CUNI Submission in WMT17: Chimera Goes Neural

Roman Sudarikov **David Mareček**
Tom Kocmi **Dušan Variš** **Ondřej Bojar**

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
surname@ufal.mff.cuni.cz

Abstract

This paper describes the neural and phrase-based machine translation systems submitted by CUNI to English-Czech News Translation Task of WMT17. We experiment with synthetic data for training and try several system combination techniques, both neural and phrase-based. Our primary submission CU-CHIMERA ends up being phrase-based backbone which incorporates neural and deep-syntactic candidate translations.

1 Introduction

The paper describes CUNI submissions for English-to-Czech WMT 2017 News Translation Task. We experimented with several neural machine translation (NMT) systems and we further developed our phrase-based statistical machine translation system Chimera, which was our primary system last year (Tamchyna et al., 2016).

This year, we planned our setup in a way that would allow us to experiment with neural system combination. To this end, we reserved the provided English-Czech parallel data for the training of the system combination and trained our "individual forward systems" on *almost only synthetic data*.

The structure of the paper is the following. In Section 2, we provide an overview of the relatively complex setup. Section 3 details how the training data for all the systems were prepared, including the description of MT systems used for back-translation. Section 4 is devoted to our individual forward translation systems, each of which could actually serve as a submission to the translation task. We do not stop there and train system combinations in Section 5. In Section 6, we present the systems we actually submitted to WMT17 and we conclude by Section 8.

2 Setup Overview

Our setup this year is motivated by the ability to use all the parallel data for system combination training. The overall sequence of system training is the following:

1. Use available monolingual data and last year's systems to prepare a synthetic parallel corpus using "back translation" (Section 3).
2. Train "individual forward systems" on this synthetic corpus (Section 4).
3. Apply individual forward systems to the source side of the genuine parallel data.
4. Train a (neural) system combination on this dataset (Section 5).
5. Apply individual forward systems to the test set and apply the trained combination system to their output (Section 5).

Each of the steps is fully described in the respective section of this paper. By "back-translated" data we mean that for English-to-Czech translation task, we created a synthetic English-Czech parallel corpus by "back-translating" Czech monolingual data into English. To distinguish back-translation Czech-to-English systems and the English-to-Czech systems to be submitted, we will call Czech-to-English systems "back-translation systems" and English-to-Czech systems "forward(-translation) systems".

3 Data Preparation

The section describes the data used for training of both Czech-to-English back-translation systems as well as English-to-Czech forward systems.

Corpus	Sentences	Tokens Cs	Tokens En
Synthetic corpora			
NematusNews	59 190 187	985 887k	1 196 366k
MosesNews	59 146 101	985 017k	1 173 839k
XenC extracted corpora			
XenCNews	20 415 268	289 472k	334 322k
XenCMonoNews	12 498 680	95 687k	103 193k
Development corpora			
Dev	2 656	46k	55k
Eval	2 999	57k	67k

Table 1: Datasets

3.1 Back-Translated Data

To create back-translated data, we used the CzEng 1.6 Czech-English parallel corpus (Bojar et al., 2016) and the Czech News Crawl articles released for WMT2017[1] (called "mononews" for short).

We used two different back-translation systems: Moses (Koehn et al., 2007) trained by ourselves, and Marian[2] (known as AmuNMT before it included NMT training; Junczys-Dowmunt et al., 2016) using the pretrained Nematus (Sennrich et al., 2017) models[3] from WMT16 News Task.[4]

We used only the non-ensembled left-to-right run (i.e. no right-to-left rescoring as done by Sennrich et al., 2016a) with beam size of 5,[5] taking just the single-best output.

The Moses-based system used only a single phrase table translating from word form to word forms and twelve 10-gram language models built on individual years of English mononews.

We took all Czech mononews corpora available this year, concatenated and translated them using both systems described above and thus created two back-translated corpora on which we planned to train our forward systems.

The "Synthetic corpora" section of Table 1 shows the numbers of sentences and tokens of the resulting corpora. Despite having started with the same Czech monolingual corpus, the number of sentences differs slightly due to minor technical issues encountered by Moses.

In the following, the synthetic corpora created by the two MT systems will be referred to as NematusNews and MosesNews, respectively.

3.2 Domain-Selected Genuine Parallel Data

For the training of forward translation systems, we used primarily the synthetic corpora described in Section 3.1 above but also some additional sources described in this section.

The first source to mention is CzEng 1.6. We did not use the whole corpus as we did in our WMT16 submission (Tamchyna et al., 2016). Instead, we used the XenC toolkit (Rousseau, 2013) to extract domain-specific data from the whole corpus (referred to as "out-of-domain", in the following). We used two modes of XenC. Both of these modes estimate two language models from in-domain and out-of-domain corpora, using SRILM toolkit (Stolcke, 2002). The first mode is a filtering process based on a simple perplexity computation utilizing only one side of the corpora so that monolingual corpora are sufficient and the second mode is based on the bilingual cross-entropy difference as described by Axelrod et al. (2011).

We took two different corpora as our in-domain data:

- News section of CzEng 1.6 – which had 197 053 parallel English-Czech sentences. The extraction was performed both monolingually (perplexity) and bilingually (bilingual cross-entropy difference).

- Concatenated mononews corpora – which had 59 190 187 Czech sentences. The extraction was performed only monolingually.

The two different in-domain corpora were used because we wanted to estimate which of them would lead to better extracted corpus – a small parallel in-domain corpus or a larger monolingual corpus.

Based on these two representatives of in-domain texts, we extracted sentences from CzEng 1.6. We took top 20% of sentence pairs extracted monolingually (see XenCMonoNews in the section "XenC extracted corpora" in Table 1) and top 20% of sentence pairs extracted monolingually and bilingually (see XenCNews) in the same table. For XenCNews corpus monolingual and bilingual sentence extractions were made separately and then the results were unioned, i.e. concatenated and duplicates removed.

For the development and evaluation purposes, we used WMT2015 and WMT2016 test sets, re-

[1] `http://www.statmt.org/wmt17/translation-task.html`

[2] `https://github.com/marian-nmt/marian`

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

[4] We decided to use Marian instead of Nematus since it was faster at the time we performed the translation.

[5] We chose beam size of 5, since our primary goal was to produce a 5-best list.

spectively, see the "Development corpora" section in Table 1.

Finally, what we are combining, are the outputs of several forward translation systems: Nematus, Neural Monkey and TectoMT. During the development, we used the outputs of these systems on the test sets of WMT 2015 and 2016. For the test run, we translated the source of WMT news test set 2017.

All the corpora were tokenized using MorphoDita (Straková et al., 2014), i.e. even for synthetic corpora and combined systems, we deBPE'd and detokenized the MT outputs and retokenized them.

4 Individual Forward Systems

This section describes our English-to-Czech systems. Each of them could be submitted to WMT17 but we combine them into just one system, see Section 5 below.

4.1 Baseline Nematus

We used Marian (formerly known as AmuNMT) (Junczys-Dowmunt et al., 2016) with pretrained English-to-Czech Nematus models[6] from WMT16 News Task as our baseline/benchmark and we also later included it in the final combined submission.

We used only the non-ensembled left-to-right run (i.e. no right-to-left rescoring as done by Sennrich et al., 2016a) with beam size of 12 (default value).

4.2 Neural Monkey

We use Neural Monkey[7] (Helcl and Libovický, 2017), an open-source neural machine translation and general sequence-to-sequence learning toolkit built using the TensorFlow machine learning library.

Neural Monkey is flexible in model configuration but for forward translation, we restrict our experiments to the standard encoder-decoder architecture with attention as preposed by Bahdanau et al. (2015). (Attempts to combine MT systems with Neural Monkey are described in Section 5.2 below.) We use the following model parameters which fit into 8GB GPU memory of NVIDIA GeForce GTX 1080. The encoder uses embeddings of size 600 and the hidden state of size 600. Dropout is turned off[8] and maximum input sentence length is set to 50 tokens. The decoder uses attention mechanism and conditional GRU cells (Firat and Cho, 2016), with the hidden state of 600. Output embedding has the size of 600, dropout is turned off as well and the maximum output length is again 50 tokens. We use batch size of 60.

To reduce vocabulary size, we use byte pair encoding (Sennrich et al., 2016c) which breaks the all words into subword units defined in the vocabulary. The vocabulary is initialized with all letters and larger units are added on the basis of corpus statistics. Frequent words make it to the vocabulary, less frequent words are (deterministically) broken into smaller units from the vocabulary.

We set the vocabulary size to 30,000 subword units. The vocabulary is constructed jointly for the source and target side of the corpus and it is then shared between encoder and decoder.

During the inference, we use either greedy decoding or beam search with beam size of 50.[9]

4.3 Chimera 2016

The last individual forward system was based on CUNI's last year submission (Tamchyna et al., 2016). We experimented with several setups, see the list in Table 2.

Chimera itself is a hybrid system combination and we used the technique both here as an individual system as well as below in Section 5.3 for our final system combination.

The main components of the individual Chimera system are:

- **Synthetic phrase table** extracted from the main training data, ie. either or both of NematusNews and MosesNews as listed in Table 1.
- **In-domain phrase table** extracted from either or both of XenCNews and XenC-MonoNews.
- **Operation Sequence Model** (Durrani et al., 2013) trained on the NematusNews corpus.

[6] `http://data.statmt.org/rsennrich/wmt16_systems`

[7] `http://ufal.mff.cuni.cz/neuralmonkey`

[8] While dropout is useful for small datasets, Sennrich et al. (2016b) observed no gain from dropout with 8M training sentence pairs. Our training data is more than 7× larger.

[9] In contrast to what Tu et al. (2017, Table 1) observe for other implementations of the Bahdanau et al. (2015) model, Neural Monkey does not exhibit degradation of the quality of the top candidate with increasing beam size. We have thus no reason to keep beam size as small as usual.

	Phrase Tables	Additional	BLEU	Avg. BLEU
1.	XenCNews + TectoMT	-	20.88	-
2.	XenCMonoNews + TectoMT	-	20.08	-
3.	NematusNews	OSM	20.60	-
4.	MosesNews + TectoMT	-	20.79	-
5.	Mix(NematusNews, XenCNews) + TectoMT	-	21.60	-
6.	Mix(NematusNews, XenCMonoNews) + TectoMT	OSM	21.70	21.6
7.	Mix(NematusNews, XenCMonoNews) + TectoMT	-	21.87	21.7
8.	Mix(MosesNews, XenCNews) + TectoMT	-	21.30	-
9.	Mix(MosesNews, XenCMonoNews) + TectoMT	-	20.96	-
10.	Mix(MosesNews, NematusNews) + TectoMT	-	21.67	-
11.	Mix(MosesNews, NematusNews, XenCMonoNews) + TectoMT	-	21.52	-
12.	Mix(Moses, Nematus, XenCMonoNews, XenCNews) + TectoMT	-	21.81	-
CHIMERA-TECTOMT-DEPFIX (secondary submission)				
Mix(NematusNews, XenCMonoNews) + TectoMT		-	21.65	21.8

Table 2: Chimera-style combinations of various individual forward systems on WMT 2016 News.

- **TectoMT phrase table** (Žabokrtský et al., 2008) – a phrase table extracted from the outputs of TectoMT, a transfer-based deep-syntactic system, applied to the source side of the development and test sets.

The common components for all the tested systems are language models, which were taken from CUNI's last year submission. For some experiments we have used up to 4 phrase tables separately as Moses alternative decoding paths, trusting MERT (Och, 2003) to estimate weights. Alternatively (or when the number of the phrase tables would be even higher), we used the standard Moses phrase table mixing technique with uniform weights. Phrase tables mixed into one before MERT are listed as "Mix($table1, table2, ...$)" in the following.

MERT was done using the WMT2015 test set, and our internal evaluation was performed on WMT2016 test set, but with a different tokenization so the scores reported here are not directly comparable to the results at `http://matrix.statmt.org/`.

We report the results in Table 2, listing the used phrase tables and optionally OSM. The column "Average BLEU" was calculated based on 5 separate MERT runs.

It seems that training only on (in-domain) synthetic data is a viable option, lines 3 and 4 in Table 2 perform reasonably good and mixing the two sources of the synthetic data into one phrase table (line 10) instead of using the two of them simultaneously lead to an improvement of almost 1 BLEU point. At the same time, genuine parallel (and again in-domain) training data is equally good as each of the synthetic corpus, even if much smaller, see lines 1 and 2 trained on up to 20M sentence pairs instead of 59M synthetic sentences. Selecting the genuine parallel sentences both bilingually and monolingually (XenCNews) works usually better than selecting them only monolingually (XenCMonoNews), but there is a significant difference in corpus size so the numbers are not directly comparable.

The best-performing setup used the synthetic corpus created by Nematus (NematusNews), the (suprisingly) monolingually selected genuine parallel data (XenCMonoNews) and TectoMT (line 7 in Table 2). We used this setup as our main phrase-based translation system and also submitted is as a contrastive system under the name CHIMERA-TECTOMT-DEPFIX. Difference between line 7 and submitted system is in the TectoMT phrase table – line 7 system had TectoMT phrase-table without WMT 2017 test set, because internal evaluation was performed prior to the release of this test set.

5 Forward System Combination

This sections describes our experiments with system combination. We tried two neural and one Chimera-style approach.

As described in Section 3, the genuine parallel training data from CzEng was not directly used for the training of the forward systems (except for Chimera) so we could use this data to train our neural combination systems. We again opted to use only domain-specific part of CzEng, so we trained the systems on XenCNews as listed in Table 1.

5.1 Concatenative Neural System Combination

We experiment with system combination made by simple concatenation of individual system outputs together, inspired by Niehues et al. (2016).

To train the neural combination system, we create a synthetic parallel corpus with the following three sentences on the source side:

- Nematus English-to-Czech translation
- Neural Monkey English-to-Czech translation
- English source sentence

The sentence triples are concatenated with spaces between them, forming a single input string of tokens. The target side remains the same, i.e. a single Czech target sentence. As shown by Niehues et al. (2016), the attention mechanism is capable of synchronously following the source and one candidate translation, so we hoped it could follow two candidate translations as well (with the obvious complication due to much longer input sequences).

The translation system trained on such data might benefit from distinguishing the words based on the translation system they come from. We therefore add labels in form of prefixes to each the token to identify the originating the system (*n-* for Nematus output, *m-* for Neural Monkey, and *s-* for the English source).

We perform three experiments:

1. without labels,
2. with labels inserted before BPE splitting, which means that only the first part of individual tokens has the prefix,
3. with labels inserted after BPE splitting.

For training, we use Nematus NMT system (Sennrich et al., 2017), using shared vocabulary of size 50,000, RNN size 1024, embedding size 500, and batch size 80. The maximum sentence length is tripled to 150, instead of standard value of 50.

The results are in Table 3. It is obvious that the additional labels do not help. The best results were achieved without using labels and more labels worsen the final BLEU score. However, the concatenative system combination did not bring any improvement over the individual systems, it is worse than the best single system Nematus by 3 BLEU points. This was partially caused by too short training time (about one week, 420,000 iterations, batch size 80).

System	BLEU
Nematus	24.4
Neural Monkey	22.9
combination without labels	21.4
combination labelled before BPE	21.2
combination labelled after BPE	20.4

Table 3: Concatenative combination BLEU scores on WMT2016 News and comparison with the single systems.

We inspected the attention scores and confirmed that the decoder used all three sentences, however it prefers the Nematus translation and the English source sentence. It pays less attention to the Neural Monkey translation, which is understandable since the translation quality is lower.

5.2 Neural Monkey System Combination

Neural Monkey supports multiple encoders and a hierarchical attention model (Libovický et al., 2016). Due to time constraints, we did not finish these experiments for WMT17 but the work is still in progress.

The idea is to use a separate encoder for each input sentence and to combine their outputs before passing them to the target sentence decoder. The final encoder states are simply concatenated (and optionally resized by a linear layer) and the hidden states are all passed to the decoder for attention computation without distinguishing which encoder generated them. Libovický and Helcl (2017) suggest also other strategies for combining attention from multiple source encoders and we plan to further investigate them in the near future.

Since we are trying to combine outputs generated by Nematus and Neural Monkey, both trained on subword units, we decided to try a character-to-character architecture as introduced in Lee et al. (2016) for system combination, expecting better results due to differences in the used architectures. In the future, we also plan comparing this approach to the subword-level multi-encoder system combination.

We trained a baseline model using GeForce GTX 1080 with 8GB memory. We used a shared vocabulary of size 500 for all encoders and decoder. We used RNN size 256 and embedding 300 for each encoder, highway depth of 2 and set of convolutional filters scaled down to fit the smaller memory and taking multiple encoders into account. The decoder RNN size was 512 and used embedding size 500. We trained the model for 10 days and obtained the BLEU score of 14.69 on the newstest2016 EN-CS development set. This is much lower than the individual combined systems.

The system performed poorly overall and we have to investigate whether the main reason for the failure is the character-to-character approach, the multi-encoder architecture, their combination, or simply some bugs in implementation. Further experiments are planned for the future to be able to draw better conclusions.

5.3 Chimera System Combination

Given the poor performance of our neural system combinations, we decided to try the same Chimera-style combination with all available systems, i.e. Nematus, Neural Monkey and Chimera 2016 described in Section 4.

We took the best phrase tables combination from Section 4.3: (1) A combination of mixed NematusNews and XenCMonoNews phrase table (called simply "Moses" in Table 4 because it is the phrase-based basis of the system), (2) phrase table generated from TectoMT output and (3) tried to add phrase tables extracted from Nematus and Neural Monkey translations of WMT2015–2017 test sets.

For Neural Monkey, we had several setups to extract phrase tables from:

- Neural Monkey – the output of the system described in Section 4.2 using greedy decoding,
- Neural Monkey 1 – decoding with beam search of 50 and taking only the first candidate translation to the phrase table,
- Neural Monkey 50 – decoding with beam search of 50 and taking all 50 candidate translations to the phrase table,

All combinations we have experimented with are shown in Table 4. The last column "Average BLEU" was calculated the same way as it was done in Section 4.3. Also the same 5 MERT runs were used for MultEval evaluation (Clark et al., 2011).

Basically, Table 4 confirms the well-know saying "more data helps". Using translations from different systems as additional phrase tables gave on average a 2.5 BLEU score boost, if we compare rows 1 or 2 and row 14.

We also see that using more than three phrase tables might lead to a lower BLEU score: Consider the system in the row 7 with four separate phrase tables (Avg. BLEU 23.7) and the system in the row 3 where three of the tables were first merged into one (Avg. BLEU 23.9). Moreover, Multeval comparison showed no significant difference between systems from rows 7 and 8, despite the effect of adding TectoMT table is generally positive. When TectoMT is added as the fourth table, MERT can probably no longer optimize the system to benefit from it.

We selected the system combination with Neural Monkey 50 as our primary submission (Avg. BLEU 24.1), because we believed, that it would be beneficial to have more translation variants. Unfortunately, we found only later that MultEval indicates a significant difference between systems from rows 1 and 2, supporting the single-best output of Neural Monkey (Avg. BLEU 24.3).

6 Results and Discussion

Our submitted systems are shown in Table 5. Depfix (Rosa et al., 2012) was applied only for the final submission. Scores in the last column are BLEU-cased evaluation results taken from `https://matrix.statmt.org`.

It is interesting to notice that Neural Monkey trained only on synthetic dataset preformed better than Moses trained on synthetic dataset with additional in-domain data.

One point of further investigation is to find out whether the combination of Moses and Neural Monkey is better because Moses provided some useful phrases or because it merely re-ranked Neural Monkey results of beam search output.

The next point is to experiment with mixing phrase tables techniques, examining e.g. non-uniform weights.

Table 6 displays the official results of English-to-Czech translation. We see that our CU-Chimera was second in terms of BLEU (20.5) and shared

	Tables	BLEU	Avg. BLEU
1.	**Moses + Mix(TectoMT, Nematus, Neural Monkey 50) ***	24.11	24.1
2.	Moses + Mix(TectoMT, Nematus, Neural Monkey 1)	24.17	24.3
3.	Moses + Mix(TectoMT, Nematus, Neural Monkey)	23.86	23.9
4.	Moses + TectoMT + Mix(Nematus, Neural Monkey)	23.82	23.9
5.	Moses + Neural Monkey + Mix(TectoMT, Nematus)	23.75	23.7
6.	Moses + Nematus + Mix(TectoMT, Neural Monkey)	23.87	23.9
7.	Moses + Nematus + Neural Monkey + TectoMT	23.82	23.7
8.	Moses + Nematus + Neural Monkey	23.82	23.7
9.	Moses + TectoMT + Nematus	23.57	-
10.	**Moses + TectoMT + Neural Monkey 50**	23.57	-
11.	Moses + TectoMT + Neural Monkey 1	23.36	-
12.	Moses + TectoMT + Neural Monkey	22.96	22.9
13.	Moses + Neural Monkey	22.43	22.4
14.	**Moses + TectoMT**	21.65	21.8

Table 4: Chimera system combination evaluation on WMT 2016 News. Submitted systems in **bold**, with the primary marked with *.

Systems	Depfix	News2017
Moses+TectoTM+Neural Monkey 50+Nematus *	+	20.5
Moses+TectoTM+Neural Monkey 1+Nematus	+	20.4
Moses+TectoTM+Neural Monkey 50	+	19.9
Neural Monkey 1	-	19.3
Moses+TectoTM	+	18.3

Table 5: Submitted systems comparison. Asterisk (*) denotes our primary submission, CU-Chimera.

the second position with limsi-factored-norm in terms of TER (0.696) but considerably lost in manual evaluation, sharing the third rank with four other systems. For us, this confirms that BLEU overvalues short sequences that our phrase-based backbone of CU-Chimera was good at.

To summarize our results, we were able to considerably improve over our setup from the last year by adding the outputs of NMT to our strong combined system. Unfortunately, we failed in implementing *neural* system combination, mainly due to technical difficulties, and our final system thus suffers from the well-known limitations of PBMT.

7 Related Work

The idea of combining phrase-based and neural systems is not novel. Our concatenative approach follows Niehues et al. (2016) who saw PBMT as a pre-processing step and added the output of PBMT to the input of NMT system, obtaining improvements over a good-performing NMT ensemble of more than 1 BLEU for two different test sets for English-German translation.

Cho et al. (2016) use a weaker approach to system combination, mixing n-best lists of several variations of NMT systems (including those that already included PBMT output)

The multi-encoder approach we describe in Section 5.2 was very recently successfully applied by Zhou et al. (2017). The main difference in the application is that we tried to use character-level encoders instead of standard sub-word units, which was clearly overly ambitious given our limited computing and time resources.

8 Conclusion

In the paper, we presented our experiments with both phrase-based and neural approaches to machine translation.

Our results document that synthetic datasets can be nearly as good as genuine in-domain parallel data.

We experimented with three different approaches to MT system combination: two neural ones and one phrase-based. Due to time and resource limitations, we were not successful with

#	Ave %	Ave z	BLEU	TER	CharacTER	BEER	System
1	62.0	0.308	22.8	0.667	0.588	0.540	uedin-nmt
2	59.7	0.240	20.1	0.703	0.612	0.519	online-B
3	55.9	0.111	20.2	0.696	0.607	0.524	limsi-factored-norm
	55.2	0.102	20.0	0.699	-	-	LIUM-FNMT
	55.2	0.090	20.2	0.701	0.605	0.522	LIUM-NMT
	54.1	0.050	20.5	0.696	0.624	0.523	CU-Chimera
	53.3	0.029	16.6	0.743	0.637	0.503	online-A
8	41.9	-0.327	16.2	0.757	0.697	0.485	PJATK

Table 6: Official results for English-to-Czech primary systems.

the neural approaches, although there are good reasons (and new evidence) that they were very promising.

CU-Chimera, our primary submission to the WMT17 News Translation Task ends up being a phrase-based backbone which includes neural and deep-syntactic candidate translations.

Acknowledgments

This work has been in part supported by the European Union's Horizon 2020 research and innovation programme under grant agreements No 644402 (HimL) and 645452 (QT21), by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16_013/0001781), by the Charles University Research Programme "Progres" Q18+Q48, by the Charles University SVV project number 260 453 and by the grant GAUK 8502/2016.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the Third International Conference on Learning Representations (ICLR 2015)*. http://arxiv.org/abs/1409.0473.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovickỳ, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered. In *International Conference on Text, Speech, and Dialogue*. Springer, pages 231–238.

Eunah Cho, Jan Niehues, Thanh-Le Ha, Matthias Sperber, Mohammed Mediani, and Alex Waibel. 2016. Adaptation and Combination of NMT Systems: The KIT Translation Systems for IWSLT 2016. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)-To be appeared*.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 176–181.

Nadir Durrani, Alexander M Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. Can markov models over minimal translation units help phrase-based smt? In *ACL (2)*. pages 399–405.

Orhan Firat and Kyunghyun Cho. 2016. Conditional Gated Recurrent Unit with Attention Mechanism. https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf. Published online, version `adbaeea`.

Jindřich Helcl and Jindřich Libovický. 2017. Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics* 107:5–17. https://doi.org/10.1515/pralin-2017-0001.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, WA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pages 177–180.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *CoRR* abs/1610.03017. http://arxiv.org/abs/1610.03017.

J. Libovický and J. Helcl. 2017. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. *ArXiv e-prints* .

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Pavel Pecina, and Ondiřej Bojar. 2016. CUNI system for WMT16 automatic post-editing and multimodal translation tasks. *CoRR* abs/1606.07481. http://arxiv.org/abs/1606.07481.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-Translation for Neural Machine Translation. *CoRR* abs/1610.05243. http://arxiv.org/abs/1610.05243.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 160–167.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A system for automatic correction of Czech MT outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 362–368.

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

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 65–68. http://aclweb.org/anthology/E17-3017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proc. of the First Conference on Machine Translation (WMT16)*. Berlin, Germany.

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

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*. volume 2, pages 901–904.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *ACL (System Demonstrations)*. pages 13–18.

Aleš Tamchyna, Roman Sudarikov, Ondřej Bojar, and Alexander Fraser. 2016. CUNI-LMU submissions in WMT2016: Chimera constrained and beaten. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*.

Zhaopeng Tu, Yang Liu, Lifeng Shang, and Xiaohua LiuAAAI 2017nd Hang Li. 2017. Neural machine translation with reconstruction. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*. AAAI Press, pages 3097–3103.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 167–170.

Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural system combination for machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 378–384. https://doi.org/10.18653/v1/P17-2060.