LIMSI@WMT'17

Franck Burlot and **Pooyan Safari** and **Matthieu Labeau**
Alexandre Allauzen and **François Yvon**
LIMSI, CNRS, Université Paris Saclay, 91 403 Orsay, France
firstname.lastname@limsi.fr

Abstract

This paper describes LIMSI's submissions to the news shared task at WMT'17 for English into Czech and Latvian, as well as related experiments. This year's novelties consist in the use of a neural machine translation system with a factored output predicting simultaneously a lemma decorated with morphological information and a fine-grained part-of-speech. Such a type of system drew our attention to the specific step of reinflection, where lemmas and parts-of-speech are transformed into fully inflected words. Finally, we ran experiments showing an efficient strategy for parameter initialization, as well as data filtering procedures.

1 Introduction

The contribution of LIMSI laboratory to the WMT 2017 News shared task consisted in the submission of different systems for English-to-Czech, as well as with this year's "guest" language pair: English-to-Latvian.

Our main focus was on translation into morphologically rich languages (MRL), a challenging question in current state-of-the-art neural machine translation (NMT) architectures. Indeed, the variety of target word forms in these languages requires the use of an open vocabulary. To tackle this issue, we have experimented with a factored neural machine translation system predicting simultaneously at each timestep a normalized word and a fine-grained part-of-speech (section 3). A normalized word (section 5.2) is a specific representation where we removed part of the morphological content of the word, keeping only the features that are relevant to the source language.

Such a factored architecture required a non-trivial step consisting in reinflecting the MT predictions, i.e. transforming normalized words and parts-of-speech into fully inflected words. To this end, we have experimented with a character-based language model that is used to select ambiguous word forms returned by a look-up table (section 5.5).

Further experiments show the use of an auto-encoder to initialize the NMT system's encoder (section 4.1), which enables a faster convergence of the parameter and therefore a lower training time.

Finally, we report experiments performed with different data filtering procedures (section 4.2) and their impact on translation quality.

2 Data and Preprocessing

The pre-processing of English data relies on in-house tools (Déchelotte et al., 2008). All the Czech data were tokenized and truecased using 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.[1] Latvian PoS-tags were obtained with the LU MII Tagger (Paikens et al., 2013). All the data used to train our systems were provided at WMT'17.[2]

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.

All systems were also trained on synthetic parallel data (Sennrich et al., 2016a). The Czech

[1] www.tilde.com

[2] www.statmt.org/wmt17

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 257–264
Copenhagen, Denmark, September 711, 2017.

monolingual corpus News-2016 was backtranslated to English using the single best system provided by the University of Edinburgh from WMT'16.[3] We then added five copies of News-commentary and 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). This resulted in about 14M parallel sentences.

The English-to-Latvian systems used all the parallel data provided at WMT'17. The DCEP corpus was filtered with the Microsoft sentence aligner[4] 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.

Training was carried on with synthetic parallel data. We used a backtranslation 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 Byte Pair Encoding (BPE) models (Sennrich et al., 2016b) for each language pair and system setup were learned on the bibtext (ie. not synthetic) parallel data used for the MT system. 90k merge operations where performed to obtain the final vocabularies.

3 System Setup

Results are reported for two NMT systems: Nematus (Sennrich et al., 2017) and NMTPY (Caglayan et al., 2017).

3.1 NMTPY

Once the data was preprocessed, only sentences of a maximum length of 50 were kept in the training data, except for the setup where cluster IDs were split in normalized words (see § 5). In this case, we set the maximum length to 100.

All NMTPY systems have an embedding dimension of 512 and hidden states of dimension 1024 for both encoder and decoder, which are implemented as GRU units. Dropout is enabled on source embeddings, encoder states, as well as output layer. When training starts, all parameters are initialized with Xavier (Glorot and Bengio, 2010). In order to slightly speed up training on bitext parallel data, the learning rate was set to 0.0004, patience to 30 with validation every 20k updates. On 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.

3.2 Nematus

The setup for Nematus is very similar to the one presented in the previous section. Training was performed on sentences with the same maximum length, the same embedding and hidden unit size. The difference lies in the fact that dropout for Nematus systems was enabled on all layers. The optimizer used was Adadelta (Zeiler, 2012) and all systems had their learning rate set to 0.0001.

4 Experiments

4.1 Parameter initialization

In order to speed up the convergence of the training procedure we tried to initialize the encoder parameters with an a priori-trained model, instead of using random initialization. For the English-to-Czech translation system, this initial model was trained to translate from English into English. In order to do so, the same English corpus was fed into the neural model on both source and target side. After few updates according to the BLEU score on the validation set (which was higher than 99) it was possible to stop the training of this model and use the encoder parameters for the initialization of the main NMT system.

4.2 Data Filtering

The English-Czech training data provided at WMT'17 was very large and some corpora contained a lot of noise. For instance, we noticed several duplicate sentences in the Czeng EU parallel corpus and entire paragraphs in it were in languages other than English-Czech. Therefore, we decided to experiment with a system not containing the Czeng EU corpus. However, this lead to a degradation in terms of BLEI (see Table 1).

In another attempt, instead of removing the EU corpus, a filtering process was performed to discard the duplicate sentences on both sides. As

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

[4] http://research.microsoft.com/apps/catalog/

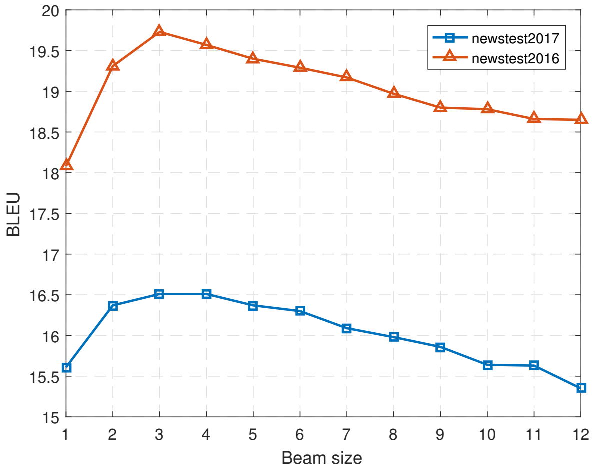

Figure 1: Comparison of different beam-size in terms of BLEU. The evaluation is performed on Newstest-2016 and Newstest-2017 English-Czech filtered data.

shown in Table 1, filtering the data results in an improvement in terms of BLEU for Newstest-2017, which is also consistent with the results we obtained on Newstest-2016 and validation set.

The filtering process was later followed by a sentence alignment check using the Microsoft sentence aligner. However, no further improvement was achieved with this method. The filtered-only data has shown the best performance on both Newstest-2016 and Newstest-2017 corpora.

Table 1: Comparison of BLEU scores of different filtering processes for English-to-Czech with Nematus systems. All the systems are evaluated with the beam search of size 2. The term "**basic**" is referred to the data without any filtering or alignment. The term **discard EU** is adopted to refer to the training without Czeng EU corpus.

data filtering	**Newstest-2016**	**Newstest-2017**
basic	18.66	15.67
discard EU	18.09	16.07
filt	**19.31**	**16.37**
filt+align	18.72	15.91

It is worthwhile to note that the model which had the best BLEU score performance on the validation data (Newstest-2015) resulted in the BLEU scores of 18.43 and 15.81 on Newstest-2016 and Newstest-2017, respectively.

Figure 1 shows the accuracy wrt. different sizes of beam during decoding. The model was trained using the English-Czech filtered data as reported in the **filt** row of the Table 1. We observed a similar trend on both Newstest-2016 and Newstest-2017, where the best performance was obtained with a beam of size 3 for both test sets.

5 Submitted systems

5.1 Factored NMT

Additionally to standard NMTPY systems (baselines), our best submissions in terms of BLEU at WMT'17 consisted in factored NMT systems.

The architecture of such systems was introduced in (García-Martínez et al., 2016). The specific setup we have used for the following factored systems consisted in an architecture that enables training towards a dual objective: at each timestep in the output sentence, a word and a PoS-tag are produced. Each one of these objectives produces a cost, that is summed in order to compute the gradients to be backpropagated.

The encoder and attention mechanism remain the same as in the baseline architecture. While in the baseline a decoder state takes as input the embedding of the prediction made at the previous step, a factored NMT decoder unit takes as input a concatenation of the two previous predictions for each factor. In this situation, the factored NMT systems deal with two sets of embeddings on target side.

Another difference lies in the hidden-to-output layer. In our setup, we have used an architecture with two different such layers: the first one takes as input the representation of the previous prediction of the first factor (word) and the second one takes the previous second factor prediction (PoS). Each layer is then passed through a last feed-forward layer leading to distinct softmax layers.

While various word representations (Burlot et al., 2017) can be used in the first factor, our system predict at each timestep on the target side a normalized word and a PoS-tag.

	fully infl.	**norm. words**
plain	kočky	kočka+Noun+7
subword	ko- čky	ko- čka- Noun+7

Table 2: Different representations of the Czech word *kočky* (cats).

5.2 Normalization of Target Morphology

Both Czech and Latvian are morphologically rich languages, as opposed to the English source. Such

	Newstest-2016			**Newstest-2017**		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
baseline	24.24	57.41	52.81	19.89	54.51	58.29
factored	23.77	57.50	52.53	19.95	54.71	58.30
+ nk-best	**24.59**	**57.95**	**52.08**	**20.54**	**54.99**	**58.06**

Table 3: Scores for English-to-Czech systems

	Newsdev-2017			**Newstest-2017**		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
baseline	22.48	57.69	52.83	14.86	52.00	62.57
+ n-best	23.11	58.13	52.21	15.22	52.37	62.08
factored	21.33	57.11	53.56	15.10	52.19	62.52
+ nk-best	**24.19**	**58.72**	**51.89**	**16.30**	**53.18**	**61.11**

Table 4: Scores for English-to-Latvian systems

differences between the source and target languages leads to difficulties. Indeed, an English adjective, that is invariable, may be translated into multiple different word forms corresponding to the same lemma. Such a variety of forms on the target side leads to serious sparsity issues and makes the estimate of reliable translation probabilities hard.

To address this issue, both Czech and Latvian vocabularies have been normalized. The normalization of a MRL consists in selecting the morpho-syntactic information that should remain encoded in a word. This selection is motivated by the fact that a target word contains more specificities than its source-side counterpart(s), leading to a lack of symmetry between both languages. For instance, when translating from English into Czech, target nouns mark grammatical case, which is removed in (Burlot et al., 2016) in order to make Czech nouns look more like their English translation(s).

Such a normalization is usually performed using hand-crafted rules and requires expert knowledge for each language pair. In this paper, normalized words are obtained with an automatic and data-driven method[5] introduced in (Burlot and Yvon, 2017a).

In a nutshell, it performs a clustering of the morphologically rich language by grouping together words that tend to share the same translation(s) in English. In order to measure this translation similarity and using word alignments, the conditional entropy of the translation probability distribution over the English vocabulary is computed for each word form. The model merges two words whenever the resulting aggregate cluster does not lead to an increase of conditional entropy, which guaranties a minimal loss of information during the clustering procedure.

The normalization model 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*), therefore a merge consists in grouping into one cluster two different tag sequences. As a result of this procedure, we obtain words represented as a lemma and a cluster identificator (ID), i.e. a coarse PoS and an arbitrary integer, like *kočka+Noun+7* in Table 2. In this example, the cluster ID *Noun+7* stands for a set of fine-grained PoS, like { *Sing+Nominative, Sing+Accusative, ...* }.

In our setup, the cluster ID was systematically split from the lemma. BPE segmentation was thus learned and applied to lemmas. Whenever the factored NMT 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 PoS that is relevant to it, thus avoiding, for instance, the prediction of a verbal PoS for the Czech noun *kočka* (cat).

The normalization of the Czech data was trained on the bibtext parallel data used to train the MT systems (see § 2), except Czeng fiction and paraweb subcorpora, which lead to over 10M sentences. As for the normalization of Latvian data it was trained on the same bitext parallel sentences used to train the MT systems.

5.3 Reinflection

The factored systems predict at each time step a normalized word and a PoS-tag, which requires a non-trivial additional step producing sentences in

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

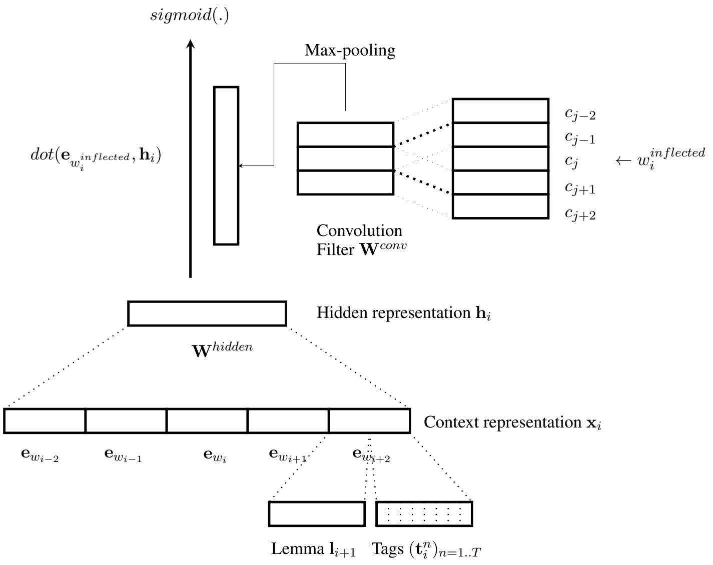
$sigmoid(.)$
Max-pooling
$dot(\mathbf{e}_{w_i^{inflected}}, \mathbf{h}_i)$
c_{j-2}
c_{j-1}
c_j
c_{j+1}
c_{j+2}
$\leftarrow w_i^{inflected}$
Convolution Filter $\mathbf{W}^{conv}$
Hidden representation $\mathbf{h}_i$
$\mathbf{W}^{hidden}$
Context representation $\mathbf{x}_i$
$\mathbf{e}_{w_{i-2}}$
$\mathbf{e}_{w_{i-1}}$
$\mathbf{e}_{w_i}$
$\mathbf{e}_{w_{i+1}}$
$\mathbf{e}_{w_{i+2}}$
Lemma $\mathbf{l}_{i+1}$
Tags $(\mathbf{t}_i^n)_{n=1..T}$

Figure 2: Architecture of the neural reinflection model

a fully inflected language. We refer to this last step as reinflection.

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

A first way to address 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 lookup, 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 a word-based system select the right word form from the dictionary. To this end, k-best hypothesis 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 hypothesis according to the unigram model mentioned above.

For Czech word form generation, 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 dictionary. As one could expect, we obtained a large quantity of word forms (nearly 2.5M), among which a lot of noise was noticed.

5.4 Experimental Results

The systems we have submitted at WMT'17 are more specifically the following:

- English-to-Czech baseline: Ensemble of 5 best models.
- English-to-Czech factored: Ensemble of 2 best models with nk-best rescoring using the single best baseline.
- English-to-Latvian baseline: Ensemble of 3 best models with n-best rescoring using the single best Nematus system.
- English-to-Latvian factored: Ensemble of 3 best models with nk-best rescoring using the single best Nematus system.

The results are reported for these systems in tables 3 and 4, using BLEU, as well as BEER (Stanojević and Sima'an, 2014) and CharacTER (Wang et al., 2016), which have shown a high correlation with human rankings for MRL (Bojar et al., 2016).

As mentioned in Section 5.3, k-best hypothesis from factored systems are rescored using a fully inflected word-based system. For Czech, we set

	Newstest-2016			Newstest-2017		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
unigrams	24.24	57.41	52.81	19.89	54.51	58.29
+ n-best	24.47	57.91	52.16	20.53	54.99	58.05
neural	21.10	56.35	53.35	17.60	53.47	59.34
+ n-best	21.52	56.36	53.52	18.12	53.64	59.21

Table 5: Scores for different English-to-Czech reinflection methods.

	Newsdev-2017			Newstest-2017		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
unigrams	22.48	57.69	52.83	14.86	52.00	62.57
+ n-best	22.06	57.58	52.92	15.34	52.52	61.98
neural	17.48	55.38	54.82	12.39	50.75	63.85
+ n-best	17.96	55.69	54.43	12.64	50.89	63.62

Table 6: Scores for different English-to-Latvian reinflection methods.

k to 10. For Latvian, the $k = 100$ best hypothesis were taken from the dictionary, in order to mitigate the poor quality of this dictionary by relying more on the rescoring system. Additionally to the k-best hypothesis from the dictionary, we also took the n-best hypothesis from the factored NMT system ($n = 30$), which lead to the rescoring of nk-best hypothesis by an inflected word based system.

The improvement given by the nk-best setups show the advantage of using a word based model to select the right word forms instead of relying on simple unigram frequencies.

5.5 Reinflection Experiments

To address the disadvantages of the reinflection methods presented in section 5.3, we investigated a neural reinflection model. The general architecture is presented in figure 2. The model first takes as input a n-gram centered on the position to reinflect. To each position corresponds a lexical unit and T PoS-tags, which are represented by embeddings $\mathbf{l}_i$ and $(\mathbf{t}_i^n)_{n=1..T}$. These are concatenated into a context representation $\mathbf{x}_i$ and transformed into a hidden representation $\mathbf{h}_i = \mathbf{W}^{hidden}\mathbf{x}_i + \mathbf{b}$.

The second input is a candidate inflected form $w_i^{inflected}$. We represent it as the sequence of its characters, and use a convolutional layer (Santos and Zadrozny, 2014) to build its vectorial representation $\mathbf{e}_{w_i^{inflected}}$. The product of these two representations goes through a *sigmoid* activation function. We train the model in a supervised way, by feeding positive and negative examples of inflected forms, with labels 1 and 0. At test time, the model is given all possible inflected forms obtained in the dictionary, and we choose the one obtaining the best score.

However, our first results show accuracies under the performances of the unigram model presented in section 5.3, for both Czech and Latvian (see Tables 5 and 6). In future work, we plan to use such a model with a beam search.

6 Morphology prediction quality

In this section, we attempt to evaluate the improvement of our factored NMT systems over the baselines. To this end, we ran the evaluation introduced in (Burlot and Yvon, 2017b) over all our WMT submissions.

The evaluation of the morphological competence of a machine translation system is performed on an automatically produced test suite. For each source test sentence from a monolingual corpus (the *base*), one (or several) *variant(s)* are generated, containing exactly one difference with the base, focusing on a specific *target* lexeme of the base. These variants differ on a feature that is expressed morphologically in the target, such as the person, number or tense of a verb; or the number or case of a noun or an adjective. This artificial test set is then translated with a machine translation system. The machine translation system is deemed correct if the translations of the base and variant differ in the same way as their respective source. Another setup focuses on a word in the *base* sentence and produces *variants* containing antonyms and synonyms of this word. The expected translation is then synonyms and antonyms bearing the same morphological features as the initial word.

There are three types of contrasts implying different sorts of evaluation:

- A: We check whether the morphological feature inserted in the source sentence has been translated (eg. plural number of a noun). Ac-

	verbs			pronouns		others		mean
System	past	future	neg.	fem.	plur.	noun nb.	compar.	
NMT baseline	92.6%	86.2%	96.0%	91.4%	79.2%	94.6%	76.2%	88.0%
Factored NMT	94.2%	88.0%	95.4%	91.2%	80.0%	96.2%	75.0%	88.6%

Table 7: Sentence pair evaluation for English-to-Czech (A-set).

	coordinated verbs			coord.n	pronouns to nouns			prep.	mean
System	number	person	tense	case	gender	number	case	case	
NMT baseline	76.6%	77.0%	69.2%	90.4%	90.8%	92.6%	92.2%	95.3%	85.5%
Factored NMT	77.6%	77.4%	70.6%	89.0%	91.4%	90.8%	91.6%	96.1%	85.6%

Table 8: Sentence pair evaluation for English-to-Czech (B-set).

	nouns	adjectives			verbs				mean
System	case	gender	number	case	number	person	tense	negation	
NMT baseline	.205	.303	.262	.301	.138	.068	.082	.054	.177
Factored NMT	.197	.287	.255	.292	.110	.062	.081	.056	.168

Table 9: Sentence group evaluation for English-to-Czech with Entropy (C-set).

	verbs		pronouns		nouns	mean
System	past	future	fem.	plur.	number	
NMT baseline	68.8%	84.6%	64.2%	86.8%	73.0%	75.5%
Factored NMT	69.6%	82.8%	62.0%	89.0%	70.6%	74.8%

Table 10: Sentence pair evaluation for English-to-Latvian (A-set).

	coordinated verbs			coord.n	pronouns to nouns			prep.	mean
System	number	person	tense	case	gender	number	case	case	
NMT baseline	69.2%	57.6%	70.4%	41.8%	40.0%	40.8%	35.8%	54.6%	51.3%
Factored NMT	72.4%	63.4%	73.2%	34.8%	43.0%	42.2%	41.4%	55.5%	53.2%

Table 11: Sentence pair evaluation for English-to-Latvian (B-set).

	nouns	adjectives			verbs			mean
System	case	gender	number	case	number	person	tense	
NMT baseline	.255	.616	.610	.644	.139	.221	.134	.374
Factored NMT	.233	.587	.582	.612	.117	.182	.113	.346

Table 12: Sentence group evaluation for English-to-Latvian with Entropy (C-set).

curacy for all morphological features is averaged over all sentences. (Tables 7 and 10)

- B: We focus on various agreement phenomena by checking whether a given morphological feature is present in both words that need to agree (eg. case of two nouns). Accuracy is computed here as well. (Tables 8 and 11)
- C: We test the consistency of morphological choices over lexical variation (eg. synonyms and antonyms all having the same tense) and measure the success based on the average normalized entropy of morphological features in the set of target sentences. (Tables 9 and 12)

The A-set focuses on the morphological adequacy of the output towards the source sentence, which does not seem to have improved with factored NMT systems. The main improvement is related to the morphological fluency of the output (B and C-sets), although the contrasts are more visible for Latvian than for Czech.

7 Conclusions

This paper described LIMSI's submissions to the News shared task at WMT2017, consisting in English-to-Czech and English-to-Latvian systems that address the issues of translating into a morphologically rich language. Further experiments reported the benefits obtained with an efficient parameter initialization procedure, as well as data filtering.

Acknowledgments

This work has been partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

References

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.

Franck Burlot, Mercedes García-Martínez, Loïc Barrault, Fethi Bougares, and François Yvon. 2017. Word Representations in Factored Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*. Association for Computational Linguistics, Copenhagen, Denmark.

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

Franck Burlot and François Yvon. 2017b. 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.

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* .

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.

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*.

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.

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

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

Cicero D. Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. JMLR Workshop and Conference Proceedings, pages 1818–1826.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 65–68.

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.

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.

Matthew D. Zeiler. 2012. Adadelta: An adaptive learning rate method. *CoRR* abs/1212.5701.