LIUM Machine Translation Systems for WMT17 News Translation Task

Mercedes García-Martínez, Ozan Caglayan[1], Walid Aransa,
Adrien Bardet, Fethi Bougares, Loïc Barrault
LIUM, University of Le Mans
[1]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

In this paper, we present Neural Machine Translation (NMT) systems trained by LIUM for WMT17 News Translation Task. This year, we mainly focus on 4 language pairs: English↔Turkish, English↔German, English→Czech and English→Latvian.

NMT systems with and without back-translation data are trained for English↔German and English↔Turkish are 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 to Czech and English to Latvian are performed using Factored NMT systems. FNMT (García-Martínez et al., 2016) 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 are trained using the open source `nmtpy`[1] framework (Caglayan et al., 2017).

2 Baseline NMT

Our baseline NMT is an attentive encoder-decoder (Bahdanau et al., 2014) implementation. A bi-directional 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 (2016), 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). The training is early stopped if validation set BLEU (Papineni et al., 2002) does not improve for a given num-

[1]github.com/lium-lst/nmtpy

ber of consecutive validations. When stated, three dropouts (Srivastava et al., 2014) are applied after source embeddings, encoder hidden states and pre-softmax activations respectively. 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* [2] 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., 2015) with 16K merge operations is learned on train+newsdev2016. The resulting training set has 200K sentences and 5.5M words (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 token 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., 2016) 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, 2016) 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-

[2] 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 solely depends on h_t similar to Sutskever et al. (2014):

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

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

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

The target probability distribution is then calculated using a softmax activation on top of this output transformed with W_o (Equation 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.

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. This indicates that a large synthetic data having a noisy machine-translated source side, may be more useful for NMT than a small human-translated corpus in terms of final generalization ability.

Second, it is evident that increasing the amount of BT data is *beneficial* to final performance 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[3] by **1.6 BLEU** (Table 4) is an ensemble of four (B6) systems two of them being *SimpleOut*.

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

4.1 Training

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 [4].

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., 2015) 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.

4.2 Submitted Systems

The results for both directions are presented in Table 5.

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

[4] data.statmt.org/rsennrich/wmt16_backtranslations

EN→DE The baseline which is an NMT with **256**-dimensional embeddings and **512**-units GRU layers, obtained 23.26 BLEU on newstest2017. 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 an NMT with 0-initialized decoder (See section 3.1).

DE→EN Our primary DE→EN system 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 the NMT system where the lemma and the Part of Speech (POS) tags of a word (i.e. factors) 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 fre-

quency words in the training can benefit from sharing the same lemma with other words with high frequency, and also from sharing the factors with other words having the same factors. The lemma and its factors can sometimes generate a new surface words which are unseen in the training data. The vocabulary of the target language contains only the lemma and POS tags but the total number of surface words that can be generated (i.e. virtual vocabulary) is bigger because of the using of external linguistic resources. This allows the system to correctly generate the words which are considered unknown words in word-based NMT system. The architecture remains as the original baseline NMT but adding output layers for the second output. The two outputs are constrained to have the same length. Two types of FNMT models are used. The first model contains a single hidden to output (*h2o*) layer which is used by the two separated softmax. The second model contains two separated *h2o* layers, each specialized for a particular output.

The results reported in Tables 6 and 7 are computed with *multi-bleu.perl* which makes them consistently lower than official evaluation matrix scores[5].

5.2 Training

The training settings are similar to the ones mentioned in Section 2 except for the following details. The embedding size is set to 512 and the dimension of the hidden states to 1024. The patience is set to 30 and the validation is performed every 20k updates. All the provided bitext is used for training the systems. They are processed by a bilingual BPE model with 90k merges operations. After applying BPE, the sentences longer than 50 tokens are again filtered out. For FNMT systems, BPE is applied on the lemma sequence and the corresponding factors are repeated (when a a split occur).

We also train the models with synthetic data in addition to the provided bitext. Those models are initialized with a previously trained model on the provided bitext only. In this case, the learning rate is set to 0.0001 and the validations are performed every 5k updates. Both values are reduced to avoid overfitting on synthetic data and forgetting the previously trained parameters. Two models with different seeds are trained for NMT and FNMT systems and are in turn ensembled.

5.3 Reranking of the n-best output

Different types of reranking of the n-best hypotheses of our best factored NMT system have been performed. The n-best hypotheses generated by the beam search are extracted. In our experiments, we empirically set the beam size to 12. For each hypothesis, we generate the surface form with the factors-to-word procedure, which can be ambiguous. Since a single pair {lemma, factors} 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 set of hypotheses is rescored with the best performing word-based model to generate the 1-best hypothesis.

For English to Latvian, we have also performed n-best list reranking with several Recurrent Neural Network Language Model (RNNLM). The first model is a simple RNNLM (Mikolov et al., 2010) and the second one, included in *nmtpy*, uses a GRU as recurrent unit. They are trained on WMT17 Latvian monolingual corpus and the target side of the available bitext (175.2M words in total). For the factored system, the log probability obtained by our best word-based NMT model is used for rescoring in addition to the RNNLM scores. The reranking is done using the *nbest* tool provided as part of CSLM toolkit[6] (Schwenk, 2010) and the weights of all these features were optimized to maximize the BLEU score on newsdev-2017 set.

5.4 English→Czech results

The English→Czech models are trained on approximately 20M sentences from the relevant news domain parallel data provided by WMT17. Early stopping is performed using newstest-2015 and newstest-2016 was used as internal test set. All the 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. For the English-Czech language pair, synthetic data is generated from monolingual corpus news-2016 provided by Sennrich et al. (2016). In order to give more weight to the provided bitext, five copies of news-commentary and the *czeng*

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

[6] github.com/hschwenk/cslm-toolkit

news dataset are added to the backtranslated data. Also, 5M sentences 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.

5.5 English→Latvian results

The English→Latvian systems used 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 consisted of 2k sentences extracted from the LETA corpus and newsdev-2017 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 pre-processing was provided by TILDE.[7] Latvian PoS-tagging was obtained with the LU MII Tagger (Paikens et al., 2013). Since there is no tool for Latvian to convert factors to words, all monolingual data available at WMT17 has been automatically tagged and kept in a dictionary. This dictionary maps the lemmas and factors with their corresponding word. After preprocessing, we filter out sentences with a maximum length of 50 or with a source/target length ratio higher than 3 from the training data.

[7] www.tilde.com

System	newsdev2017	newstest2017
NMT		
FNMT		
(LS1) Baseline	15.25	10.36
(LS2) LS1 + synthetic	21.88	15.26
(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	13.98
(LS7) LS6 n-best reranking	21.24	15.28
(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. Underlined and **bold** scores represent contrastive and primary submissions. Ensemble(Sn) correspond to the ensemble of 2 systems Sn trained with different seeds.

5.6 Analysis

We can 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 (*E1-*E4). The FNMT system n-best reranking (systems CS6 and LS6) 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 available tool for Czech (morphodita) 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 (R1 and R3). As a follow up work, after submission, we ensembled two models applying reranking for Latvian and got improvements (*E1-*E4). 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. The human evaluation might reveal this.

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 sig-

nificantly 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[8].

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv: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.

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. `github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf`.

[8] `http://m2cr.univ-lemans.fr`

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 *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*.

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* .

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, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Meeting of the Association for Computational Linguistics*. pages 177–180.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*. volume 2, page 3.

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.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of The 30th International Conference on Machine Learning*. pages 1310–1318.

Ofir Press and Lior Wolf. 2016. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859* .

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

Holger Schwenk. 2010. Continuous space language models for statistical machine translation. In *The Prague Bulletin of Mathematical Linguistics, (93):137–146.*.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch-Mayne, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel LÃd'ubli, Antonio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. *Nematus: a Toolkit for Neural Machine Translation*, Association for Computational Linguistics (ACL), pages 65–68.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* .

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. *Improving Neural Machine Translation Models with Monolingual Data*, Association for Computational Linguistics (ACL), pages 86–96.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR* abs/1409.3215. http://arxiv.org/abs/1409.3215.