just a few more.

Table 3: BLEU scores for direct translations (en2cs) and two-step translations (en2cx2cs), re-inflecting n-best hypothesis from Ncode with different data sets (# parallel sentences / # monolingual sentences).

| **Model** | **10k/10k** | **117k/117k** | **242k/242k** | **885k/885k** | **1M/1M** |
|---|---|---|---|---|---|
| en2cs | 10.62 | 15.77 | 16.06 | 16.94 | 17.15 |
| LM | 10.42 (-0.20) | 15.47 (-0.30) | 15.81 (-0.25) | 16.64 (-0.30) | 16.72 (-0.43) |
| CRF | 12.39 (+1.77) | 17.31 (+1.54) | 17.17 (+1.11) | 18.24 (+1.30) | 18.23 (+1.08) |
| + CRF k-best | **12.47** (+1.85) | 17.22 (+1.45) | 17.37 (+1.31) | **18.55** (+1.61) | **18.62** (+1.47) |
| Greedy | 12.39 (+1.77) | **17.49** (+1.72) | **17.65** (+1.59) | 18.31 (+1.37) | 18.55 (+1.40) |
| **Model** | **885k/5M** | **885k/10M** | **885k/50M** | **885k/90M** | **885k/200M** |
| en2cs | 17.91 | 18.38 | 18.96 | 19.59 | 21.13 |
| LM | 17.91 (+0.00) | 18.30 (-0.08) | 19.20 (+0.24) | 19.81 (+0.22) | 21.29 (+0.16) |
| CRF | 18.81 (+0.90) | 19.23 (+0.85) | 19.50 (+0.54) | 20.02 (+0.43) | 21.07 (-0.06) |
| + CRF k-best | 19.17 (+1.26) | 19.35 (+0.97) | **19.90** (+0.94) | **20.24** (+0.65) | **21.40** (+0.27) |
| Greedy | **19.23** (+1.32) | **19.54** (+1.16) | 19.84 (+0.88) | 20.23 (+0.64) | 21.35 (+0.22) |

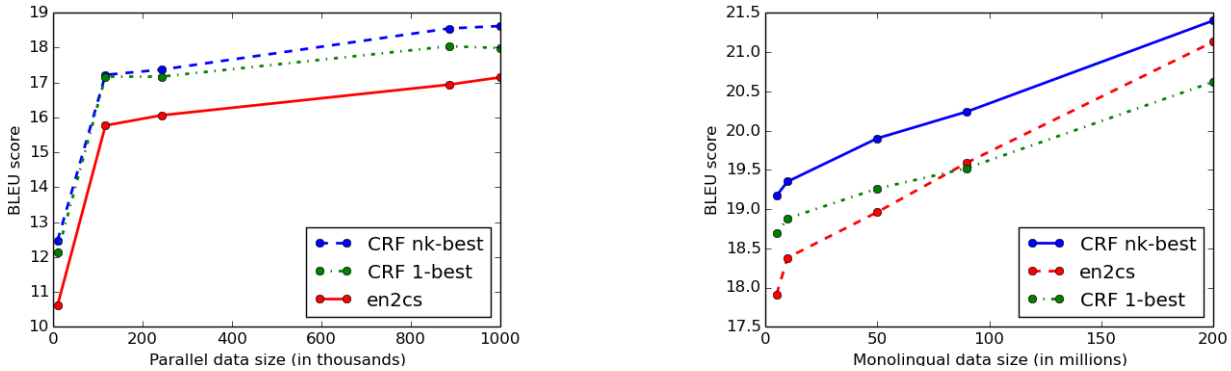


Figure 1: Scores for CRF re-inflexion of 1-best and nk-best hypothesis over increasing parallel and monolingual data size.

## 6. Taking advantage of larger data

Two-step MT is not limited to improvements in low resource conditions. With the systems we have described so far, the prediction of morphology is done within the scope of a fixed set of words in a fixed order, since we only re-inflect the best hypothesis of the MT system. There actually are situations where we could make a better prediction for a word on the condition that this word itself or its position changes in the sentence. We allow such a variation in the output by considering the n-best hypothesis of the MT system ($n = 300$).

We introduce results obtained by re-inflecting the n-best hypothesis from the MT system (now Ncode only) in Table 3. The re-inflected n-best hypothesis are rescored using MIRA and a language model trained over the monolingual data used for the MT system, except now in naturally inflected Czech. We also have an additional setup where the CRF outputs its k-best predictions ($k = 5$), leading to the rescoring of nk-best translation hypothesis. Using for this purpose a character-based neural language model provides only slight improvements over an n-gram language model (see Burlot et al. [29]).

We see that both classifiers still show a significant improvement over large systems and start decreasing only after 200M. Figure 1 shows the scores of the CRF for the re-inflection of 1-best and nk-best hypothesis. While using up to 242k parallel sentences for the MT training, the re-inflection of Ncode n-best hypothesis shows no significant improvement in BLEU over the 1-best re-inflection. Furthermore, n-best re-inflection performs better as the amount of data grows. Indeed, with larger language models, the space of the n-best list provides more useful alternatives, of which the morphology prediction models can take advantage. An example of this is shown in Table 4, where the 1-best hypothesis provided an ungrammatical verb frame with a future tense constructed on the auxiliary verb *být* and a perfective verb, leading to a bad prediction of the dative form for the pronoun. Exploring the n-best hypothesis for re-inflection allowed the model to make the right prediction (accusative) according to a correct verbal frame (with the imperfective verb).

It seems that the deterioration given by the classifiers with bigger MT systems is mainly due to a reduction of the translation quality improvement in the first step. Figure 2 shows the improvement over the baseline obtained with the normalized output (the BLEU score is computed over the normalized reference translation). We understand this BLEU score as a simulation of an ideal situation where

Table 4: Better morphological predictions with nk-best hypothesis (885k/90M system).

| | |
|---|---|
| Source | I will bypass you |
| CRF 1-best | budu **ti** obejít |
| | will you-Dative bypass-Perfective |
| CRF nk-best | budu **tě** obcházet |
| | will you-Accusative bypass-Imperfective |

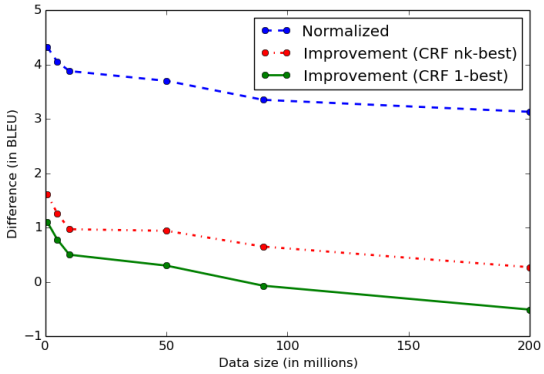


Figure 2: Difference in BLEU score between baseline (cs) and both normalized (cx) and re-inflected outputs (cx2cs) with growing monolingual data.

we are able to predict the same inflections as the reference. Thus, this shows the maximum improvement we can theoretically achieve with the output re-inflection. Here also, as the amount of data grows, improvement decreases: the score of the normalized output goes from a potential improvement of over 4 BLEU points with small data, to barely 3 with larger data. We note that the actual improvement given by the systems is highly correlated to this maximum achievable BLEU score. In a larger data setup, the low improvement is therefore mainly due to the quality of the first step (translation), that seems to take less advantage of Czech normalization. Therefore, making the translation step easier by normalizing the target side helps less when very large data is available.

## 7. Conclusions

We have presented a complete study on the effect of different amounts of parallel and monolingual data on a translation system with re-inflexion. The results suggest that re-inflexion is more effective when corpora are a scarce resource as with under-resourced languages or domain specific translation. In our experiments we found that, in such a context, even when vast amounts of monolingual data is available, a two-steps MT is still the best choice if we switch from a supervised morphological prediction to an LM when needed.

We have also studied the impact of using the n-best lists from the MT system and showed that, when enough monolingual data is available for an effective rescoring, they improve the overall performance of the system making relevant the use of a re-inflexion system in bigger configurations.

These results explain some previous unsuccessful attempts. Weller et al. [11] and Marie et al. [13] obtain small to no improvements on translation into French and German with similar setups. In such cases, a large amount of parallel and monolingual data were used, making classifier predictions useless. Fraser et al. [8] unsuccessfully used the n-best list of the MT system, but on a small system. On such scale, the LM used for the rescoring of the re-inflected n-best is too small to be efficient.

Future works will include the exploration of alternative sequence predictors like the greedy one, which can better capture long range dependencies as our experiments demonstrate, as well as ways to integrate knowledge of the source sentence to improve the predictions. We also plan to investigate automatic ways to perform the normalization instead of our manual selection of the attributes to keep.

## 8. Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work has been partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

## 9. References

[1] H. S. Einat Minkov, Kristina Toutanova, "Generating complex morphology for machine translation," in *Proceedings of ACL*. ACL, June 2007.

[2] K. Toutanova, H. Suzuki, and A. Ruopp, "Applying morphology generation models to machine translation," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: ACL, June 2008, pp. 514–522.

[3] V. Chahuneau, E. Schlinger, N. A. Smith, and C. Dyer, "Translating into morphologically rich languages with synthetic phrases," in *EMNLP*. ACL, 2013, pp. 1677–1687.

[4] O. Bojar, "English-to-Czech factored machine translation," in *Proc. of the 2nd WMT*, Prague, Czech Republic, 2007, pp. 232–239.

[5] O. Bojar and K. Kos, "2010 failures in English-Czech phrase-based MT," in *Proc. WMT and MetricsMATR*, ser. WMT'10, Stroudsburg, PA, USA, 2010, pp. 60–66.

[6] O. Bojar, B. Jawaid, and A. Kamran, "Probes in a taxonomy of factored phrase-based models," in *Proc. WMT*, ser. WMT '12, Stroudsburg, PA, USA, 2012, pp. 253–260.

[7] B. Jawaid and O. Bojar, "Two-step machine translation with lattices," in *Proc. LREC*. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[8] A. Fraser, M. Weller, A. Cahill, and F. Cap, "Modeling inflection and word-formation in SMT," in *Proc. EACL*, Avignon, France, 2012, pp. 664–674.

[9] A. El Kholy and N. Habash, "Translate, predict or generate: Modeling rich morphology in statistical machine translation," in *Proc. EAMT*, Trento, Italy, 2012, pp. 27–34.

[10] ——, "Rich morphology generation using statistical machine translation," in *Proceedings of the Seventh International Natural Language Generation Conference*, ser. INLG'12. Stroudsburg, PA, USA: ACL, 2012, pp. 90–94.

[11] M. Weller, A. M. Fraser, and S. S. im Walde, "Using subcategorization knowledge to improve case prediction for translation to german." in *ACL (1)*. ACL, 2013, pp. 593–603.

[12] M. Weller-Di Marco, A. Fraser, and S. Schulte im Walde, "Modeling complement types in phrase-based smt," in *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: ACL, August 2016, pp. 43–53.

[13] B. Marie, A. Allauzen, F. Burlot, Q.-K. Do, J. Ive, e. knyazeva, M. Labeau, T. Lavergne, K. Löser, N. Pécheux, and F. Yvon, "LIMSI@WMT'15: Translation task," in *Proc. WMT*, Lisbon, Portugal, 2015, pp. 145–151.

[14] A. Allauzen, L. Aufrant, F. Burlot, O. Lacroix, E. Knyazeva, T. Lavergne, G. Wisniewski, and F. Yvon, "LIMSI@WMT16: Machine Translation of News," in *Proc. WMT*, Berlin, Germany, August 2016, pp. 239–245.

[15] M. Popovic and H. Ney, "Towards the use of word stems and suffixes for statistical machine translation," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*, 2004.

[16] S. Goldwater and D. McClosky, "Improving statistical MT through morphological analysis," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05, 2005, pp. 676–683.

[17] I. Durgar El-Kahlout and F. Yvon, "The pay-offs of preprocessing for German-English Statistical Machine Translation," in *Proc. IWSLT*, M. Federico, I. Lane, M. Paul, and F. Yvon, Eds., 2010, pp. 251–258.

[18] J. Straková, M. Straka, and J. Hajič, "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition," in *Proc. ACL: System Demos*, Baltimore, Maryland, 2014, pp. 13–18.

[19] L. Guillou and C. Hardmeier, "Protest: A test suite for evaluating pronouns in machine translation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), may 2016.

[20] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2001, pp. 282–289.

[21] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Proc. ACL*. ACL, July 2010, pp. 504–513.

[22] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, June 2008.

[23] H. D. III, J. Langford, and D. Marcu, "Search-based structured prediction," *Machine Learning Journal*, 2009.

[24] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proc. ACL:Systems Demos*, Prague, Czech Republic, 2007.

[25] J. M. Crego, F. Yvon, and J. B. Mariño, "N-code: an open-source Bilingual N-gram SMT Toolkit," *Prague Bulletin of Mathematical Linguistics*, vol. 96, pp. 49–58, 2011.

[26] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proc. WMT*, Edinburgh, Scotland, 2011, pp. 187–197.

[27] D. Déchelotte, G. Adda, A. Allauzen, O. Galibert, J.-L. Gauvain, H. Maynard, and F. Yvon, "LIMSI's statistical translation systems for WMT'08," in *Proceedings of NAACL-HTL Statistical Machine Translation Workshop*, Columbus, Ohio, 2008.

[28] A. Rousseau, "XenC: An open-source tool for data selection in natural language processing," *The Prague Bulletin of Mathematical Linguistics*, no. 100, pp. 73–82, 2013.

[29] F. Burlot, M. Labeau, E. Knyazeva, T. Lavergne, A. Allauzen, and F. Yvon, "LIMSI@IWSLT'16: MT Track," in *Proc. IWSLT*, ser. IWSLT'16, Seattle, USA, 2016, to appear.