LIUM Machine Translation Systems for WMT17 News Translation Task

Mercedes Garcia-Martinez, Ozan Caglayan, Walid Aransa, Adrien Bardet, Fethi Bougares, Loïc Barrault

To cite this version:

HAL Id: hal-01742378
https://hal-univ-lemans.archives-ouvertes.fr/hal-01742378
Submitted on 26 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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† are trained using the open source nmtpy‡ 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 (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

†Backtranslations and other data can be found at http://github.com/lium-lst/wmt17-newstask
‡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* 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 \( \sim 63\% \) and \( \sim 50\% \) of English and Turkish vocabularies is composed of a common set of tokens.

<table>
<thead>
<tr>
<th>Language</th>
<th># BPE Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>10041 = 6285 Common + 3756 En</td>
</tr>
<tr>
<td>Turkish</td>
<td>12433 = 6285 Common + 6148 Tr</td>
</tr>
<tr>
<td>Combined</td>
<td>16189</td>
</tr>
</tbody>
</table>

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 \( P_{\text{drop}} \) 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 \( \sim 1/4 \) 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 \( \sim 150K \) 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 \( \sim 16K \) 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 \( \sim 2 \) 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 (\( P_{\text{drop}} = 0.2 \)) 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 \( \sim 3.7 \) BLEU over the baseline sys-
system 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 the decoder does not seem to incur a significant change in BLEU. As a side note, 3WT reduces the number of parameters by \( \sim 10\% \) (12M→10.8M).

\[
\begin{align*}
\sigma_t &= \tanh(W_h h_t + y_{t-1} + W_c c_t) \quad (1) \\
\sigma_t &= \tanh(W_h h_t) \quad (2) \\
P(y_t) &= \text{softmax}(W_o \sigma_t) \quad (3)
\end{align*}
\]

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 only \( 3.2 \) 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.

<table>
<thead>
<tr>
<th>System</th>
<th>3WT</th>
<th>nt2016</th>
<th>nt2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>(E1) Baseline (200K)</td>
<td>×</td>
<td>14.2</td>
<td>14.2</td>
</tr>
<tr>
<td>(E2) E1 + 150K-BT</td>
<td>×</td>
<td>16.6</td>
<td>16.1</td>
</tr>
<tr>
<td>(E3) E1 + 150K-BT ((P_{drop} = 0.2))</td>
<td>×</td>
<td>16.4</td>
<td>16.3</td>
</tr>
<tr>
<td>Ensemble (2xE2 + 2xE3)</td>
<td>×</td>
<td>18.1</td>
<td>17.9</td>
</tr>
<tr>
<td>(T1) Baseline (200K)</td>
<td>×</td>
<td>10.9</td>
<td>11.1</td>
</tr>
<tr>
<td>(T2) T1 + 150K-BT</td>
<td>×</td>
<td>12.7</td>
<td>13.6</td>
</tr>
<tr>
<td>(T3) T1 + 150K-BT + Ini0</td>
<td>×</td>
<td>12.8</td>
<td>13.5</td>
</tr>
<tr>
<td>(T4) Baseline (200K)</td>
<td>✓</td>
<td>11.5</td>
<td>11.6</td>
</tr>
<tr>
<td>(T5) T4 + 150K-BT</td>
<td>✓</td>
<td>13.4</td>
<td>14.2</td>
</tr>
<tr>
<td>(T6) T4 + 150K-BT + Ini0</td>
<td>✓</td>
<td>13.3</td>
<td>14.0</td>
</tr>
<tr>
<td>Ensemble (2xT5 + 2xT6)</td>
<td>✓</td>
<td>14.7</td>
<td>16.0</td>
</tr>
</tbody>
</table>

Table 2: EN↔TR: Underlined 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 \( W_o \).
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 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).

### 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.

### 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. 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.

#### EN→TR

<table>
<thead>
<tr>
<th>System</th>
<th>Single</th>
<th>Ens</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIUM</td>
<td>-</td>
<td>16.0</td>
</tr>
<tr>
<td>UEDIN</td>
<td>-</td>
<td>16.5</td>
</tr>
<tr>
<td>(B1) Only BT</td>
<td>14.8</td>
<td>16.3</td>
</tr>
<tr>
<td>(B6) SETIMES2 + BT</td>
<td>16.2</td>
<td>17.5</td>
</tr>
<tr>
<td>(B6) + SimpleOut</td>
<td>16.6</td>
<td>17.6</td>
</tr>
<tr>
<td>Ensemble (2xB6 + 2xB6-SimpleOut)</td>
<td>-</td>
<td><strong>18.1</strong></td>
</tr>
</tbody>
</table>

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).

### 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.

#### 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)
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. 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. 

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.

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

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

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 LUMII 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 ensemble two models applying reranking for Latvian and get 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 (\(\sim 11M\) 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


