Word Representations in Factored Neural Machine Translation

Franck Burlot[*]
LIMSI, CNRS, Université Paris Saclay

Mercedes García-Martínez[*]
LIUM, University of Le Mans

Loïc Barrault
LIUM, University of Le Mans

Fethi Bougares
LIUM, University of Le Mans

François Yvon
LIMSI, CNRS, Université Paris Saclay

Abstract

Translation into a morphologically rich language requires a large output vocabulary to model various morphological phenomena, which is a challenge for neural machine translation architectures. To address this issue, the present paper investigates the impact of having two output factors with a system able to generate separately two distinct representations of the target words. Within this framework, we investigate several word representations that correspond to different distributions of morpho-syntactic information across both factors. We report experiments for translation from English into two morphologically rich languages, Czech and Latvian, and show the importance of explicitly modeling target morphology.

1 Introduction

Open vocabularies remain a challenge for Neural Machine Translation (NMT) (Cho et al., 2014; Bahdanau et al., 2015), both for linguistic and computational reasons. From a linguistic standpoint, morphological variation and lexical productivity cause word forms unseen in training to occur in source texts, which may also require to generate novel target word forms. Using very large input/output vocabularies partially mitigates these issues, yet may cause serious instability (when computing embeddings of rare or unseen words) and complexity issues (when dealing with large softmax layers).

Several proposals have been put forward to address these problems, which are particularly harmful when one language is a morphologically rich language (MRL), exhibiting larger token/type ratio than is observed for English. One strategy is to improve NMT's internal procedures: for instance by using a structured output layer (Mnih and Hinton, 2008) or by altering the training or decoding criteria (Jean et al., 2015). An alternative approach is to work with representations designed to remove some variations via source-side or target-side normalization procedures; or more radically to consider character-based representations (Ling et al., 2015; Luong and Manning, 2016; Costa-jussà and Fonollosa, 2016), which are however much more costly to train, and make long distance dependencies even longer.

None has however been as successful as the recent proposal of Sennrich et al. (2016b) which seems to achieve a right balance between a limited vocabulary size and an ability to translate a fully open vocabulary. In a nutshell, this approach decomposes source and target tokens into smaller units of variable length (using what is now termed as a "Byte Pair Encoding" or BPE in short): this means that (a) all source tokens can be represented as a sequence of such units, which crucially are all seen in training; (b) all possible target words can also be generated; (c) the size of the output layer can be set to remain within tractable limits; (d) most frequent words are kept as BPE units, which preserves the locality of many dependencies.

In this work, we consider possible ways to extend this approach by also supplying target-side linguistic information in order to help the system generate correct target word forms. Our proposal relies on two distinct components (a) linguistically or data-driven normalization procedures manipulating various source and target word segmentations, as well as eg. multiple factors on the target side (see § 4), and (b) a neural architecture equipped with a dual output layer to predict the target in two simpler tasks generating the lexi-

[*]Both authors have contributed equally to this work.

Proceedings of the Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 20–31
Copenhagen, Denmark, September 711, 2017.

cal unit and the morphological information (§ 3). These components are assessed separately and in conjunction using translation from English into two MRLs: Czech and Latvian. Our experiments show improvement over a strong (Denkowski and Neubig, 2017) BPE-to-BPE baseline, incorporating ensemble of models and backtranslated data (§ 5). Overall, they suggest that BPE representations, which loosely simulates concatenative morphological processes, is complementary to feature-based morphological representations.

2 Related Work

Translating from and into MRLs has recently attracted some attention from the research community, as these languages compound a number of difficulties for automatic translation, such as the need to analyze or generate word forms unseen in training, or to handle variation in word order.

To mitigate the unknown word problem, a first approach consists in translating into target *stems* (Minkov et al., 2007; Toutanova et al., 2008); the right form is then selected from the full paradigms in a second step using a classifier. Target words may also be represented as lemmas complemented with side information. Bojar (2007); Bojar and Kos (2010); Bojar et al. (2012) use such a representation for two statistical MT 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.

Fraser et al. (2012) propose a target morphology normalization for German words represented as lemmas followed by a sequence of morphological tags and introduce a linguistically motivated selection of these when translating from English. The selection step consists in predicting the tags that have been removed during normalization, using a specific Conditional Random Field (CRF) model for each morphological attribute to predict. Finally, word forms are produced via look-up in a morphological dictionary. This approach is extended by Weller et al. (2013), who takes verbal subcategorization frames into account, thus enabling the CRFs to make better predictions. Note that Burlot et al. (2016) and El Kholy and Habash (2012b,a) propose related approaches respectively for translating into Czech and Arabic.

Factored word representations have also been considered in neural language models (Niehues et al., 2016; Alexandrescu and Kirchhoff, 2006; Wu et al., 2012), and more recently in a neural machine translation architecture as input features (Sennrich and Haddow, 2016) and in the output by separating the lemma and morphological factors (García-Martínez et al., 2016). One contribution of the current paper is the investigation of new variants of the latter architecture. There have been other attempts with dual training objectives in NMT. In (Chen et al., 2016), a *guided alignment training* using topic information of the sentence as a second objective helps the decoder to improve the translation. Multi-task and multilingual learning in NMT have also been considered in several papers (Luong et al., 2015; Dong et al., 2015; Firat et al., 2016), where training batches have to carefully balance tasks and language pairs. In contrast to these approaches, our factored NMT (FNMT) system produces several outputs *simultaneously*.

3 Model Architectures

The baseline NMT system used in this paper is an implementation of a standard NMT model with attention mechanism (Bahdanau et al., 2015). It consists of a sequence to sequence encoder-decoder of two recurrent neural networks (RNN), one used by the encoder and the other by the decoder. This architecture integrates a bidirectional RNN encoder (see bottom left part with green background of Figure 1). Each input sentence word x_i ($i \in 1 \dots N$ with N the source sequence length) is encoded into an annotation a_i by concatenating the hidden states of a forward and a backward RNN. Each annotation $a_1 \dots a_N$ thus represents the whole sentence with a focus on the word(s) being processed. The decoder is based on a conditional gated recurrent unit (GRU) (Firat and Cho, 2016) made of two GRUs interleaved with the attention mechanism. The attention mechanism computes a context vector C_j as a convex combination of annotation vectors, where the weights of each annotation are computed locally using a feed-forward network. The decoder RNN takes as input the embedding of the previous output word in the first GRU, the context vector C_j in the second GRU and its hidden state. The softmax output layer is connected to the network through a non-linear layer which takes as input the embedding of the previous output word as well as the context vector and the output of the decoder from the second GRU (both adapted with a linear trans-

formation, respectively, L_C and L_R). Finally, the output probabilities for each word in the target vocabulary are computed with a softmax. The word with the highest probability is the translation output at each time step. The encoder and the decoder are trained jointly to maximize the conditional probability of the reference translation.

The Factored NMT system of García-Martínez et al. (2016) is an extension of the standard NMT architecture that allows the system *to generate several output symbols at the same time*, as presented in Figure 1.

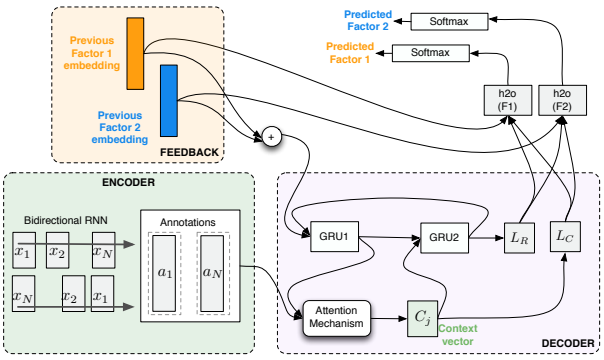

Figure 1: Factored NMT system.

The encoder and the attention mechanism of the Factored NMT are the same as the standard NMT model. However, the decoder has been modified to produce multiple outputs. The two outputs are constrained to have the same length. The decoder feedback is also modified to use information from the multiple output streams. The concatenation of the embeddings of the pair of generated symbols is used to feed the decoder's cGRU at each timestep.

Two types of FNMT models have been used for this work. Their architecture differ after the generation of the decoder state. The first model contains a single hidden-to-output (*h2o*) layer which is used by the two separate softmax. This layer uses the context vector, the decoder's hidden state and the concatenation of the embeddings of the previous generated tokens. The second model is one contribution of the current work. As shown in Figure 1), it contains two separated *h2o* layers. They are similar to the *h2o* layer in the first model except that instead of using the concatenation of the embeddings of the previously generated factors, each *h2o* layer receives only the embedding of the factor it is generating. The two separated *h2o* layers allow the system to have more weights specialized for each output.

4 Word Representations

This paper focuses on the question of word representations, which we understand not only in terms of word segmentation, but also as the quantity of morpho-syntactic information encoded in a word. We introduce three representations varying in the quantity of grammatical information they contain:

- **fully inflected words**: this is a baseline setup where all the lexical and grammatical information is encoded in a single factor.
- **normalized words**: only a well chosen subset of morphological features is kept in the first factor; the second factor corresponds to the Part of Speech (PoS).
- **lemmas**: the output splits the lexical content of the word (first factor: lemma) and its grammatical content (second factor: PoS).

These differences are illustrated in Table 1.

4.1 Normalizing Word Forms

Translating from English into a MRL is made difficult by linguistic divergences, as English lacks many of the morphological contrasts that exist in the MRL. Normalization is needed to reduce the morphological variability on the MRL side so as to limit the number of types in the target, and to mitigate sparsity issues. This strategy is used for instance by Burlot et al. (2016) who remove the case mark from Czech nouns, which is not predictable from their English counterpart(s).

Normalization is usually performed using handcrafted rules and requires expert knowledge for each language pair. In this paper, normalized words are obtained with an automatic data-driven method[1] introduced in (Burlot and Yvon, 2017b).

In a nutshell, this method performs a clustering of the MRL vocabulary by grouping together words that tend to share the same translation(s) in English. This translational similarity is based on the conditional entropy of lexical translation models estimated, for each MRL word form, using automatic word alignments. The clustering procedure merges two words whenever the resulting cluster does not increase the conditional entropy, which ensures a minimal loss of information during the whole process.

[1]The source code is available at github.com/franckbrl/bilingual_morph_normalizer

The actual normalization algorithm is delexicalized and operates at the level of PoS. Each word is represented as a lemma, a coarse PoS and a sequence of morphological tags (e.g. *kočka+Noun+Sing+Accusative*). Translational similarities are computed on such words and are combined to provide a PoS-level similarity between two tag sequences. Successive merge operations group into one cluster different such tag sequences. As a result of this procedure, we represent words as a lemma and a cluster identificator (ID) taking the form of a coarse PoS and an arbitrary integer, such as *kočka+Noun+7* in Table 1. In this example, the cluster ID *Noun+7* stands for a set of fine-grained PoS, such as {*Sing+Nominative, Sing+Accusative,* ...}.

This representation introduces a direct correspondence between the first and the second factor in our architecture, since the former (the cluster ID) constraints the set of possible values of the latter (the fine-grained PoS), which is notably used in our constrained decoding procedure (§ 5.4).

4.2 Word Representation Setup

The example of Table 1 shows that words are also varying along a second dimension: in addition to considering unsegmented lexical units (be it fully inflected words, normalized words or lemmas), we also investigate the impact of a segmentation of these units using BPE (Sennrich et al., 2016b).

In this scenario, BPE segmentation is performed on fully inflected words and lemmas. For its application to normalized words, the cluster ID was considered as a minimal unit that cannot be segmented (just like any other character), in order to avoid segmentations like *kočka+No- un+7*. For these setups, the PoS information (second factor) is replicated for all subparts of a word.

We finally use an alternative representation with normalized words to which BPE segmentation is applied and cluster IDs are systematically split from the lemma. Whenever the FNMT system predicts a lemma in the first factor, it is forced to predict a null PoS in the second factor. On the other hand, when a split cluster ID is predicted, the second factor should output an actual PoS. This specific treatment of the second factor is expected to give the system a better ability to map a word to a compatible PoS, thus avoiding, for instance, the prediction of a verbal PoS for the Czech noun *kočka* (cat).

These different word representations imply a progressive reduction of the target vocabulary. We computed the vocabulary size of Czech on the parallel data used to train the systems (§ 5.1) over unsegmented words. We thus have 2.1M fully inflected words, 1.9M normalized words, 1.5M normalized words with split clusters (lemmas and clusters), and 1.4M lemmas.

5 Experiments

We introduce here the experimental setup for all the reported systems translating from English into Czech and Latvian.

5.1 Data and Preprocessing

Our experimental setting follows the guidelines of the WMT'17[2] news translation task. The preprocessing of English data relies on in-house tools (Déchelotte et al., 2008). All the Czech data were tokenized and truecased the Moses toolkit (Koehn et al., 2007). PoS-tagging was performed with Morphodita (Straková et al., 2014). The pre-processing of Latvian was provided by Tilde.[3] Latvian PoS-tags were obtained with the LU MII Tagger (Paikens et al., 2013).

For English-to-Czech, the parallel data used consisted in nearly 20M sentences from a subset of WMT data relevant to the news domain: News-commentary, Europarl and specific categories of the Czeng corpus (news, paraweb, EU, fiction). Newstest-2015 was used for validation and the systems are tested on Newstest-2016 and 2017. The normalization of the Czech data was trained on the parallel data used to train the MT systems, except Czeng fiction and paraweb subcorpora, which amounts to over 10M sentences.

A part of these systems was also trained on synthetic parallel data (Sennrich et al., 2016a) (see § 6). The Czech monolingual corpus News-2016 was backtranslated to English using the single best system provided by the University of Edinburgh from WMT'16.[4] In order to prevent learning from being too biased towards the synthetic source of this set, we used initial bitext parallel data as well. We added five copies[5] of News-commentary and

[2] www.statmt.org/wmt17

[3] www.tilde.com

[4] http://data.statmt.org/rsennrich/wmt16_systems/

[5] Adding multiple copies of the same corpus into the training set can be seen as a coarse way to weight different corpora and favor in-domain bibtext.

	fully infl.	norm. words		lemmas	
	Single factor	factor 1	factor 2	factor 1	factor 2
plain	kočky	kočka+N+7	N+Pl+Nom	kočka	N+Pl+Nom
BPE	ko- čky	ko- čka+N+7	N+Pl+Nom N+Pl+Nom	ko- čka	N+Pl+Nom N+Pl+Nom
+ split cls		ko- čka- N+7	null null N+Pl+Nom		

Table 1: Multiple representations for the Czech word *kočky* (cats). *N* stands for noun, *Pl* for plural and *Nom* for nominative case.

the news subcorpus from Czeng, as well as 5M sentences from the Czeng EU corpus randomly selected after running modified Moore-Lewis filtering with XenC (Rousseau, 2013).

The English-to-Latvian systems used all the parallel data provided at WMT'17. The DCEP corpus was filtered with the Microsoft sentence aligner[6] and using modified Moore-Lewis. We kept the best 1M sentences, which led to a total of almost 2M parallel sentences. The systems were validated on 2k sentences held out from the LETA corpus and we report results on Newsdev-2017 and newstest-2017. The normalization of Latvian data was trained on the same parallel sentences used to train the MT systems.

Training was carried out for a part of these systems on synthetic parallel data. We used a back-translation of the monolingual corpora news-2015 and 2016 provided by the University of Edinburgh (Moses system). To these corpora were added 10 copies of the LETA corpus, as well as 2 copies of Europarl and Rapid.

Bilingual BPE models for each language pair and system setup were learned on the bitext parallel data. 90k merge operations were performed to obtain the final vocabularies. For (F)NMT models, the vocabulary size of the second factors is only 1.5k for Czech and 376 for Latvian. The number of parameters in (F)NMT systems increases around 2.5% for Czech and 7% in Latvian.

5.2 System Setup

Only sentences with a maximum length of 50 were kept in the training data, except for the setup where cluster IDs were split in normalized words. In this case, we set the maximum length to 100. For the training of all models, we used NMTPY, a Python toolkit based on Theano (Caglayan et al., 2017) and available as free software[7]. We used the standard NMT system on fully inflected words and the FNMT architecture described in § 3 on all other word representations.

All systems (F)NMT systems have an embedding dimension of 512 and hidden states of dimension 1024 for both the encoder and the decoder. Dropout is enabled on source embeddings, context vector, as well as output layers. When training starts, all parameters are initialized with Xavier (Glorot and Bengio, 2010). In order to slightly speed up the training on the actual parallel data, the learning rate was set to 0.0004, patience to 30 with validation every 20k updates. On the synthetic data, we finally set the learning rate to 0.0001 and performed validation every 5k updates. These systems were tuned with Adam optimizer (Kingma and Ba, 2014) and have been training for approximately 1 month.

5.3 Reinflection

The factored systems predict at each time step a lexical unit and a PoS-tag, which requires a non-trivial additional step producing sentences in a fully inflected language. We refer to this process as reinflection.

Given a lexical unit and a PoS-tag, word forms are retrieved with a dictionary look-up. In the context of MRL, deterministic mappings from a lemma and a PoS to a form are very rare. Instead, the dictionary often contains several word forms corresponding to the same lexical unit and morphological analysis.

A first way to solve this ambiguity is to simply compute unigram frequencies of each word form, which was done over all the monolingual data available at WMT'17 for both Czech and Latvian. During a dictionary look-up, ambiguities can then be solved by taking the most frequent word form. The downside of this procedure is that it ignores important information given by the target monolingual context. For instance, the Czech preposition *s* (with) will have different forms according to the right-side context: ***s** tebou* (with you), but ***se** mnou* (with me). A solution is to let an inflected-

[6] https://www.microsoft.com/en-us/download/details.aspx?id=52608

[7] https://github.com/lium-lst/nmtpy

word-based system select the correct word form from the dictionary. To this end, k-best hypotheses from the dictionary are generated. Given a sentence containing lemmas and PoS, we perform a beam search going through each word and keeping at each step the k-best reinflection hypotheses according to the unigram model mentioned above.

For Czech reinflection, we used the Morphodita generator (Straková et al., 2014). Since we had no such tool for Latvian, all monolingual data available at WMT'17 were automatically tagged using the LU MII Tagger (Paikens et al., 2013) and we gathered the result in a look-up table. As one could expect, we obtained a large table (nearly 2.5M forms) in which we observed a lot of noise.

5.4 Constrained Decoding

The factored system described in § 3 outputs a lexical unit and a PoS-tag at each time step. A peculiarity of this system is that the predictions of both factors are independent. There is only a weak dependency due to the fact that both share the same decoder state and context vector. As a consequence, the best hypothesis for the first factor can well be incompatible with the best hypothesis for the second factor, and the risks of such mismatches only get worse when top-n hypotheses are considered, as in beam search.

Our constrained decoding procedure aims at enforcing a strong consistency between factors. Each word in the target vocabulary is first associated with a specific set of PoS-tags. The decoding procedure is modified as follows: for each candidate target word, we only retain the compatible PoS tags, and select the top-n hypotheses to be kept in the beam from this filtered list. This constraint ensures that the beam search does not evaluate incompatible pairs of factors. (e.g. the PoS *Preposition* and the word *cat*).

With a dictionary, creating such a mapping is trivial for full lemmas, but less obvious in the case of BPE units. Since the latter can be generated from different words having different grammatical classes, the size of the set of possible PoS can grow quickly. For normalized words, things are much easier and do not even require a dictionary, as the mapping between cluster IDs and compatible PoS is learnt during the normalization process (see § 4.1). Thus constrained decoding was only performed for (a) unsegmented lemmas, and (b) unsegmented and segmented normalized words.

6 Automatic Evaluation

Results are reported using the following automatic metrics: BLEU (Papineni et al., 2002), BEER (Stanojević and Sima'an, 2014) which tunes a large number of features to maximize the human ranking correlation at sentence level and CharacTER (Wang et al., 2016), a character-level version of TER which has shown a high correlation with human rankings (Bojar et al., 2016). Each score on fully inflected word systems is averaged from two independent runs (for both single and ensembled models).

6.1 Experiments with Bitext

The results using the bitext provided at the WMT'17 the evaluation campaign are presented in Table 2 for English-to-Czech [8] and in Table 3 for English-to-Latvian.

We can observe that using the constrained decoding consistently improves the results, except when using split clusters. In this last case, the system is forced to predict a PoS in the second factor whenever it has generated a cluster ID in the first factor. Since there is a reduced quantity of such cluster IDs, the model has no difficulty to learn the constraints by itself and therefore to map a cluster ID exclusively to a specific subset of PoS. In the Latvian lemma setup, we observe that the improvement using constrained decoding is lower than for Czech (see Table 3), which is probably due to the quality of the noisy look-up table we have created for Latvian (see § 5.1). Note that we have no such dependency on the lexical resources at decoding time for the normalized word setups, where improvements are comparable across both language pairs.

The systems using BPE tokens significantly outperform word-level systems, which confirms the analysis of Sennrich et al. (2016b). The results show that BPE units are even more efficient when applied to normalized words, providing significant improvements over segmented inflected words of 1.79 and 1.85 BLEU points for Czech, and 0.78 and 1.06 for Latvian.

The lemma representation was tested with the two FNMT models presented in § 3, one model using a single hidden-to-output layer (*single h2o layer*) and the other model using two separated hidden-to-output layers (*separated h2o layers*).

[8] At decoding time, Czech systems performed better with a beam size of 2, which was used to provide these results.

	Newstest-2016			Newstest-2017		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
word-to-word						
fully inflected w.	15.74	47.29	74.79	12.76	44.81	78.90
factored norm						
sep. h2o layers	16.63	49.78	68.02	13.70	47.13	72.81
+ constrained dec.	17.71	50.38	66.94	14.88	47.81	71.44
factored lemmas						
single h2o layer	16.73	50.50	65.51	14.09	48.15	69.85
+ constrained dec.	17.42	50.94	64.95	14.93	48.76	69.26
sep. h2o layers	16.54	50.12	66.35	13.89	47.78	70.63
+ constrained dec.	17.56	50.73	65.48	14.66	48.26	69.96
BPE-to-BPE						
fully inflected w.	18.24	52.29	60.05	15.08	49.54	65.38
factored norm						
sep. h2o layers	18.59	53.01	59.95	15.89	50.49	66.75
+ constrained dec.	20.03	53.96	58.90	16.93	51.14	64.13
split clusters	19.74	53.90	59.95	16.31	50.73	64.49
+ constrained dec.	19.71	53.96	59.85	16.38	50.83	64.35
factored lemmas						
single h2o layer	17.30	51.82	61.19	14.19	48.98	66.28
sep. h2o layers	17.34	52.22	60.62	14.73	49.61	65.34

Table 2: Scores for English-to-Czech systems trained on official bitext data

We observe mixed results, here: the system with the *single h2o* layer has slightly better results for the word-to-word systems, but the BPE-to-BPE factored lemma system obtains better performance with the *separated h2o layers* architecture. For that reason, we decided to only use the *separated h2o layers* architecture for the next set of experiments involving synthetic data which is the aim of the next section.

6.1.1 Experiments with Selected Bitext and Synthetic Data

Table 4 and 5 show the results of using selected parts of bitext and synthetic parallel data (see section 5.1) for both language pairs. Each model trained with a selection of bitext and synthetic data was initialized with the parameters of its counterpart trained on bitext. The BPE vocabulary used was the same as in the model used for initialization, which led the systems to generate unknown words. In our experiments, we forced the decoder to avoid unknown token generation.

By using synthetic data, we are able to obtain a large improvement for all systems, which is in line with (Sennrich et al., 2016a). We notice that the contrasts present in the previous section between the various word representations are less clear now. The baseline system (first two rows) is the system which benefits the most from the additional data with +5.7 and +6.9 BLEU for Czech and Latvian. The performance of factored systems has also increased, but to a lesser extent, leading to slightly worse results compared to the baseline system. This situation changes when the reinflected hypotheses are rescored. We are then able to surpass the baseline system with normalized words.

The two language pairs react differently to k-best hypotheses rescoring (*+k-best rescored* in the tables). For Czech, this has nearly no impact on translation quality according to the metrics, whereas it provides an important improvement in Latvian: +2.03 and +0.84 BLEU in the split cluster setup. Note that this specific setup gives the best score we could achieve on newsdev-2017, without n-best rescoring or model ensembling. We interpret this situation as a result of the difference in quality observed for the Czech and Latvian dictionaries used for reinflection. Indeed, since Morphodita contains exclusively useful Czech reinflection candidates, a simple unigram model is sufficient to select the right word forms, making the generation of 10-best reinflection hypotheses useless.[9] On the other hand, the hypotheses returned by the look-up table we have used to generate Latvian word forms were noisy and required a rescoring from an MT system based on fully inflected words.[10] We obtained the best results for

[9] Our experiments with 50-best and 100-best reinflections did not show any improvement.

[10] We assume that the word form generation at this step requires information from the monolingual context only, and could be modeled with a simple target language model, although this needs to be confirmed empirically.

	Newsdev-2017			Newstest-2017		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
words-to-words						
fully inflected w.	15.15	48.18	76.97	10.61	43.44	85.67
factored norm						
sep. h2o layers	14.91	50.56	69.49	10.42	45.94	78.83
+ constrained dec.	15.57	50.78	69.65	11.38	46.28	78.93
factored lemmas						
single h2o layer	13.96	49.53	68.36	9.68	45.24	77.07
+ constrained dec.	14.02	49.48	69.97	9.94	45.21	78.11
sep. h2o layers	13.92	49.93	68.45	9.71	45.10	77.51
+ constrained dec.	14.38	49.74	70.04	10.07	45.26	78.08
BPEs-to-BPEs						
fully inflected w.	16.22	51.63	64.44	11.29	47.02	71.95
factored norm						
sep. h2o layers	15.69	52.35	64.14	10.94	47.80	73.51
+ constrained dec.	16.81	52.72	64.02	12.16	48.25	72.93
split clusters	16.99	52.95	64.65	12.35	48.64	72.40
+ constrained dec.	17.00	52.96	64.61	12.35	48.65	72.32
factored lemmas						
single h2o layer	14.45	50.86	67.14	10.45	46.36	72.25
sep. h2o layers	14.39	50.72	66.05	10.69	46.44	72.96

Table 3: Scores for English-to-Latvian systems trained on official bitext.

	Newstest-2016			Newstest-2017		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
fully inflected w.	23.94	57.30	52.77	20.00	54.45	58.40
+ ensemble	24.34	57.51	52.48	20.16	54.62	58.22
factored norm						
sep. h2o layers	22.26	56.49	53.43	18.74	53.76	59.18
+ constrained dec.	23.02	56.76	53.29	19.34	54.03	58.67
split clusters	23.37	57.44	52.66	19.77	54.58	58.44
+ constrained dec.	23.39	57.43	52.71	19.76	54.59	58.51
+ k-best rescored	23.43	57.45	52.64	19.79	54.60	58.49
+ n-best rescored	24.19	57.88	52.19	20.56	54.99	57.96
+ ensemble	24.55	58.00	51.97	20.68	55.08	57.93
factored lemmas						
sep. h2o layers	22.30	56.63	53.46	19.34	54.16	58.76
+ k-best rescored	22.35	56.60	53.49	19.36	54.17	58.71
+ n-best rescored	23.39	57.25	52.73	19.83	54.57	58.35
+ ensemble	24.05	57.59	52.27	20.22	54.80	57.89

Table 4: Scores for English-to-Czech systems (BPE-to-BPE) trained on selected bitext and synthetic parallel data.

this Latvian setup by generating the 100-best reinflection hypotheses, which provides less dependency on the quality of the dictionary and relies more on the knowledge learned by a word-form-aware system. Despite the fact that such a rescoring procedure is costly in terms of computational time, we observe that it can be a helpful solution when no resources of quality are available.

Czech n-best reinflection, as opposed to k-best, turned out to be efficient, bringing the lemma-based systems to the level of the baselines and even above for the normalized word setups. Whereas it does not improve with Latvian normalized words, we observe a positive impact on the lemma-based systems. We assume that rescoring the n-best list is a way to rely on an inflected-word-based system to make important decisions related to translation, as opposed to the much simpler monolingual process of reinflection mentioned above. Latvian split-cluster models seem to have nothing to learn from such systems.

Factored norm performs best among all the presented models, showing consistent BLEU improvements over the baselines of 0.25 and 0.56 for Czech, and 0.57 and 0.89 for Latvian. We finally notice that ensembling two models slightly reduces those contrasts, and lemma-based systems are the ones that benefit the most from model ensembling. Conclusions are not easy to draw, since across the different setups, the level of indepen-

	Newsdev-2017			Newstest-2017		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
fully inflected w.	22.05	57.34	53.32	14.84	51.78	63.08
+ ensemble	22.41	57.78	52.67	15.12	52.11	62.64
factored norm						
sep. h2o layers	18.81	55.65	56.07	13.57	50.94	64.24
+ constrained dec.	20.05	56.14	56.13	14.44	51.26	63.60
split clusters	20.85	56.77	54.13	14.50	51.84	63.04
+ constrained dec.	20.86	56.80	54.02	14.57	51.87	62.96
+ k-best rescored	22.89	57.88	52.77	15.41	52.39	62.40
+ n-best rescored	22.62	57.43	53.66	15.73	52.77	61.78
+ ensemble	22.69	57.61	52.91	16.04	52.99	61.41
factored lemmas						
sep. h2o layers	18.93	56.01	54.36	13.98	51.26	63.9
+ k-best rescored	20.56	56.94	53.42	14.80	51.78	63.19
+ n-best rescored	21.59	57.62	52.83	15.31	52.34	62.64
+ ensemble	21.90	57.83	52.38	15.35	52.31	62.46

Table 5: Scores for English-to-Latvian systems (BPEs-to-BPEs) trained on selected bitext and synthetic parallel data.

dence of the two ensembled models is suspected to be quite different. [11]

It is important to note that automatic metrics may not do justice to the lexical and grammatical choices made by the factored systems. In an attempt to focus on the grammaticality of the FNMT systems, we conducted a qualitative analysis of the outputs.

7 Qualitative Evaluation

7.1 Attention in Factored Systems

In a factored NMT setup, the attention mechanism distributes weights across all positions in the input sentence in order to make two predictions, one for each factor, which is an important difference from single-objective NMT. An illustration of the impact of this difference is shown in Figure 2 for the ensembles of two English-to-Czech models introduced in § 6.

In this sentence, the system based on fully inflected words (translation on the top) erroneously predicts the verbal present tense in *nevyhýbá* (does not avoid). We can see that the target subword unit *nevy@@* is rather strongly linked to the source *didn't*, which allowed the system to correctly predict negative polarity. On the other hand, the ending of the verb *á* is not linked by attention to this same source word, from which the morphological feature of past should have been conveyed. We observe in (a) that the lemma-based system attention aligns the target position to both the source auxiliary *didn't* and the lexical verb's first subword unit *shir@@*, which enables the successful prediction of the right lemma and morphology, i.e. negation (N) and past (R). The normalized word based system in (b) shows an even more explicit modelization of this morphological phenomenon. While the lemma *nevyhýbat@@* is strongly related to the same English segment *shir@@*, it is only slightly linked to the English auxiliary. *didn't* is instead clearly associated to the cluster ID *V+20* that gathers negative past tense PoS-tags, enabling the right prediction in the second factor. In this last setup, the system has to deal, at each time step in the output sentence, with either a lexical phenomenon or a grammatical one.

Target-side grammatical phenomena being more explicitly modeled in factored NMT, it is generally easier for the attention mechanism to spot an English grammatical word (auxiliary, preposition, negative particle, etc.), which enables a better prediction in the second factor output. We assume that this peculiarity ensures a better source-to-target grammatical adequacy.

7.2 Measuring Morphological Consistency

We provide here an attempt to understand more systematically whether an *a priori* intuition of factored NMT systems is verified. The intuition is that dividing the task of translating a sentence into two easier joint tasks, namely the prediction of a lexical unit and of a set of morphological features, should encourage the system to produce a higher level of grammaticality.

To this end, we have used a part of the test suite

[11] Performing independently two system runs for ensembling would have given results easier to analyze, which we were not able to provide due to the cost of such practice.

		nouns	adjectives			verbs				mean
target	system	case	gender	number	case	number	person	tense	polarity	
Czech	**fully inflected w.**	.208	.295	.272	.310	.125	.070	.086	.061	.178
	factored norm.	**.165**	.308	**.236**	.273	**.105**	**.059**	**.067**	**.042**	**.157**
	factored lemmas	.206	**.278**	.240	**.269**	.125	.074	.090	.067	.169
Latvian	**fully inflected w.**	.263	.640	.623	.669	.140	.233	.142		.387
	factored norm.	.220	**.580**	**.577**	**.617**	.108	.170	.111		**.340**
	factored lemmas	**.213**	.608	.606	.643	**.099**	**.163**	**.092**		.346

Table 6: Morphological prediction consistency (Entropy).

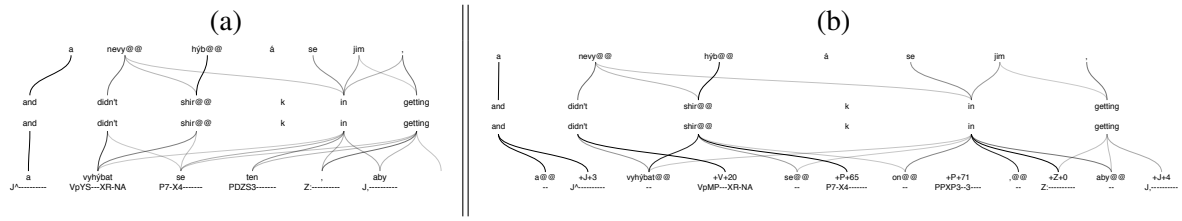

Figure 2: An example of attention weight distribution in FNMT (bottom) and fully inflected words (top) output systems aligned to the source sentence (middle) for English-to-Czech. (a) corresponds to the factored lemmas system and (b) factored norm system

provided by Burlot and Yvon (2017a), who propose an evaluation of the morphological competence of a machine translation system performed on an automatically produced test suite. For each source test sentence from a monolingual corpus (the *base*), several *variants* are generated, containing exactly one difference with the base, and focusing on a specific *target* lexeme of the base. We took the part of the test labeled as "C-set" that focuses on a word in the *base* sentence and produces *variants* containing synonyms and antonyms of this word. Thus the consistency of morphological choices is tested over lexical variation (eg. synonyms and antonyms all having the same tense) and the success is measured based on the average normalized entropy of morphological features in the set of target sentences. The systems used are the ensembles of two models introduced in § 6 (the inflected word system is our best system for each language pair).

The results of this procedure are shown in Table 6. Entropy demonstrates how confident a system is *wrt.* a specific morphological feature across synonyms and antonyms. While NMT systems on fully inflected words are well-known to produce fluent outputs, we always observe a lower entropy with the factored systems over all features, except for the lemma-based system on Czech verbs. This tends to show that the prediction of any morphological feature is more confident when it is explicitly modeled by a separate objective focused on morphology, disregarding lexical variations.

8 Conclusion

In this paper, we have presented various models integrating factored word representations for neural machine translation systems. Additionally to results with automatic metrics reporting significant improvements over a strong baseline, we provided a qualitative analysis focusing on the grammatical competence of FNTM systems that showed the benefits of explicitly modeling morpho-syntactic information.

Our experiments have shown that the cluster ID from the morphological normalization of target words brings useful information to the system by enabling a better correspondence of both factors' predictions. This specificity, as well as the improvements given by constrained decoding, brings us to future work focusing on the modelization of a stronger dependency of the second factor towards the first one in the FNMT architecture.

Acknowledgments

This work has been partly funded by the European Unions Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21) and the French National Research Agency (ANR) through the CHIST-ERA M2CR project, under the contract number ANR-15-CHR2-0006-01.

References

Andrei Alexandrescu and Katrin Kirchhoff. 2006. Factored neural language models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. NAACL-Short '06, pages 1–4.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR 2015*. San Diego, CA.

Ondřej Bojar. 2007. English-to-Czech factored machine translation. In *Proc. of the 2nd WMT*. Prague, Czech Republic, pages 232–239.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 Metrics Shared Task. In *Proc. WMT*. Berlin, Germany, pages 199–231.

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. 2012. Probes in a taxonomy of factored phrase-based models. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Stroudsburg, PA, USA, WMT '12, pages 253–260.

Ondřej Bojar and Kamil Kos. 2010. Failures in English-Czech phrase-based MT. In *Proc. of the 5th WMT*. pages 60–66.

Franck Burlot, Elena Knyazeva, Thomas Lavergne, and François Yvon. 2016. Two-Step MT: Predicting Target Morphology. In *Proc. IWSLT*. Seattle, USA.

Franck Burlot and François Yvon. 2017a. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*. Association for Computational Linguistics, Copenhagen, Denmark.

Franck Burlot and François Yvon. 2017b. Learning morphological normalization for translation from and into morphologically rich language. *The Prague Bulletin of Mathematical Linguistics (Proc. EAMT)* (108):49–60.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *arXiv preprint arXiv:1706.00457* .

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. *CoRR* abs/1607.01628.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. SSST@EMNLP*. Doha, Qatar, pages 103–111.

R. Marta Costa-jussà and R. José A. Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pages 357–361.

Daniel Déchelotte, Gilles Adda, Alexandre Allauzen, Olivier Galibert, Jean-Luc Gauvain, Hélène Maynard, and François Yvon. 2008. LIMSI's statistical translation systems for WMT'08. In *Proceedings of NAACL-HLT Statistical Machine Translation Workshop*. Columbus, Ohio.

Michael Denkowski and Graham Neubig. 2017. Stronger Baselines for Trustable Results in Neural Machine Translation. *arXiv preprint arXiv:1706.09733* .

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL*.

Ahmed El Kholy and Nizar Habash. 2012a. Rich morphology generation using statistical machine translation. In *Proc. INLG*. pages 90–94.

Ahmed El Kholy and Nizar Habash. 2012b. Translate, predict or generate: Modeling rich morphology in statistical machine translation. In *Proc. EAMT*. Trento, Italy, pages 27–34.

Orhan Firat and Kyunghyun Cho. 2016. Conditional gated recurrent unit with attention mechanism. `github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf`.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073* .

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling inflection and word-formation in SMT. In *Proc. EACL*. Avignon, France, pages 664–674.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation architectures. In *Proceedings of the International Workshop on Spoken Language Translation*. Seattle, USA, IWSLT'16.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China, pages 1–10.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical MT. In *Proc. ACL:Systems Demos*. Prague, Czech Republic, pages 177–180.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015. Character-based neural machine translation. *CoRR* abs/1511.04586. http://arxiv.org/abs/1511.04586.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *CoRR* abs/1511.06114.

Minh-Thang Luong and D. Christopher Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 1054–1063.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proc. ACL*. Prague, Czech Republic, pages 128–135.

A. Mnih and G.E. Hinton. 2008. A scalable hierarchical distributed language model. In *Advances in Neural Information Processing Systems 21*. volume 21, pages 1081–1088.

Jan Niehues, Thanh-Le Ha, Eunah Cho, and Alex Waibel. 2016. Using factored word representation in neural network language models. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 74–82.

Peteris Paikens, Laura Rituma, and Lauma Pretkalnina. 2013. Morphological analysis with limited resources: Latvian example. In *Proc. NODALIDA*. pages 267–277.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA, ACL '02, pages 311–318.

Anthony Rousseau. 2013. XenC: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics* 100:73–82.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. *CoRR* abs/1606.02892.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proc. ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1715–1725.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proc. EMNLP*. Doha, Qatar, pages 202–206.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proc. ACL: System Demos*. Baltimore, MA, pages 13–18.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proc. ACL-08: HLT*. Columbus, OH, pages 514–522.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation Edit Rate on Character Level. In *Proc. WMT*. Berlin, Germany, pages 505–510.

Marion Weller, Alexander M. Fraser, and Sabine Schulte im Walde. 2013. Using subcategorization knowledge to improve case prediction for translation to german. In *ACL (1)*. The Association for Computer Linguistics, pages 593–603.

Youzheng Wu, Hitoshi Yamamoto, Xugang Lu, Shigeki Matsuda, Chiori Hori, and Hideki Kashioka. 2012. Factored recurrent neural network language model in TED lecture transcription. In *IWSLT*.