LIUM Machine Translation Systems for WMT17 News Translation Task

Mercedes García-Martínez, Ozan Caglayan[†], Walid Aransa
Adrien Bardet, Fethi Bougares, Loïc Barrault
LIUM, University of Le Mans
[†]ozancag@gmail.com
FirstName.LastName@univ-lemans.fr

Abstract

This paper describes LIUM submissions to WMT17 News Translation Task for English↔German, English↔Turkish, English→Czech and English→Latvian language pairs. We train BPE-based attentive Neural Machine Translation systems with and without factored outputs using the open source *nmtpy* framework. Competitive scores were obtained by ensembling various systems and exploiting the availability of target monolingual corpora for back-translation. The impact of back-translation quantity and quality is also analyzed for English→Turkish where our post-deadline submission surpassed the best entry by +1.6 BLEU.

1 Introduction

This paper describes LIUM Neural Machine Translation (NMT) submissions to WMT17 News Translation Task for English↔German, English↔Turkish, English→Czech and English→Latvian language pairs. We experimented with and without back-translation data for English↔German and English↔Turkish which are respectively described in Sections 3 and 4. For the latter pair, we also present an analysis about the impact of back-translation quality and quantity as well as two architectural ablations regarding the initialization and the output of recurrent decoder (Section 3).

Experiments for English→Czech and English→Latvian are performed using Factored NMT (FNMT) (García-Martínez et al., 2016) systems. FNMT is an extension of NMT which aims at simultaneously predicting the canonical form of a word and its morphological information needed to generate the final surface form. The details and results are presented in section 5. All submitted systems[1] are trained using the open source `nmtpy`[2] framework (Caglayan et al., 2017).

2 Baseline NMT

Our baseline NMT is an attentive encoder-decoder (Bahdanau et al., 2014) implementation. A bidirectional Gated Recurrent Unit (GRU) (Chung et al., 2014) encoder is used to compute source sentence annotation vectors. We equipped the encoder with layer normalization (Ba et al., 2016), a technique which adaptively normalizes the incoming activations of each hidden unit with a learnable gain and bias, after empirically observing that it improves both convergence speed and translation performance.

A conditional GRU (CGRU) (Firat and Cho, 2016; Sennrich et al., 2017) decoder with attention mechanism is used to generate a probability distribution over target tokens for each decoding step t. The hidden state of the CGRU is initialized using a non-linear transformation of the average encoder state produced by the encoder. Following Inan et al. (2016); Press and Wolf (2017), the feedback embeddings (input to the decoder) and the output embeddings are **tied** to enforce learning a single target representation and decrease the number of total parameters by target vocabulary size $\times$ embedding size.

We used Adam (Kingma and Ba, 2014) as the optimizer with a learning rate of $4e{-}4$. Weights are initialized with Xavier scheme (Glorot and Bengio, 2010) and the total gradient norm is clipped to 5 (Pascanu et al., 2013). When stated, three dropouts (Srivastava et al., 2014) are applied after source embeddings, encoder hidden

[1] Backtranslations and other data can be found at `http://github.com/lium-lst/wmt17-newstask`
[2] `http://github.com/lium-lst/nmtpy`

states and pre-softmax activations respectively. The training is early stopped if validation set BLEU (Papineni et al., 2002) does not improve for a given number of consecutive validations. A beam size of **12** is used for beam-search decoding. Other hyper-parameters including layer dimensions and dropout probabilities are detailed for each language pair in relevant sections.

3 English↔Turkish

3.1 Training

We use SETIMES2 which consists of 207K parallel sentences for training, newsdev2016 for early-stopping, and newstest2016 for model selection (internal test). All sentences are normalized and tokenized using *normalize-punctuation* and *tokenizer* [3] from Moses (Koehn et al., 2007). Training sentences that have less than 3 and more than 50 words are filtered out and a joint Byte Pair Encoding (BPE) model (Sennrich et al., 2016b) with 16K merge operations is learned on train+newsdev2016. The resulting training set has 200K sentences and 5.5M tokens (Table 1) where ∼63% and ∼50% of English and Turkish vocabularies is composed of a common set of tokens.

Language	# BPE Tokens
English	10041 = 6285 Common + 3756 En
Turkish	12433 = 6285 Common + 6148 Tr
Combined	16189

Table 1: Sub-word statistics for English, Turkish and Combined vocabularies.

All models use **200**-dimensional embeddings and GRU layers with **500** hidden units. The dropout probability P_{drop} is used for all 3 dropouts and set to 0.2 and 0.3 for EN→TR and TR→EN respectively. The validation BLEU is computed after each ∼1/4 epoch and the training stops if no improvement is achieved after 20 consecutive validations.

Data Augmentation Due to the low-resource characteristic of EN↔TR, additional training data has been constructed using back-translations (BT) (Sennrich et al., 2016a) where target-side monolingual data is translated to source language to form a Source→Target synthetic corpus. newscrawl2016 (1.7M sentences) and newscrawl2014 (3.1M sentences) are used as monolingual data for Turkish and English respectively. Although we kept the amount of synthetic data around ∼150K sentences for submitted systems to preserve *original-to-synthetic* ratio, we present an analysis about the impact of synthetic data quantity/quality as a follow-up study in Section 3.3. All back-translations are produced using the NMT systems described in this study.

3-way Tying (3WT) In addition to tying feedback and output embeddings (Section 2), we experiment with 3-way tying (3WT) (Press and Wolf, 2017) only for **EN→TR** where we use the **same** embeddings for source, feedback and output embeddings. A *combined* vocabulary of ∼16K tokens (Table 1) is then used to form a bilingual representation space.

Init-0 Decoder The attention mechanism (Bahdanau et al., 2014) introduces a time-dependent context vector (weighted sum of encoder states) as an auxiliary input to the decoder allowing implicit encoder-to-decoder connection through which the error back-propagates towards source embeddings. Although this makes it unnecessary to initialize the decoder, the first hidden state of the decoder is generally derived from the last (Bahdanau et al., 2014) or average encoder state (Sennrich et al., 2017) in common practice. To understand the impact of this, we train additional **Init-0** EN→TR systems where the decoder is initialized with an all-zero vector instead of average encoder state.

3.2 Submitted Systems

Each system is trained twice with different seeds and the one with better newstest2016 BLEU is kept when reporting single systems. Ensembles by default use the best early-stop checkpoints of both seeds unless otherwise stated. Results for *both* directions are presented in Table 2.

TR→EN baseline (E1) achieves 14.2 BLEU on newstest2017. The (E2) system trained with additional 150K BT data surpasses the baseline by ∼2 BLEU on newstest2017. The EN→TR system used for BT is a single (T5) system which is itself a BT-enhanced NMT. A contrastive system (E3) with less dropout ($P_{drop} = 0.2$) is used for our final submission which is an ensemble of 4 systems (2 runs of E2 + 2 runs of E3). In overall, an improvement of ∼3.7 BLEU over the baseline sys-

[3] The tokenizer is slightly modified to fix handling of apostrophe splitting in Turkish.

tem is achieved by making use of a small quantity of BT data and ensembling.

EN→TR baseline (T1) achieves 11.1 BLEU on newstest2017 (Table 2). (T2) which is augmented with 150K synthetic data, improves over (T1) by 2.5 BLEU. It can be seen that once 3-way tying (3WT) is enabled, a consistent improvement of up to 0.6 BLEU is obtained on newstest2017. We conjecture that 3WT is beneficiary (especially in a low-resource regime) when the intersection of vocabularies is a large set since the embedding of a common token will now receive as many updates as its occurrence count in both sides of the corpus. On the other hand, the initialization method of the decoder does not seem to incur a significant change in BLEU. Finally, using an ensemble of 4 3WT-150K-BT systems with different decoder initializations (2xT5 + 2xT6), an overall improvement of 4.9 BLEU is obtained over (T1). As a side note, 3WT reduces the number of parameters by ∼10% (12M→10.8M).

System	3WT	nt2016	nt2017
TR→EN ($P_{drop} = 0.3$)			
(E1) Baseline (200K)	×	14.2	14.2
(E2) E1 + 150K-BT	×	16.6	<u>16.1</u>
(E3) E1 + 150K-BT ($P_{drop} = 0.2$)	×	16.4	16.3
Ensemble (2xE2 + 2xE3)	×	18.1	**17.9**
EN→TR ($P_{drop} = 0.2$)			
(T1) Baseline (200K)	×	10.9	11.1
(T2) T1 + 150K-BT	×	12.7	13.6
(T3) T1 + 150K-BT + Init0	×	12.8	13.5
(T4) Baseline (200K)	✓	11.5	11.6
(T5) T4 + 150K-BT	✓	13.4	<u>14.2</u>
(T6) T4 + 150K-BT + Init0	✓	13.3	14.0
Ensemble (2xT5 + 2xT6)	✓	14.7	**16.0**

Table 2: EN↔TR: <u>Underlined</u> and **bold** scores represent contrastive and primary submissions respectively.

3.3 Follow-up Work

We dissect the output layer of CGRU NMT (Sennrich et al., 2017) which is conditioned (Equation 1) on the hidden state h_t of the decoder, the feedback embedding y_{t-1} and the weighted context vector c_t. We experiment with a *simple output* (Equation 2) which depends only on h_t similar to Sutskever et al. (2014). The target probability distribution is computed (Equation 3) using softmax on top of this output transformed with W_o.

$$o_t = \tanh(\mathbf{W_h} h_t + y_{t-1} + \mathbf{W_c} c_t) \quad (1)$$

$$o_t = \tanh(\mathbf{W_h} h_t) \quad (2)$$

$$P(y_t) = \text{softmax}(\mathbf{W_o} o_t) \quad (3)$$

System	# Sents	nt2016		nt2017	
		Single	Ens	Single	Ens
(B0) Only SETIMES2	200K	11.5	12.8	11.6	13.0
(B1) Only 1.0M-BT-E1	1.0M	13.6	14.5	14.8	16.3
(B2) B0 + 150K-BT-E1	350K	13.2	14.2	14.3	15.4
(B3) BT-E2		13.4	14.1	14.2	14.9
(B4) B0 + 690K-BT-E1	890K	14.8	15.4	15.9	17.1
(B5) BT-E2		14.7	15.6	16.1	16.9
(B6) B0 + 1.0M-BT-E1	1.2M	**14.9**	**15.6**	**16.2**	**17.5**
(B7) BT-E2		14.9	15.5	16.0	17.0
(B8) B0 + 1.7M-BT-E1	1.9M	14.7	15.4	16.4	17.1
(B9) BT-E2		14.8	15.7	16.1	16.7

Table 3: Impact of back-translation quantity and quality for EN→TR: all systems are 3WT, (B0) is the same as (T4) from Table 2.

As a second follow-up experiment, we analyse the impact of BT data quantity and quality on final performance. Four training sets are constructed by taking the original 200K training set and gradually growing it with BT data of size 150K, 690K, 1.0M and 1.7M (all-BT) sentences respectively. The source side of the monolingual Turkish data used to create the synthetic corpus are translated to English using two different TR→EN systems namely (E1) and (E2) where the latter is better than former on newstest2016 by 2.4 BLEU (Table 2).

The results are presented in Table 3 and 4. First, (B1) trained with *only* synthetic data turns out to be superior than the baseline (B0) by 3.2 BLEU. The ensemble of (B1) even surpasses our primary submission. Although this may indicate the impact of training set size for NMT where a large corpus with synthetic source sentences leads to better performance than a human-translated but small corpus, a detailed analysis would be necessary to reveal other possible reasons.

Second, it is evident that increasing the amount of BT data is beneficial regardless of *original-to-synthetic* ratio: the system (B6) achieves +4.6 BLEU compared to (B0) on newstest2017 (11.6→16.2). The single (B6) is even slightly better than our ensemble submission (Table 4). The +2.4 BLEU gap between back-translators E1 and E2 does not seem to affect final performance where both groups achieve more or less the same scores.

Finally, the *Simple Output* seems to perform slightly better than the original output formulation. In fact, our final *post-deadline* submission which surpasses the winning UEDIN system[4] by 1.6 BLEU (Table 4) is an ensemble of four (B6) systems two of them being *SimpleOut*. Conditioning the target distribution over the weighted context vector c_t creates an auxiliary gradient flow from the cross-entropy loss to the encoder by skipping the decoder. We conjecture that conditioning only over the decoder's hidden state h_t forces the network (especially the decoder) to better learn the target distribution. Same gradient flow also happens for feedback embeddings in the original formulation (Equation 1).

System	Single	Ens
LIUM	-	16.0
UEDIN	-	**16.5**
(B1) Only BT	14.8	16.3
(B6) SETIMES2 + BT	16.2	17.5
(B6) + *SimpleOut*	16.6	17.6
Ensemble (2xB6 + 2xB6-*SimpleOut*)	-	**18.1**

Table 4: Summary of follow-up results for EN→TR newstest2017: UEDIN is the best WMT17 matrix entry before deadline while LIUM is our primary submission (Table 2).

4 English↔German

We train two types of model: first is trained with only parallel data provided by WMT17 (5.6M sentences), the second uses the concatenation (9.3M sentences) of the provided parallel data and UEDIN WMT16 back-translation corpus [5]. Prior to training, all sentences are normalized, tokenized and truecased using *normalize-punctuation*, *tokenizer* and *truecaser* from Moses (Koehn et al., 2007). Training sentences with less than 2 and more than 100 units are filtered out. A joint Byte Pair Encoding (BPE) model (Sennrich et al., 2016b) with 50K merge operations is learned on the *training data*. This results in a vocabulary of 50K and 53K tokens for English and German respectively.

The training is stopped if no improvement is observed during 30 consecutive validations on *newstest2015*. Final systems are selected based on *newstest2016* BLEU.

4.1 Submitted Systems

EN→DE The baseline which is an NMT with **256**-dimensional embeddings and **512**-units GRU layers, obtained 23.26 BLEU on newstest2017 (Table 5). The addition of BT data improved this baseline by 1.7 BLEU (23.26→24.94). Our primary submission which achieved 26.60 BLEU is an ensemble of 4 systems: 2 best checkpoints of an NMT and 2 best checkpoints of an NMT with 0-initialized decoder (See section 3.1).

DE→EN Our primary DE→EN system (Table 5) is an ensemble without back-translation (No-BT) of two NMT systems with different dimensions: 256-512 and 384-640 for embeddings and GRU hidden units respectively. Our post-deadline submission which is an ensemble with back-translation (BT) improved over our primary system by +4.5 BLEU and obtained 33.9 BLEU on newstest2017. This ensemble consists of 6 different systems (by varying the seed and the embedding and the GRU hidden unit size) trained with WMT17 and back-translation data.

System	# Params	nt2016	nt2017
EN→DE Baseline	35.0M	29.11	23.26
+ synthetic		31.08	24.94
primary ensemble		33.89	**26.60**
DE→EN Baseline	52.9M	33.13	29.42
primary ensemble (No-BT)		33.63	*30.10*
+ synthetic		37.36	32.20
post-deadline ensemble (BT)		**39.07**	**33.90**

Table 5: BLEU scores computed with *mteval-v13a.pl* for EN↔DE systems on newstest2016 and newstest2017.

5 English→{Czech,Latvian}

The language pairs English→Czech and English→Latvian are translated using a Factored NMT (FNMT) system where two symbols are generated at the same time. The FNMT systems are compared to a baseline NMT system similar to the one described in Section 2.

5.1 Factored NMT systems

The FNMT system (García-Martínez et al., 2016) is an extension of NMT where the lemma and the Part of Speech (PoS) tags of a word (i.e. factors)

[4] http://matrix.statmt.org

[5] http://data.statmt.org/rsennrich/wmt16_backtranslations

are produced at the output instead of its surface form. The two output symbols are then combined to generate the word using external linguistic resources. The low frequency words in the training set can benefit from sharing the same lemma with other high frequency words, and also from sharing the factors with other words having the same factors. The lemma and its factors can sometimes generate new surface words which are unseen in the training data. The vocabulary of the target language contains only lemmas and PoS tags but the total number of surface words that can be generated (i.e. virtual vocabulary) is larger because of the external linguistic resources that are used. This allows the system to correctly generate words which are considered unknown words in word-based NMT systems.

We experimented with two types of FNMT systems which have a second output in contrast to baseline NMT. The first one contains a single hidden to output layer (*h2o*) which is then used by two separate softmaxes while the second one contains two separate *h2o* layers each specialized for a particular output. The lemma and factor sequences generated by these two outputs are constrained to have the same length.

The results reported in Tables 6 and 7 are computed with *multi-bleu.perl* which makes them consistently lower than official evaluation matrix scores[6].

5.2 Training

All models use **512**-dimensional embeddings and GRU layers with **1024** hidden units. The validation BLEU is computed after each 20K updates and the training stops if no improvement is achieved after 30 consecutive validations. The rest of the hyperparameters are the same as Section 2.

The NMT systems are trained using all the provided bitext processed by a joint BPE model with 90K merge operations. The sentences longer than 50 tokens are filtered out after BPE segmentation. For FNMT systems, BPE is applied on the lemma sequence and the corresponding factors are repeated when a split occurs.

We also trained systems with synthetic data which are initialized with a previously trained model on the provided bitext only. For these systems, the learning rate is set to 0.0001 and the validations are performed every 5K updates in order to avoid overfitting on synthetic data and forgetting the previously learned weights. Two models with different seeds are trained for NMT and FNMT systems for ensembling purposes.

5.3 N-best Reranking

We experimented with different types of N-best reranking of hypotheses generated with beam search (beam size = 12) using our best FNMT. For each hypothesis, we generate the surface form with the factors-to-word procedure, which can be ambiguous. Since a single {lemma, factors} pair may lead to multiple possible words, k possible words are considered for each pair (with k being 10 for Czech and 100 for Latvian). Finally, the hypotheses are rescored with our best word-based NMT model to select the 1-best hypothesis.

For English→Latvian, we have also performed N-best reranking with two Recurrent Neural Network Language Models (RNNLM), a simple RNNLM (Mikolov et al., 2010) and GRU-based RNNLM included in *nmtpy*. The RNNLMs are trained on WMT17 Latvian monolingual corpus and the target side of the available bitext (175.2M words in total). For the FNMT system, the log probability obtained by our best word-based NMT model is also used in addition to the RNNLM scores. The reranking is done using the *nbest* tool provided by the CSLM toolkit[7] (Schwenk, 2010). (The score weights were optimized with CONDOR (Vanden Berghen and Bersini, 2005) to maximize the BLEU score on newsdev2017 set.)

5.4 English→Czech

The English→Czech systems are trained using approximately 20M sentences from the relevant news domain parallel data provided by WMT17. Early stopping is performed using newstest2015 and newstest2016 is used as internal test set. All datasets are tokenized and truecased using the Moses toolkit (Koehn et al., 2007). PoS tagging is performed with Morphodita toolkit (Straková et al., 2014) as well as the reinflection to go from factored representation to word. Synthetic data is generated from news-2016 monolingual corpus provided by Sennrich et al. (2016a). In order to focus more on the provided bitext, five copies of news-commentary and the *czeng* news dataset are added to the backtranslated data. Also, 5M sen-

[6] http://matrix.statmt.org

[7] http://github.com/hschwenk/cslm-toolkit

tences from the *czeng* EU corpus applying modified Moore-Lewis filtering with XenC (Rousseau, 2013). We end up with about 14M sentences and 322M words for English and 292M for Czech.

System	newstest2016	newstest2017
NMT		
(CS1) Baseline	18.30	14.90
(CS2) CS1 + synthetic	24.18	20.26
(CE1) Ensemble(CS2)	24.52	**20.44**
FNMT		
(CS3) single h2o layer	17.30	14.19
(CS4) sep. h2o layers	17.34	14.73
(CS5) CS4 + synthetic	22.30	19.34
(CS6) CS5 n-best reranking	23.39	19.83
(CE2) Ensemble(CS5) n-best reranking	24.05	**20.22**

Table 6: EN→CS. **Bold** scores represent primary submissions. Ensemble(CSn) correspond to the ensemble of 2 systems CSn trained with different seeds.

System	newsdev2017	newstest2017
NMT		
(LS1) Baseline	15.25	10.36
(LS2) LS1 + synthetic	21.88	<u>15.26</u>
(LS3) LS2 RNNLM reranking	21.98	**15.59**
(LE1) Ensemble(LS2)	22.34	15.46
(LE2) Ensemble(LS2) RNNLM reranking	22.46	16.04
FNMT		
(LS4) single h2o layer	14.45	10.45
(LS5) sep. h2o layers	14.39	10.69
(LS6) LS5 + synthetic	18.93	<u>13.98</u>
(LS7) LS6 n-best reranking	21.24	<u>15.28</u>
(LS8) LS6 RNNLM reranking	21.79	**15.51**
(LE3) Ensemble(LS6) n-best reranking	21.90	15.35
(LE4) Ensemble(LS6) RNNLM reranking	21.87	15.53

Table 7: EN→LV. <u>Underlined</u> and **bold** scores represent contrastive and primary submissions. Ensemble(Sn) correspond to the ensemble of 2 systems Sn trained with different seeds.

5.5 English→Latvian

The English→Latvian systems are trained using all the parallel data available for the WMT17 evaluation campaign. Data selection was applied to the DCEP corpus resulting in 2M parallel sentences.The validation set consists of 2K sentences extracted from the LETA corpus and newsdev2017 is used as internal test set.

Monolingual corpora news-2015 and 2016 were backtranslated with a Moses system Koehn et al. (2007). Similarly to Czech, we added ten copies of the LETA corpus and two copies of Europarl and *rapid* to perform corpus weighting. The final corpus contains 7M sentences and 172M words for English and 143M for Latvian.

All the Latvian preprocessing was provided by TILDE.[8] Latvian PoS-tagging is done with the LU MII Tagger (Paikens et al., 2013). Since there is no tool for Latvian to convert factors to words, all the available WMT17 monolingual data has been automatically tagged and kept in a dictionary. This dictionary maps the lemmas and factors to their corresponding word. After preprocessing, we filter out training sentences with a maximum length of 50 or with a source/target length ratio higher than 3.

5.6 Analysis

We observe that including the synthetic parallel data in addition to the provided bitext results in a big improvement in NMT and FNMT for both language pairs (see systems CS2 and CS5 in Table 6 and LS2 and LS6 in Table 7). Applying the ensemble of several models also gives improvement for all systems (CS1-CS2 and LS1-LS4). N-best reranking of FNMT systems (systems CS6 and LS7) shows bigger improvement when translating into Latvian than into Czech. This is due to the quality of the dictionary used for reinflection in each language. The Morphodita tool for Czech includes only good candidates, besides a similar tool is not available for Latvian. The reranking with RNNLM gives an improvement for the NMT and FNMT systems when translating Latvian (LS3 and LS8). As a follow-up work after submission, we ensembled two models applying reranking for Latvian and got improvements (LE2-LE4). Finally, the submitted translations for NMT and FNMT systems obtain very similar automatic scores. However, FNMT systems explicitly model some grammatical information leading to different lexical choices, which might not be captured by the BLEU score. Human evaluation shows for EN-LV task that NMT system obtained 43% of standardized mean direct assessment score and FNMT system obtained 43.2% showing a small improvement in FNMT system. Both systems obtained 55.2% in EN-CS task. Other analysis has been done (Burlot and Yvon, 2017) about morphology strength showing good results in EN-LV task. FNMT system helps when the corpus is not huge, this is the case of EN-LV task but EN-CS dataset is huge. Therefore, NMT system has already the information to learn the morphology.

[8] www.tilde.com

6 Conclusion and Discussion

In this paper, we presented LIUM machine translation systems for WMT17 news translation task which are among the top submissions according the official evaluation matrix. All systems are trained using additional synthetic data which significantly improved final translation quality.

For English→Turkish, we obtained (post-deadline) state-of-the-art results with a small model (∼11M params) by tying all the embeddings in the network and simplifying the output of the recurrent decoder. One other interesting observation is that the model trained using *only* synthetic data surpassed the one trained on genuine translation corpus. This may indicate that for low-resource pairs, the amount of training data is much more important than the correctness of source-side sentences.

For English→Czech and English→Latvian pairs, the best factored NMT systems performed equally well compared to NMT systems. However, it is important to note that automatic metrics may not be suited to assess better lexical and grammatical choices made by the factored systems.

Acknowledgments

This work was supported by the French National Research Agency (ANR) through the CHIST-ERA M2CR project, under the contract number ANR-15-CHR2-0006-01[9] and also partially supported by the MAGMAT project.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* http://arxiv.org/abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Franck Burlot and FranÃğois Yvon. 2017. 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* http://arxiv.org/abs/1706.00457.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555. http://arxiv.org/abs/1412.3555.

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

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 Thirteenth International Conference on Artificial Intelligence and Statistics*. PMLR, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256. http://proceedings.mlr.press/v9/glorot10a.html.

Hakan Inan, Khashayar Khosravi, and Richard Socher. 2016. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv preprint arXiv:1611.01462* http://arxiv.org/abs/1611.01462.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* http://arxiv.org/abs/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 machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180. http://dl.acm.org/citation.cfm?id=1557769.1557821.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. pages 1045–1048. http://www.isca-speech.org/archive/interspeech_2010/i10_1045.html.

Pēteris Paikens, Laura Rituma, and Lauma Pretkalniņa. 2013. Morphological analysis with limited resources: Latvian example. In *Proceedings of the*

[9] http://m2cr.univ-lemans.fr

19th Nordic Conference of Computational Linguistics (NODALIDA 2013). Linköping University Electronic Press, Sweden, Oslo, Norway, pages 267–277. http://www.aclweb.org/anthology/W13-5624.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. JMLR.org, ICML'13, pages III–1310–III–1318. http://dl.acm.org/citation.cfm?id=3042817.3043083.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 157–163. http://www.aclweb.org/anthology/E17-2025.

Anthony Rousseau. 2013. XenC: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics* (100):73–82. http://ufal.mff.cuni.cz/pbml/100/art-rousseau.pdf.

Holger Schwenk. 2010. Continuous-space language models for statistical machine translation. *The Prague Bulletin of Mathematical Linguistics* (93):137–146. http://ufal.mff.cuni.cz/pbml/93/art-schwenk.pdf.

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. http://aclweb.org/anthology/E17-3017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 86–96. http://www.aclweb.org/anthology/P16-1009.

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, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15(1):1929–1958. http://dl.acm.org/citation.cfm?id=2627435.2670313.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-source tools for morphology, lemmatization, pos tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 13–18. http://www.aclweb.org/anthology/P14-5003.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS'14, pages 3104–3112. http://dl.acm.org/citation.cfm?id=2969033.2969173.

Frank Vanden Berghen and Hugues Bersini. 2005. Condor, a new parallel, constrained extension of powell's uobyqa algorithm: Experimental results and comparison with the dfo algorithm. *J. Comput. Appl. Math.* 181(1):157–175. https://doi.org/10.1016/j.cam.2004.11.029.