The Helsinki Neural Machine Translation System*

Robert Östling
Department of Linguistics
Stockholm University

Yves Scherrer and **Jörg Tiedemann**
Department of Modern Languages
University of Helsinki

Gongbo Tang
Department of Linguistics and Philology
Uppsala University

Tommi Nieminen
Department of Modern Languages
University of Helsinki

Abstract

We introduce the Helsinki Neural Machine Translation system (HNMT) and how it is applied in the news translation task at WMT 2017, where it ranked first in both the human and automatic evaluations for English–Finnish. We discuss the success of English–Finnish translations and the overall advantage of NMT over a strong SMT baseline. We also discuss our submissions for English–Latvian, English–Chinese and Chinese–English.

1 Introduction

The Helsinki Neural Machine Translation system (HNMT) is a full-featured system for neural machine translation, with a particular focus on morphologically rich languages. We participated in the WMT 2017 shared task on news translation, obtaining the highest BLEU score for English–Finnish translation, while also performing well on English–Latvian and acceptably on English–Chinese and Chinese–English.

In addition to our participation in the shared task, this paper also details some of the other methods we have implemented and evaluated with HNMT, many of which yielded negative results and were subsequently not used in our submissions for the shared task.

1.1 HNMT

HNMT is based on the attentional encoder–decoder model due to Bahdanau et al. (2014). This is a rather minimalistic framework for NMT, and many extensions have been proposed. Of particular interest are those that allow proper and efficient handling of morphologically rich languages, such as Finnish. We combine two such approaches: the hybrid character/word model of Luong and Manning (2016), which is used for the source language encoder, and the byte-pair encoding (BPE) technique of Sennrich et al. (2016c), which is used for the target language decoder and has been successfully used for Finnish previously (Sánchez-Cartagena and Toral, 2016). As BPE can be added as a simple pre- and post-processing step, it does not affect the structure of the translation model. This means that our system can be used with character, word and BPE level generation on the target side. The structure of the network, thus, consists of three Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) layers:

1. A character-level encoder that transforms out-of-vocabulary source tokens into the same vector space as the source word embeddings.

2. A token-level bidirectional encoder that transforms a sequence of source word embeddings (or outputs from network (1) in case of OOV items) into an encoded sequence of the same length.

3. A character-, token- or BPE-level decoder that works as language model conditioned (via an attention mechanism) on the encoded source sequence from (2).

HNMT is implemented using Theano (Theano Development Team, 2016), which supports efficient training with a GPU. For optimization we use minibatch stochastic gradient descent with Adam (Kingma and Ba, 2015) for learning rate adaptation.

*The software is available from https://github.com/robertostling/hnmt under the GNU General Public License version 3.

2 Tricks from the NMT arsenal

We have implemented and evaluated a number of proposed extensions to the basic attentional encoder–decoder model. Basic experiments were carried out on English–Finnish data, unless specified otherwise.

2.1 Layer normalization

Layer normalization (Ba et al., 2016) has been proposed as a technique for speeding up training of recurrent network models. We have implemented it into HNMT as the modified LSTM described by equations (20)–(22) in Ba et al. (2016). However, as preliminary experiments did not indicate any consistent effect of using layer normalization we did not include it in our evaluation.

2.2 Variational dropout

Gal and Ghahramani (2016) proposed a method for regularization of recurrent neural networks. This has also been implemented in HNMT, but preliminary experiments on Finnish did not indicate any improvement over the baseline system. While Sennrich et al. (2016a) reported large improvements for the Romanian news translation task at WMT 2016, the amount of training data is lower than what is available for Finnish, which should explain some of the difference. They also apply dropout on the word level, whereas the HNMT application currently only drops recurrent states.

2.3 Context gates

Context gates (Tu et al., 2017) introduce an explicit model for selecting to which extent the target sentence generation should focus on the source sentence or the target context, giving the network a chance to tune the balance between adequacy and fluency. While we obtained better cross-entropy on the development set, particularly early during training, the BLEU and chrF3 evaluations on development data made us decide against the slower context gates in the final run.

2.4 Coverage decoder

Wu et al. (2016) present an empirically determined method for using the attention vectors produced during decoding in the search algorithm, to bias the decoder towards translations with reasonable length and good coverage of the source sentence.[1] We performed a grid search of the parameter space for the length, coverage and overtranslation penalties, but did not find any that resulted in higher BLEU scores on the development set than the decoder without penalties.

2.5 Forward-Backward reranking

It is trivial to train a translation model to generate translations either from the beginning or the end of the target sentence. HNMT supports selecting translation direction, and combined with its n-best list and reranking features it is simple to generate candidate translations in both directions and to combine them based on their scores. This led to some minor improvements in our English–Finnish translations.

2.6 Ensembling

HNMT supports two general modes of ensembling, as well as their combination:

- Proper ensembling where $p(w) = \frac{1}{M}\sum_{m=1}^{M} p_m(w)$ is used to predict target symbol w, given predictions $p_m(w)$ for each model m in the ensemble.
- Parameter averaging where the model's parameter vector θ is computed as $\frac{1}{N}\sum_{m=1}^{M} \theta_m$ for each model m. This only works if the different θ_m are relatively similar, typically because they were saved at different points during the same training process.

The overhead for proper ensembling is linear in the number of ensembled systems, both for training (assuming one is building an ensemble of separately trained models) and inference, while parameter averaging is essentially free. HNMT allows proper ensembling of groups of models where the parameters are averaged within each group. This flexible structure allows a number of setups, which are explored further in Section 3.2.

3 English–Finnish

In our experiments, we used all English–Finnish parallel data sets provided by WMT except the Wiki headlines, which is a small and rather noisy data set that did not contribute anything in our experiments from last year. We also added substantial amounts of backtranslated data that has

[1] The HNMT implementation of this was contributed by Stig-Arne Grönroos.

been shown to help especially in neural machine translation (Sennrich et al., 2016b) but also in statistical MT (Tiedemann et al., 2016). Table 1 lists some basic statistics of the backtranslated data sets we created out of WMT's monolingual Finnish news data from 2014 and from 2016. We applied our best constrained phrase-based SMT model for Finnish to English from last year (Tiedemann et al., 2016) that uses a factored model with multiple translation paths, morphological tags and pseudo-tokens for case-markers that correspond to English prepositions (Tiedemann et al., 2015). The system scored 20.5% lowercased BLEU on the newstest 2016 data, which was the second-best system for the task in 2016.

	sentences	Finnish	English
news2014	1,378,833	17,117,137	23,818,547
news2016	4,144,406	55,637,304	76,161,439

Table 1: Backtranslated Finnish news data.

3.1 Preprocessing and postprocessing

We trained our models on tokenized and truecased data, except for the character-level models which were trained on raw untokenized data. For the former, we applied Moses tools for Unicode/punctuation normalisation, tokenization and truecasing using a model trained on the parallel training data.

We tested three different types of word segmentation: basic word-based segmentation, supervised morphological segmentation using OMorFi (Pirinen, 2015) and byte-pair encoding (BPE) (Sennrich et al., 2016c). For the latter, we opted for a fine-grained segmentation that results in a small vocabulary of 20,000 tokens when trained on the parallel data, expecting BPE to handle various cases of compound splitting and morphological segmentation. We always used the same BPE-based segmentation and did not try to optimize the BPE parameters in any way.

During development, we observed that the English development files contained a lot of verb form contractions (of the type *wouldn't*), but that such contracted forms appear rarely in the training data. Therefore, we also added a preprocessing routine to transform the contracted forms to their uncontracted equivalents.

Finally, we found that our tokenizer/detokenizer pipeline for Finnish did not handle the hyphen/dash distinction correctly. In Finnish, the '-' sign can be used with spaces on both sides, without spaces, with a space only on the left, and with a space only on the right, as in the following examples:

(1) a. Draamaa Riossa - suomalaisnostaja pyörtyi...
'Drama in Rio - Finnish lifter fainted...'
b. Kempinski-hotelli
'Kempinski[-]hotel'
c. kissa ja hiiri -leikkiä
'cat and mouse [-]game'
d. öljy- ja kaasutoiminnot
'oil[-] and gas functions'

The tokenizer always introduces spaces on both sides, which means that the detokenizer is then unable to retrieve the original configuration. In order to remedy this problem, we applied a postprocessing step to the translated data. After detokenizing the output, for every hyphen sign, the four tokenization variants were generated and scored by the hybrid-to-character system; we then chose the tokenization variant with the highest score.

3.2 NMT Models

In preliminary experiments, we focused on different segmentation strategies for the source and target sides as well as on different proportions of backtranslations and parallel data. The models were evaluated on *newsdev2015* using lowercased BLEU and chrF3.[2] Table 2 shows some results.

In these experiments, we found BPE to be useful on the target side, but not so much on the source side. Character-level decoders are favoured by character-level evaluation scores such as chrF3, whereas BLEU favours decoders using larger units such as BPE. The best results were obtained with a combination of backtranslated and parallel data; using all backtranslations was slightly better than restricting the amount of backtranslations to match the size of the parallel data. The model based on supervised morphological segmentation followed by BPE encoding (OMorFi) yielded promising chrF3 results, but lagged behind in terms of BLEU. Further investigation is needed on the benefits and shortcomings of combining these segmentation approaches.

[2]The HNMT-internal BLEU computation is based on https://github.com/vikasnar/Bleu and on the NLTK tokenizer. The reported results are thus not directly comparable with official WMT results.

		BLEU				chrF3			
Encoder	Decoder	None	Only	Balanced	All	None	Only	Balanced	All
BPE	BPE	11.9	**14.4**	**15.7**	**15.5**	43.7	47.2	48.3	48.5
BPE	Char	9.2	13.0	13.7	14.0	41.0	**47.8**	**48.4**	48.6
Hybrid	BPE	**12.2**	13.8	15.4	15.3	43.4	47.0	48.1	49.0
Hybrid	Char	11.6	13.1	14.1	14.2	**46.3**	47.2	48.2	49.0
Hybrid	OMorFi	—	—	—	14.3	—	—	—	**49.2**

Table 2: Development results with different segmentation strategies for the source language encoder and the target language decoder and different proportions of backtranslated and parallel data (None = 2.5M sentences of parallel data + 0 sentences of backtranslated data; Only = 0 + 5.5M; Balanced = 2.5M + 2.5M; All = 2.5M + 5.5M).

BLEU	chrF3	M	SP/M	AVG
12.8	48.8	1	1	N/A
13.6	49.7	1	4	+
13.8	49.8	1	4	−
14.1	50.0	3	1	N/A
14.4	50.2	3	4	+
14.6	50.4	3	4	−

Table 3: Development results with different ensembling setups. Each configuration consists of M models, with SP/M savepoints per model, where the savepoints may be averaged (+AVG) or included as equal members in the ensemble (-AVG).

The model based on a hybrid encoder and a BPE decoder did not yield the best results in this preliminary evaluation, but showed the most robust performance across different evaluation types, training configurations and evaluation data (in particular, it outperformed other models on the *newstest2015* set). Therefore, four of our five submissions use that configuration. For comparison, we also submitted a system based on a character-level decoder.

We also investigated the effect of different ensembling combinations, and the result can be found in Table 3. In general, proper ensembling is better than savepoint averaging, but savepoint averaging is better than nothing. Further experiments revealed that the difference between an *ensemble of averaged savepoints from independent models* setup (second row from the bottom) and an *ensemble of several savepoints each from independent models* (bottom row) is not consistent, so we use the former (faster) variant for our official submissions.

The submitted character-decoder system uses 256 dimensions for word embeddings, 64 for character embeddings, 512 encoder state dimensions, 1024 decoder state dimensions, and 256 attention dimensions. We train four independent models for 72h each, and the savepoint with the best heldout chrF3 score is used (in practice we do not observe any significant overfitting, so this amounts to using nearly 72h of training for all models). Training data are the unprocessed versions of all parallel and backtranslated data. For decoding, we used proper ensembling of the four models, and averaging of the four last savepoints of each model (states were saved after each 5000 training batches).

The submitted BPE-decoder systems use the same model size as the character-decoder system. Again, we train four independent models for 72h each, using the preprocessed and BPE-encoded data, with hyphen retokenization applied as a postprocessing step. We provide two contrastive systems: one without input normalization, which shows a decrease of 0.3 BLEU, and one without hyphen retokenization, which shows a decrease of 0.9 BLEU (see Table 4).

We also propose an extended system that is based on the four models above and four additional backwards models (i.e., trained right-to-left). At test time, we generate a 10-best list of forward translations and a similar one of backward translations. We choose the best translation that occurred in both lists, or if the lists are disjoint, the translation with the highest likelihood according to the model (forward or backward) that generated it. This reranking only provided +0.1 BLEU; 48% of translations were chosen from the forward model, 22% from the backward model, and 30% occurred in both lists. This system has been ranked first in the automatic and manual evaluations.

3.3 SMT Baselines

Besides the neural MT models, we also trained various phrase-based SMT models to contrast our results with another popular paradigm. In particular, we were interested to see the effect of BPE segmentation and backtranslation on statis-

Decoder	IN	HR	Direction	BLEU
Char	+	N/A	fw	19.1
BPE	+	+	fw	20.6
BPE	−	+	fw	20.3
BPE	+	−	fw	19.7
BPE	+	+	fw+bw	**20.7**

Table 4: Submitted HNMT systems with official results. They vary with respect to decoder type, input normalization (IN), hyphen retokenization (HR), direction (forward or backward). The best result was submitted for manual evaluation, where it ranked #1 (tied with one unconstrained system).

tical MT. Both techniques are popular in neural MT but their impact on statistical MT has not been evaluated properly before. Therefore, we started a systematic comparison of different setups including various types of segmentations and data collections. All systems are based on Moses (Koehn et al., 2007) and we use standard configurations for training non-factored phrase-based SMT models using KenLM for language modeling (Heafield, 2011) and BLEU-based MERT for tuning. The only difference to the standard pipeline is the use of efmaral (Östling and Tiedemann, 2016), an efficient implementation of fertility-based IBM word alignment models with a Bayesian extension and Gibbs sampling.[3] Table 5 summarizes the results of our SMT experiments during development.

The first observation is that BPE (and also supervised morphological segmentation) is not very helpful. This is somewhat surprising as we expect a similar problem as with neural MT in the sense that the productive and rich morphology in Finnish causes problems due to data sparseness. We can see that some models benefit from BPE (see *back* and *opus*) especially if tuning is done on the word level and not on BPE-segmented output. However, we have to admit that we did not attempt to optimize the segmentation level and it can well be that the small BPE vocabulary in our setup is not working well for SMT.

Another observation is that the operation-sequence model does not lead to significant (or any) improvements. This is in contrast to related work and may be due to data sparseness again due to the morphological richness of Finnish.

The biggest surprise is the positive effect of backtranslated data. The models trained on those data sets only are in fact better than the ones trained on the official training data provided by WMT. This demonstrates the strong domain mismatch between training and test data and the use of in-domain data, even very noisy ones, seems to lead to visible benefits. In combination, we can see substantial improvements over the individual models, which demonstrates the use of backtranslation even for SMT.

[3]Software available from https://github.com/robertostling/efmaral.

newsdev15	segmentation			LM	
data	src	trg	tuning	news	+CC
WMT	word	word	word	**12.51**	**13.74**
WMT	word	BPE	word	12.16	–
WMT	word	morf	word	11.58	–
WMT	BPE	BPE	BPE	11.91	–
WMT	BPE	BPE	word	12.24	12.95
back	word	word	word	12.69	**13.69**
back	BPE	BPE	BPE	12.73	–
back	BPE	BPE	word	**12.92**	13.50
WMT+back	word	word	word	–	**14.62**
WMT+back	BPE	BPE	BPE	12.94	–
WMT+back	BPE	BPE	word	13.40	14.44
+osm	word	word	word	–	14.04
+osm	BPE	BPE	word	12.85	14.58
opus	word	word	word	14.05	15.54
opus	BPE	BPE	word	14.45	15.63
+osm	word	word	word	–	**15.82**
+osm	BPE	BPE	word	–	15.57
newstest17					
WMT+back	BPE	BPE	word	–	**16.2**
opus+osm	BPE	BPE	word	–	**17.3**

Table 5: Phrase-based SMT tested on newsdev 2015 and newstest 2017 (lowercased BLEU). Different types of segmentation in source language text (src), target language text (trg) and during minimum-error rate training (tuning): word-based, byte-pair encoding (BPE) and OMorFi-based (morf). Different data sets for training: Europarl and Rapid2016 (WMT), backtranslated Finnish news (back) and all available data sets including parallel corpora from OPUS (opus). Additional component: operation-sequence model (osm).

Another common outcome in SMT is the strong impact of language models. We can confirm this once again. Adding a second language model trained on common-crawl data (CC) has a strong influence on translation quality as we can see by the BLEU scores in Table 5.

In the manual evaluation, our best SMT system shared 6th rank with four other systems (interestingly a mix of phrase-based, rule-based and neural systems), of which two were constrained like ours.

3.4 NMT with Pre-translated Data

We were also interested in the combination of SMT and NMT using the pre-translation approach proposed by Niehues et al. (2016). In their model, SMT-based translations of the source text are simply concatenated to the input to make it possible for an NMT system to draw information from other MT models. Niehues et al. show that the attention model is capable of learning the connections between the pre-translated part and the original source language input to jointly influence the generated target language translations. The approach is straightforward and interesting because it may improve the faithfulness (or adequacy) of the translation engine, which can be a problem in neural encoder–decoder models.

One challenge is that training data has to be translated completely to make it possible to learn the final NMT model. One of the problems discussed by Niehues et al. is the issue of overfitting to the SMT-based translation if the SMT model is trained on the same data set as will be used for learning the NMT model. They propose to weaken the phrase table by removing longer segments and, hence, reducing the capacity of the SMT model to create very generic translation options.

In our setup, we use a different strategy: Instead of using the same data sets for training and translating, we use the backtranslated news data to train a model that can be used to translate the parallel WMT data (Europarl and Rapid2016). With this, we get the same domain-mismatch as during test time with a realistically weak model that avoids over-trusting its capacity when training the NMT model in pre-translated data. Furthermore, we use a WMT-model trained on Europarl and Rapid2016 to translate the backtranslated news data from English back to Finnish again. The latter may be a problem because of the significant noise added due to the double backtranslation but we do not want to discard the important news data completely.

Another difference in our setup is that we use BPE-segmented SMT models to obtain segmented output that we can use directly to be concatenated with the original (BPE-segmented) source. We mark the pre-translated part with a special suffix and then train a standard attention-based NMT model. We use similar parameters as for our standard NMT experiments: 256-dimensional word embeddings, encoder states and attentions, 512-dimensional decoder states, and a vocabulary of

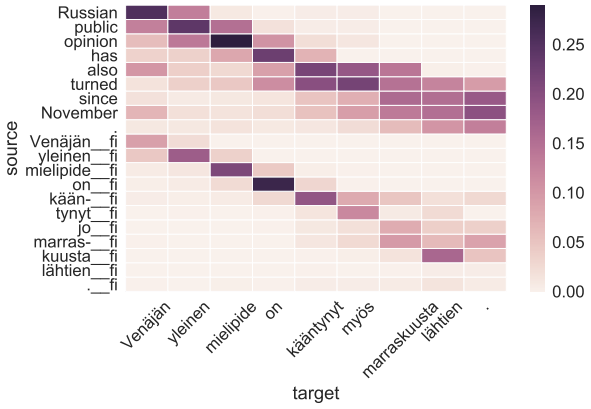

Figure 1: Attention with pre-translated data.

50,000 in source and target language. It turns out that, indeed, the model learns to look at both the source language and the pre-translated text, as we can see in the attention plot in Figure 1.

Unfortunately, the training process is very slow due to the extended input sequence and, hence, converges very slowly. No useful model could be submitted before the official deadline. Our final system tested after the official submission is an ensemble of four independently trained models with savepoint averaging over the last four savepoints and reaches a lowercased BLEU score of 17.34% on newsdev 2015 and 20.92% on newstest 2017 in our internal evaluations (but only 19.8% BLEU in the official on-line system). Even tough this looks quite encouraging compared to the SMT scores, it is still below the plain NMT models, which is, of course, a bit disappointing. However, the results are not directly comparable and there is some variation that needs to be accounted for. More detailed analyses are required to study the possible contributions by the pre-translations. Further investigations of attention plots may reveal whether the model still overfits to the SMT output, which could be a good reason why it underperforms in the end. The additional complexity and the increased length of the input sequences are certainly other reasons for the negative outcome. It also seems that the strong performance of the NMT model also with respect to adequacy make it difficult to improve it further with a weaker SMT model.

3.5 Manual Evaluation

The outputs of the best SMT and NMT systems were partially reviewed and compared by a profes-

sional translator. The impression of the reviewer was that the perceived quality of NMT far exceeds that of SMT, mainly due to the superior fluency of NMT. The BLEU scores of the systems also indicate a significant quality difference in favor of NMT. However, single-reference BLEU scores are known to be unreliable indicators of quality for morphologically complex languages (Bojar et al., 2010), and they are also known to favor SMT over other MT methods (Callison-Burch et al., 2006). Due to this, it is possible that the BLEU scores, impressive as they are, do not reflect the real qualitative impact of NMT for English–Finnish MT.

To explore whether single-reference evaluation underestimates NMT quality, a sample of 68 sentences was extracted from the test set. Both SMT and NMT translations of the sample were postedited with minimal changes to the same quality level as the reference translation. The minimally edited MT was then used as a TER reference to obtain a more reliable estimate of the MT quality. The sample was chosen from sentences where SMT has a sentence-level TER that is at least 10 points lower than the corresponding NMT TER, since such differences can indicate evaluation errors. The sample was also restricted to sentences with an SMT TER lower than 40 to reduce postediting workload and filter out low-quality MT.

When postedited MT was used as a reference, total TER/BLEU for the sample changed from 24.7/50.2 to 12.5/76.0 for SMT and from 48.4/25.0 to 18.3/70.5 for NMT. While the score improved for both SMT and NMT, the improvement is clearly much larger for NMT. The test was then repeated for another sample of 68 sentences from the test set, this time selected from the sentences where NMT had lower sentence-level TER. The purpose of this sample was to see if evaluation errors affect single-reference scores for SMT to the same extent as for NMT. With the second sample, total TER/BLEU changed from 58.9/22.1 to 42.5/39.3 for SMT and from 28.2/48.5 to 12.1/77.01 for NMT, so the result was even more favorable for NMT. While the sample size was small, these results strongly suggest that single-reference BLEU scores indeed underestimate NMT quality.

4 English–Latvian

Training models for English–Latvian was a rather spontaneous decision and we did not spend a lot of time optimizing our settings. Backtranslations were produced with simplistic Latvian–English models. We used a quickly trained character-level NMT model to translate Latvian news data from 2016 and a standard phrase-based SMT model to translate parts of 2014-2016 news data. The statistics of the backtranslations are given in Table 6.

SMT	Sentences	Latvian	English
news2014	330,152	6,469,914	7,611,259
news2015	330,644	6,484,318	7,624,202
news2016	313,180	6,161,332	7,239,953
NMT	Sentences	Latvian	English
news2016	2,059,647	33,447,392	45,262,908

Table 6: Backtranslated Latvian news data using SMT and NMT.

4.1 NMT Models

We submitted one NMT system that follows the basic BPE-decoder system for English–Finnish in terms of model size and training settings. It is trained on the preprocessed versions of the parallel data and the NMT-based backtranslations. This system yielded a case-sensitive BLEU score of 16.8. We again applied hyphen retokenization as a postprocessing step, although it was less useful here than for Finnish (+0.1 BLEU). Again, we trained four independent models and used savepoint-averaging. For time reasons and given the low impact of forward-backward reranking observed for Finnish, we refrained from submitting such a system for English–Latvian.

4.2 SMT Baselines

The SMT models we trained use the provided data sets for training translation models and language models (including a second language model based on common crawl data) with the same tools as for our English–Finnish systems. We applied BPE to all data sets again with a rather fine-grained segmentation into 20,000 types on training data. Table 7 summarizes the results of our models on the *newstest* data from 2017.

We can see that the backtranslated data sets do not work very well in the Latvian case. A small improvement can be observed when combined with the provided training data but the quality of the backtranslations is too poor to have a strong impact on translation quality.

newstest 2017	BLEU
SMT WMT	13.29
SMT back	11.94
SMT WMT+back	13.74
SMT official score (WMT+back)	14.7
NMT official score (WMT+back)	17.3

Table 7: Statistical MT for English–Latvian tested on newstest 2017 (lowercased BLEU). The *official* score in the on-line evaluation system (lowercased) is surprisingly different from our own evaluations. The manual evaluation for English–Latvian produced no statistically significant ranking.

5 English–Chinese and Chinese–English

For English/Chinese, we performed experiments with the HNMT system only. We trained both English–Chinese and Chinese–English models, using all of the available parallel training data from the WMT/CWMT news translation task. After cleaning, 24,954,952 sentence pairs remained. Using the standard Moses tools, we tokenized and truecased the English data. Two methods were used for Chinese word segmentation, as detailed below.

All the models are trained by a hybrid character–word level encoder and a character-level decoder. The final submissions are generated by ensembles with parameters averaging. The official BLEU scores of these two tasks are shown in Table 8. The manual evaluation ranked our system in a shared last place (shared with four other systems) for Chinese–English, while it was ranked #9 (better than two unconstrained online systems) for English–Chinese.

5.1 Translating Chinese into English

Chinese is a language without word boundaries, so word segmentation is necessary before using our hybrid encoder with Chinese source sentences. There are different segmentation methods at different granularities, and they will lead to different translations. In the work of Su et al. (2017), they proposed a lattice-based recurrent encoder which applied three segmentations at different granularities (from the CTB, PKU and MSRA corpora). In our model, we just tried two segmentations: One is a fine-grained method implemented in Zpar (Zhang and Clark, 2011), the other is a coarse-granularity method by THULAC (Sun et al., 2016). The model with THULAC segmentation achieved a slightly lower BLEU score compared to the model with Zpar segmentation. Thus, we did not train more models on THULAC segmentation data after 6-day training. Unlike our results with English–Finnish translation, our experiments with BPE using a 30,000 size vocabulary did not yield any improvements.

newstest 2017	BLEU
English–Chinese	23.9
Chinese–English	15.9

Table 8: HNMT official results on English–Chinese language pair news translation task.

The final submission uses Zpar for segmentation, a hybrid encoder with 60,000 item vocabulary, and a character-level decoder. We use dimensionalities of 256 for both word and character embedding, encoder LSTM and attention. The decoder uses an LSTM of size 512. We use a single model with parameters averaged from savepoints at 6, 9, 10, 12 and 14 days to generate the final submission. This is a rather unusual setup and different from the Finnish and Latvian submissions, but it shows parameter averaging works even when days have passed between savepoints. The beam size in the decoding is set to 10.

5.2 Translating English into Chinese

In addition to translating English into Chinese orthography (using Chinese characters, Hanzi), we also explored translating into romanized Chinese (using the Pinyin system), and then disambiguating the Pinyin to Hanzi with a 3-gram language model. This reduces the vocabulary to the circa 1300 syllables in Standard Mandarin. However, the final disambiguation step introduces new errors that were not outweighed by the easier task of predicting Pinyin output, and we did not pursue this method.

For our official submission, we used a hybrid encoder with 50,000 vocabulary size, and a character-level decoder. Again, we used a single model with parameters averaged from savepoints at 6, 7.5 and 11.5 days.

6 Conclusions

This paper introduces the Helsinki Neural Machine Translation system (HNMT) and its succesful application to the news translation task in WMT 2017. The models we trained handle well the translation into morphologically complex lan-

guages such as Finnish and our submission scored best among the participants in the English–Finnish task. The evaluations show that the neural models are superior to the strong SMT baselines that exploit the same tricks such as backtranslated data and automatic word segmentation. Manual inspections suggest that the advantage of NMT is even underestimated by single-reference BLEU scores. We also applied our models to English–Latvian and English–Chinese (in both directions) with a more moderate success. This is not very surprising for Latvian, for which we only invested about a week to set up the experiments and to train the models. For Chinese, manual evaluation will be important to judge the outcome of our systems fairly.

Acknowledgments

We wish to thank the anonymous reviewers, one of whom provided exceptionally thorough comments. The Finnish IT Center for Science (CSC) provided computational resources. We would also like to acknowledge the support by NVIDIA and their GPU grant. Gongbo Tang is supported by China Scholarship Council (No. 201607110016).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv e-prints* .

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

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling sparse data issue in machine translation evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, Uppsala, Sweden, pages 86–91. http://www.aclweb.org/anthology/P10-2016.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *In EACL*. pages 249–256.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., pages 1019–1027. http://papers.nips.cc/paper/6241-a-theoretically-grounded-application-of-dropout-in-recurrent-neural-networks.pdf.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, UK, pages 187–197.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*.

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 *ACL'07: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demonstration Session*. Association for Computational Linguistics, pages 177–180.

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

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 1828–1836. http://aclweb.org/anthology/C16-1172.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics* 106:125–146. http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf.

Tommi A Pirinen. 2015. Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. *SKY Journal of Linguistics* 28:381–393.

Vıctor M Sánchez-Cartagena and Antonio Toral. 2016. Abu-MaTran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 371–376. http://www.aclweb.org/anthology/W16-2323.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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. 2016c. 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.

Jinsong Su, Zhixing Tan, Deyi Xiong, Rongrong Ji, Xiaodong Shi, and Yang Liu. 2017. Lattice-based recurrent neural network encoders for neural machine translation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. pages 3302–3308.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. Thulac: An efficient lexical analyzer for chinese .

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints* abs/1605.02688. http://arxiv.org/abs/1605.02688.

Jörg Tiedemann, Fabienne Cap, Jenna Kanerva, Filip Ginter, Sara Stymne, Robert Östling, and Marion Weller-Di Marco. 2016. Phrase-based SMT for finnish with more data, better models and alternative alignment and translation tools. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 391–398. http://www.aclweb.org/anthology/W16-2326.

Jörg Tiedemann, Filip Ginter, and Jenna Kanerva. 2015. Morphological segmentation and opus for finnish-english machine translation. In *WMT'15: Proceedings of the Tenth Workshop on Statistical Machine Translation*. pages 177–183.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics* 5:87–99. https://www.transacl.org/ojs/index.php/tacl/article/view/948.

Yonghui Wu et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144. http://arxiv.org/abs/1609.08144.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics* 37:105–151.