The QT21 Combined Machine Translation System for English to Latvian

Jan-Thorsten Peter[1], Hermann Ney[1], Ondřej Bojar[2],
Ngoc-Quan Pham[3], Jan Niehues [3], Alex Waibel[3],
Franck Burlot[4], François Yvon[4], Mārcis Pinnis[5], Valters Šics[5],
Joost Bastings[6], Miguel Rios[6], Wilker Aziz[6],
Philip Williams[7], Frédéric Blain[8], Lucia Specia[8]

[1]RWTH Aachen University, Aachen, Germany
[2]Charles University, Prague, Czech Republic
[3]Karlsruhe Institute of Technology, Karlsruhe, Germany
[4]LIMSI, CNRS, Université Paris Saclay, 91 403 Orsay, France
[5]Tilde, Riga, Latvia
[6]University of Amsterdam, Amsterdam, Netherlands
[7]University of Edinburgh, Edinburgh, UK
[8]University of Sheffield, Sheffield, UK

Abstract

This paper describes the joint submission of the QT21 projects for the English→Latvian translation task of the *EMNLP 2017 Second Conference on Machine Translation* (WMT 2017). The submission is a system combination which combines seven different statistical machine translation systems provided by the different groups.

The systems are combined using either RWTH's system combination approach, or USFD's consensus-based system-selection approach. The final submission shows an improvement of 0.5 BLEU compared to the best single system on newstest2017.

1 Introduction

Quality Translation 21 (QT21) is a European machine translation research project with the aim of substantially improving statistical and machine learning based translation models for challenging languages and low-resource scenarios.

Members of the QT21 project have jointly built a combined statistical machine translation system, in order to achieve high-quality machine translation from English into Latvian.

Core components of the QT21 combined system for the WMT 2017 shared task for machine translation of news [1] are seven individual

---

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

English→Latvian translation engines which have been set up by different project partners.

The outputs of all these individual engines are combined using the system combination approach as implemented in Jane, RWTH's open source statistical machine translation toolkit (Freitag et al., 2014a). The Jane system combination is a mature implementation which previously has been successfully employed in other collaborative projects and for different language pairs (Peter et al., 2016; Freitag et al., 2013, 2014b,c).

As an alternative way of combining our systems, all outputs have been merged as the form of a n-best list and a consensus-based system-selection applied to obtain as best translation hypothesis the candidate that is most similar to the most likely translations amongst those systems.

2 Preprocessing

The training data was pre-processed using a custom language-specific tokeniser and the Moses truecaser (*truecase.perl*). For tokenisation, we used the Tilde's regular expression-based tokeniser for Latvian and English that takes into account language-specific characteristics (e.g., abbreviations, contractions, date, time, and numerical expressions, etc.) and non-translatable entities (e.g., phone numbers, e-mail addresses, XML tags, URLs, file paths, various identifiers and codes, etc.). Only the first word in each sentence was truecased.

The data (backtranslation included) is further cleaned using a simple language identifier from Shuyo (2010). We simply removed sentence pairs whose targets cannot be identified by the

tool. The number of sentences being removed is approximately 50000.

3 Translation Systems

Each group contributed one or more systems. In this section the systems are presented in alphabetic order.

3.1 CUNI

The CUNI component of the system was built using Neural Monkey[2] (Helcl and Libovický, 2017), a flexible sequence-to-sequence toolkit implementing primarily the Bahdanau et al. (2015) model but useful also in multi-modal translation and multi-task training.

We used essentially the baseline setup of the system as released for the WMT17 NMT Training Task[3] (Bojar et al., 2017) for an 8GB GPU card. This involves BPE (Sennrich et al., 2016) with 30k merges, maximum sentence length for both source and target limited to 50 (BPE) tokens, no dropout and embeddings (both source and target) of 600, vocabulary shared between encoder and decoder, attention and conditional GRU (Firat and Cho, 2016). We experimented with the RNN size of the encoder and decoder and increased them to 800 instead of 600, at the expense of reducing batch size to 10. The batch size of 30 with this enlarged model would still fit into our GPU card but this run was prematurely interrupted due to a hardware failure and we noticed that it converges slower in terms of sentence pairs (not in terms of wallclock time), so we opted for a more efficient use of the training data by taking the smaller batch.

We trained on 5245514 sentence pairs mixing the genuine parallel data and synthetic data, as described in Section 2. Neural Monkey does not shuffle the corpus, so we shuffled it beforehand and kept the order identical for all training epochs.

The training ran for 15 days on NVIDIA GeForce GTX 1080 and processed 4.7 epochs but the best model (according to BLEU scores on the development set, "devset-b") was actually reached after 11M sentence pairs (early epoch 3), after 7 days.

Neither ensembling nor beam-search was used for the run, because they were not yet available

in Neural Monkey. Instead, the translations were generated using greedy search.

3.2 KIT

The neural machine translation models from KIT are built with the OpenNMT framework (Klein et al., 2017), which is a multi-layer LSTM encoder decoder network. We trained the models with 2.1 million parallel sentence pairs concatenated with 2.8 million pairs from backtranslation provided by University of Edinburgh. The networks have 1024 hidden units for each of 2 LSTM layers for both encoder and decoder. Furthermore, we experiment a number of features with the baseline:

First, we found out that using a context gate to mask activities between the decoder hidden state and the source context vector before producing the distribution at each time step (Tu et al., 2016a) is simple yet beneficial for performance. Second, we strengthen the attentional network with a coverage vector accumulating the previous attentional information, similar to the work of Mi et al. (2016) and Tu et al. (2016b).

Using the two techniques helps improve the BLEU score on the newsdev2017 set by 1.1 (tokenized) BLEU. By using ensembling 3 networks with different configs and rescoring using a model trained with reversed target sentences, we managed to reach 26.96 BLEU score for the development set, which yields 2.8 point of improvement compared to the baseline model. Details about the effect of each technique is described in Pham et al. (2017)

3.3 LIMSI

LIMSI's intput to this system combination consists of two NMT systems, both trained with the NMTPY framework (Caglayan et al., 2017) on bitext, then on synthetic parallel data. All of them were rescored with a Nematus system (Sennrich et al., 2017b). More details about these systems can be found in (Burlot et al., 2017b,a).

The first system, named baseline, is a BPE-to-BPE system. Bilingual sub-word units (Sennrich et al., 2016) were trained on the bitext parallel data with 90k merge operations. All the parameters of the neural network were initialized with Xavier. The system was optimized with Adam, dropout was enabled on source embeddings, encoder states, as well as output layer. The whole training process took approximately 1.5 months.

Individual Systems	newsdev2017/1				newsdev2017/2				newstest2017			
	BLEU	TER	CTER	BEER	BLEU	TER	CTER	BEER	BLEU	TER	CTER	BEER
Tilde smt	21.3	59.6	57.8	56.5	20.8	61.4	58.5	56.0	15.3	70.6	67.2	51.6
CUNI neuralmonkey	18.9	63.2	62.6	54.8	19.8	63.9	62.1	54.6	13.6	73.2	69.3	50.2
UvA	16.6	70.0	71.9	52.2	16.4	68.8	71.3	51.9	12.0	78.1	76.4	47.9
KIT	26.8	53.4	49.6	60.2	26.8	54.5	49.5	60.0	18.3	66.5	60.5	54.4
UEDIN NMT	25.7	55.1	51.6	59.5	25.6	56.4	50.4	59.4	17.8	67.9	66.7	53.5
UEDIN rescored by KIT	25.9	54.8	52.0	59.6	26.3	55.8	51.3	59.6	17.9	67.7	64.5	53.8
LIMSI factored	24.3	57.3	53.2	58.6	24.8	57.4	52.1	58.5	17.1	69.0	61.5	53.3
System Combination	27.4	53.1	50.9	60.2	27.9	53.9	51.0	59.9	18.8	66.0	67.8	54.3

Table 1: Results of the individual systems for the English→Latvian task. BLEU [%] and TER [%] scores are case-sensitive.

The results shown in Table 1 correspond to an ensemble of our three best models, which produced n-best hypothesis. Finally, these hypothesis were rescored using a Nematus system trained on the same data as the baseline and with similar hyper-parameters.

The second system is an experiment with factored NMT, which is part of the NMTPY framework (García-Martínez et al., 2016). The hyper-parameters mentioned above for the baseline also hold for this system. The specific setup we have used consisted in an architecture that enables training towards a dual objective: at each time-step in the output sentence, a normalized word and a PoS-tag are produced. To obtain the first factor vocabulary, all target words have been normalized (Burlot and Yvon, 2017a), i.e. all grammatical information that is redundant *wrt.* English has been removed from the words. In a nutshell, the normalization system performs a clustering of the morphologically rich language by grouping together words that tend to share the same translation(s) in English. As a result, words are represented by a lemma and a cluster identificator containing the morphological features that have been merged. In our setup, the cluster identificator was systematically split from the lemma. BPE segmentation was thus learnt and applied to lemmas.

Given a lexical unit and a PoS-tag, word forms are retrieved with a dictionary lookup. In the context of morphologically rich languages, deterministic mappings from a lemma and a PoS to a form are very rare. Instead, the dictionary often proposes several word forms corresponding to the same lexical unit and morphological analysis. To address this issue, we let a word-based system select the right word form from the dictionary. To this end, k-best hypothesis from the dictionary were generated, as well as the n-best hypothesis

from the factored NMT system, leading to nk-best rescoring.

Our factored NMT system is an ensemble of two best models and rescoring is performed with our single best Nematus model.

3.4 Tilde

The Tilde system is a Moses phrase-based SMT system that was trained on the Tilde MT platform (Vasiļjevs et al., 2012). The system was trained using all available parallel data - 1.74 million unique sentence pairs after filtering, and 3 million unique sentence pairs that were acquired by re-translating a random selection of in-domain monolingual sentences with a neural machine translation system (Pinnis et al., 2017). The system has a 5-gram language model that was trained using KenLM (Heafield, 2011) on all available monolingual data (27.83 million unique sentences).

3.5 UEDIN

The University of Edinburgh's system is an attentional encoder-decoder (Bahdanau et al., 2015), trained using the Nematus toolkit (Sennrich et al., 2017c).

As training data, we used all parallel and synthetic data, which was tokenized, truecased, and filtered as described in Section 2. After filtering, the data was segmented into subword units using byte-pair-encoding (BPE), for which we used 90,000 operations, jointly learned over both sides of the parallel corpora.

We used word embeddings of size 512 and hidden layers of size 1024, with the size of the source and target network vocabularies fixed to the size of the respective BPE vocabularies. In order to reduce the size of the models, the target-side embedding weights were tied with the transpose of

the output weight matrix (Press and Wolf, 2017). We used a deep transition architecture inspired by the one proposed by Zilly et al. (2016) for language modelling. In experiments conducted during feature development, we found that this gave consistent improvements across multiple language pairs. We also applied layer normalisation (Ba et al., 2016) to all recurrent and feed-forward layers, except for layers that are followed by a softmax. In preliminary experiments, we found that using layer normalisation led to faster convergence and resulted in slightly better performance.

We trained the models with adam (Kingma and Ba, 2015), using a learning rate of 0.0001 and mini-batch size of 80. Training was automatically stopped when the validation cross-entropy failed to reach a new minimum for 10 consecutive save-points (saving every 10000 updates).

For our final system, we trained eight independent models: four left-to-right and four right-to-left. We used results on newsdev2017 to select one checkpoint from each model. An ensemble of the four left-to-right models was used to generate a 50-best list, which was rescored using the right-to-left models.

For a more detailed description of the system, see Sennrich et al. (2017a).

3.6 UvA: syntactically aware NMT with GCNs

We focus on exploiting structural information on the source side, i.e. in the encoder. We hypothesize that an encoder that incorporates syntax will lead to more informative representations of words, and that these representations, when used as context vectors by the decoder, will lead to an improvement in translation quality. Our model (Bastings et al., 2017) is an attentive encoder-decoder (Bahdanau et al., 2015) where in the encoder side we exploit the power of GCNs (Kipf and Welling, 2016) to induce syntactically-aware representations (Marcheggiani and Titov, 2017). GCNs operate by convolving nodes in a neighbourhood defined by a graph. In our case, a node corresponds to a position in the source sentence which is initially represented by a BiRNN hidden state. We then define a syntactic neighbourhood by following edges in an automatically produced dependency parse. Instead of relying on linear order only (as the BiRNN does), the GCN allows the encoder to 'teleport' over parts of the source
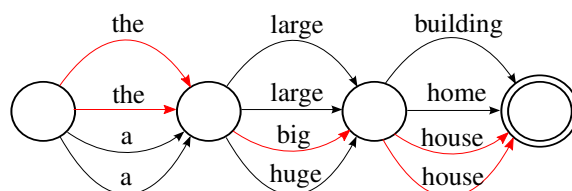

Figure 1: System A: *the large building*; System B: *the large home*; System C: *a big house*; System D: *a huge house*; Reference: *the big house*.

sentence connecting words that are potentially far apart. The model might not only benefit from this teleporting capability however; also the nature of the relations between words (i.e. dependency relation types and directionality) may be useful, and the GCN exploits this information.

4 System Combination

We conducted experiments with two methods for system combination that only require the translated hypotheses. This allows us choose the contributing systems without any restrictions.

4.1 Confusion Network

System combination produces consensus translations from multiple hypotheses which are obtained from different translation approaches, i.e., the systems described in the previous section. A system combination implementation developed at RWTH Aachen University (Freitag et al., 2014a) is used to combine the outputs of the different engines. The consensus translations outperform the individual hypotheses in terms of translation quality.

The first step in system combination is the generation of confusion networks (CN) from I input translation hypotheses. We need pairwise alignments between the input hypotheses, which are obtained from METEOR (Banerjee and Lavie, 2005). The hypotheses are then reordered to match a selected skeleton hypothesis in terms of word ordering. We generate I different CNs, each having one of the input systems as the skeleton hypothesis, and the final lattice is the union of all I generated CNs. In Figure 1 an example of a confusion network with $I = 4$ input translations is depicted. Decoding of a confusion network finds the best path in the network. Each arc is assigned a score of a linear model combination of M different models, which includes word penalty, 3-gram language model trained on the input hypotheses, a binary primary system feature that marks the primary hy-

Systems	newsdev2017/1				newsdev2017/2				newstest2017			
	BLEU	TER	CTER	BEER	BLEU	TER	CTER	BEER	BLEU	TER	CTER	BEER
KIT 20best Bleu	25.8	54.5	51.3	59.3	26.0	55.4	51.0	58.8	17.8	66.9	61.6	53.8
KIT 20best ChrF	25.4	55.0	50.5	59.6	25.7	56.0	50.4	59.2	17.6	68.0	60.9	53.9
KIT 20best Beer	26.0	54.1	50.2	60.0	26.3	54.1	50.6	59.6	18.0	66.7	60.8	54.2
LIMSI Factored 12best Bleu	19.7	60.4	55.7	55.4	19.9	61.0	54.8	55.3	14.2	71.7	63.9	50.8
LIMSI factored 12best ChrF	19.7	60.3	55.4	55.6	19.8	61.1	54.5	55.4	14.2	71.7	63.7	50.9
LIMSI factored 12best Beer	19.8	60.3	55.5	55.6	19.8	61.0	54.7	55.4	14.2	71.7	63.8	50.9
LIMSI factored 100best Bleu	21.5	59.1	55.5	55.9	21.3	59.8	54.8	55.7	15.3	70.7	63.6	51.2
LIMSI factored 100best ChrF	21.9	58.6	54.3	57.1	21.7	59.4	53.6	56.9	15.4	70.5	62.9	52.0
LIMSI factored 100best Beer	21.7	58.7	54.3	57.1	21.6	59.4	53.7	56.9	15.5	70.4	62.9	52.1
Consensus-based System-selection Bleu	19.8	72.5	60.1	51.8	20.5	72.9	59.7	51.6	17.4	69.7	61.9	53.3
Consensus-based System-selection ChrF	26.5	54.1	49.3	60.4	26.8	54.6	48.9	60.2	18.3	67.1	59.9	54.5
Consensus-based System-selection Beer	27.1	53.0	49.6	60.5	27.3	53.8	49.1	60.3	18.6	66.2	60.0	54.6
System Combination	27.4	53.1	50.9	60.2	27.9	53.9	51.0	59.9	18.8	66.0	67.8	54.3
System Combination + Cons-based Beer	27.4	52.7	50.0	60.5	27.7	53.6	51.7	60.2	18.7	66.1	62.0	54.4

Table 2: USFD rescoring and combination experiments English→Latvian task. BLEU [%] and TER [%] scores are case-sensitive.

pothesis, and a binary voting feature for each system. The binary voting feature for a system is 1 if and only if the decoded word is from that system, and 0 otherwise. The different model weights for system combination are trained with MERT (Och, 2003) and optimized towards $8 \cdot \text{BLEU} - \text{TER}$.

4.2 Consensus-based System Selection

As a secondary solution for system combination, we used USFD's consensus-based n-nbest list selection approach (Blain et al., 2017) for system combination by combining each system's output in the form of a n-best list. Inspired by DeNero et al. (2009)'s work on consensus-based Minimum Bayes Risk (MBR) decoding which compares different types of similarity metrics (BLEU, WER, etc.) under a SMT setup, USFD designed a re-ranking approach to empirically evaluate the effect of consensus on the varying n-best list in NMT.

Given a n-best list, each translation hypothesis is scored against the other MT candidates of the search space towards an automatic metric. In our experiment we considered three automatic metrics amongst the most widely used and which have been shown to be well correlated with human judgments (Bojar et al., 2016): BLEU, BEER (Stanojevic and Simaan, 2014) or CHRF (Popovic, 2015). The entire list of MT candidates is then entirely re-ranked according to the averaged score of each candidate. Different from most re-ranking approaches which make use of additional information usually treated as new model components and combined with the existing ones, we here focus only on the MT candidates. The difference between the consensus-based n-best list selection and an oracle translation is the absence

of reference translation: each translation hypothesis is scored against all the other hypotheses used as references while in an oracle translation each translation hypothesis is scored against a single reference. This results in obtaining as best translation hypothesis the candidate that is most similar to the most likely translations.

5 Experimental Evaluation

Since only one development set was provided we split the given development set into two parts: newsdev2017/1 and newsdev2017/2. The first part was used as development set while the second part was our internal test set. The single systems and the system combintaion are optimized for the newsdev2017/1 set.

The single system scores in Table 2 show that the KIT system is the strongest single system closely followed by the UEDIN NMT system. The rescoreing of the UEDIN NMT nbest lists by KIT showed only a small improvement on newstest2017. The system combination of all these systems showed an improvement of 1.1 BLEU on newsdev2017/2 and 0.5 BLEU on official test set, newstest2017.

Table 3 shows a comparison between all systems by scoring the translation output against each other in TER and BLEU. We see that the outputs of the two best performing systems KIT and UEDIN are very close.

6 Morphology Evaluation

In order to get some insight regarding the quality of the morphological correctness of the outputs produced by the systems involved in the combina-

	CUNI	KIT	LIMSI	Tilde	UEDIN	UEDIN r.	UvA	USFD	Average
CUNI	-	38.1	32.4	23.9	37.8	38.2	22.3	40.2	33.3
KIT	43.8	-	49.3	29.9	60.1	62.6	28.3	77.0	49.3
LIMSI	49.8	33.2	-	26.1	48.6	49.4	27.5	56.7	41.4
Tilde	57.0	47.4	52.7	-	30.1	30.2	19.4	31.6	27.3
UEDIN	45.1	25.6	35.1	48.9	-	91.1	28.5	76.2	53.2
UEDIN rescored by KIT	44.5	23.8	34.3	48.4	5.4	-	28.7	78.4	54.1
UvA	62.9	56.6	57.5	65.8	57.1	56.7	-	30.3	26.4
USFD	42.0	13.9	28.1	45.9	15.8	14.2	54.7	-	55.8
Average	49.3	34.9	41.5	52.3	33.3	32.5	58.8	30.7	

Table 3: Comparison of system outputs against each other, generated by computing BLEU and TER on the system translations for newstest2017. One system in a pair is used as the reference, the other as candidate translation; we report the average over both directions. The USFD system is similar to the "Consensus-based System-selection Beer" in Table 2. The upper-right half lists BLEU [%] scores, the lower-left half TER [%] scores.

	verbs		pronouns		nouns	mean
System	past	future	fem.	plur.	number	
Tilde smt	68.8%	70.4%	56.0%	71.8%	65.0%	66.4%
UvA	75.2%	84.2%	46.4%	80.8%	66.8%	70.7%
UEDIN NMT	74.6%	83.6%	57.0%	88.6%	69.4%	74.6%
LIMSI NMT	68.8%	84.6%	64.2%	86.8%	73.0%	75.5%
LIMSI factored	69.6%	82.8%	62.0%	89.0%	70.6%	74.8%
KIT	74.2%	89.0%	56.6%	89.8%	71.6%	76.2%

Table 4: Sentence pair evaluation (A-set).

	coordinated verbs			coord.n	pronouns to nouns			prep.	mean
System	number	person	tense	case	gender	number	case	case	
Tilde smt	49.6%	32.8%	50.2%	47.6%	24.0%	25.4%	19.0%	48.5%	37.1%
UvA	61.8%	52.4%	63.2%	31.6%	36.8%	38.8%	36.6%	50.9%	46.5%
UEDIN NMT	70.6%	60.8%	72.0%	30.2%	46.4%	44.8%	43.4%	56.7%	53.1%
LIMSI NMT	69.2%	57.6%	70.4%	41.8%	40.0%	40.8%	35.8%	54.6%	51.3%
LIMSI factored	72.4%	63.4%	73.2%	34.8%	43.0%	42.2%	41.4%	55.5%	53.2%
KIT	73.4%	64.8%	74.0%	37.4%	51.4%	49.8%	48.8%	55.0%	56.8%

Table 5: Sentence pair evaluation (B-set).

	nouns	adjectives			verbs			mean
System	case	gender	number	case	number	person	tense	
Tilde smt	.436	.755	.735	.768	.254	.337	.258	.506
UvA	.295	.629	.613	.643	.157	.187	.160	.383
UEDIN	.234	.598	.596	.628	.115	.190	.114	.354
LIMSI NMT	.255	.616	.610	.644	.139	.221	.134	.374
LIMSI factored	.233	.587	.582	.612	.117	.182	.113	.346
KIT	.244	.599	.594	.633	.102	.186	.108	.352

Table 6: Sentence group evaluation with Entropy (C-set).

tion, we ran the evaluation method introduced in (Burlot and Yvon, 2017b). The evaluation of the morphological competence of a machine translation system is performed on an automatically produced test suite. For each source test sentence from a monolingual corpus (the *base*), one (or several) *variant(s)* are generated, containing exactly one difference with the base, focusing on a specific *target* lexeme of the base. These variants differ on a feature that is expressed morphologically in the target, such as the person, number or tense of a verb; or the number or case of a noun or an adjective. This artificial test set is then translated with a machine translation system. The machine translation system is deemed correct if the translations of the base and variant differ in the same way as their respective source. Another setup focuses on a word in the *base* sentence and produces *variants* containing antonyms and synonyms of this word. The expected translation is then synonyms and antonyms bearing the same morphological features as the initial word.

There are three types of contrasts implying different sorts of evaluation:

- A: We check whether the morphological feature inserted in the source sentence has been translated (eg. plural number of a noun). Accuracy for all morphological features is averaged over all sentences.

- B: We focus on various agreement phenomena by checking whether a given morphological feature is present in both words that need to agree (eg. case of two nouns). Accuracy is computed here as well.

- C: We test the consistency of morphological choices over lexical variation (eg. synonyms and antonyms all having the same tense) and measure the success based on the average normalized entropy of morphological features in the set of target sentences.

The results for the A-set are shown in Table 4 and reflect the adequacy of an output towards the source, or the quantity of morphological information that has been well conveyed from the source. Certain morphological features indicate rather low contrasts between statistical and neural systems (verb tense and pronoun gender), which shows the relevance of SMT systems in the combination.

Sets B and C are more forcused on target monolingual phenomena, such as agreement, and assess the level of fluency of a system output. Here, the observed contrasts between statistical and neural systems are far more obvious: all B-set SMT scores are below 50%, whereas NMT scores are always above. Here again, the superior performance of KIT is noticed, at least for sets A and B. As for the C-set, LIMSI factored, KIT and UEDIN show a comparable high confidence in their morphology predictions across lexical variety.

6.1 Consensus-based re-ranking

We report in Table 2 the results of the consensus-based approach for either system re-ranking or system combination.

First, we applied our approach on both KIT and LIMSI-factored outputs. While we never outperform original systems' performances, we observe that increasing the n-best size does help with a significant difference between LIMSI's system 12- or 100-best. One would note that in both cases, consensus-based n-best list re-ranking with BEER seems to be performing the best amongst all metrics.

Then, we applied our approach at system-level by combining the outputs of all systems described in Section 3. Once again, we observe better performance with BEER compared to the other two metrics, reaching similar results as the system combination based on confusion network. The only noticeable exception being the CTER score on newstest2017 which is significantly lower compared to the other system combination, most likely the benefit of using character-based metrics.

Finally, we combined both consensus-based selection confusion-based combination and although we observe similar performance to each system individually but a worse CTER.

7 Conclusion

Our combined effort shows again that the combination of different SMT systems results in a better overall system. The final result improved by 0.5 BLEU points. Consensus-based re-ranking showed a performance close to the confusion network approach.

Acknowledgments

354

References

Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *CoRR* abs/1607.06450.

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Arbor, MI, USA, pages 65–72.

Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675* Under review at EMNLP17.

Frédéric Blain, Pranava Swaroop Madhyastha, and Lucia Specia. 2017. Exploring hypotheses spaces in neural machine translation. In Asia-Pacific Association for Machine Translation (AAMT), editor, *Machine Translation Summit XVI*. Nagoya, Japan.

Ondrej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojevic. 2016. Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*.

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovický, and Tomáš Musil. 2017. Results of the WMT17 Neural MT Training Task. In *Proceedings of the Second Conference on Machine Translation (WMT17)*. Copenhagen, Denmark.

Franck Burlot, Mercedes García-Martínez, Loïc Barrault, Fethi Bougares, and François Yvon. 2017a. Word Representations in Factored Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*. Association for Computational Linguistics, Copenhagen, Denmark.

Franck Burlot, Pooyan Safari, Matthieu Labeau, Alexandre Allauzen, and François Yvon. 2017b. LIMSI@WMT'17. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*. Association for Computational Linguistics, Copenhagen, Denmark.

Franck Burlot and François Yvon. 2017a. Learning morphological normalization for translation from and into morphologically rich language. *The Prague Bulletin of Mathematical Linguistics (Proc. EAMT)* (108):49–60.

Franck Burlot and Franois Yvon. 2017b. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation (WMT'17)*. Association for Computational Linguistics, Copenhagen, Denmark.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *arXiv preprint arXiv:1706.00457* http://arxiv.org/abs/1706.00457.

John DeNero, David Chiang, and Kevin Knight. 2009. Fast consensus decoding over translation forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, pages 567–575.

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`.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014a. Jane: Open Source Machine Translation System Combination. In *Proc. of the Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL)*. Gothenberg, Sweden, pages 29–32.

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Nadir Durrani, Matthias Huck, Philipp Koehn, Thanh-Le Ha, Jan Niehues, Mohammed Mediani, Teresa Herrmann, Alex Waibel, Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*. Heidelberg, Germany, pages 128–135.

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Matthias Huck, Rico Sennrich, Nadir Durrani, Maria Nadejde, Philip Williams, Philipp Koehn, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2014b. EU-BRIDGE MT: Combined Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*. Baltimore, MD, USA, pages 105–113.

Markus Freitag, Joern Wuebker, Stephan Peitz, Hermann Ney, Matthias Huck, Alexandra Birch, Nadir Durrani, Philipp Koehn, Mohammed Mediani, Isabel Slawik, Jan Niehues, Eunah Cho, Alex Waibel, Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2014c. Combined Spoken Language Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*. Lake Tahoe, CA, USA, pages 57–64.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation architectures. In *Proceedings of the International Workshop on Spoken Language Translation*. Seattle, USA, IWSLT'16.

Kenneth Heafield. 2011. KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2009, pages 187–197.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *The International Conference on Learning Representations*. San Diego, California, USA.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *CoRR* abs/1609.02907. http://arxiv.org/abs/1609.02907.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints* .

Diego Marcheggiani and Ivan Titov. 2017. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. *ArXiv e-prints* https://arxiv.org/abs/1703.04826.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. *arXiv preprint arXiv:1605.03148* .

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*. Sapporo, Japan, pages 160–167.

Jan-Thorsten Peter, Tamer Alkhouli, Hermann Ney, Matthias Huck, Fabienne Braune, Alexander Fraser, Ale Tamchyna, Ondrej Bojar, Barry Haddow, Rico Sennrich, Frdric Blain, Lucia Specia, Jan Niehues, Alex Waibel, Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, Elena Knyazeva, Thomas Lavergne, Franois Yvon, Stella Frank, and Mārcis Pinnis. 2016. The qt21/himl combined machine translation system. In *ACL 2016 First Conference on Machine Translation*. Berlin, Germany, page 344355.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, Eunah Cho, Matthias Sperber, and Alexander H Waibel. 2017. Kit system description for wmt 2017. In *WMT*.

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. 2017. Tilde's Machine Translation Systems for WMT 2017. In *Proceedings of the Second Conference on Machine Translation (WMT 2017), Volume 2: Shared Task Papers*.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*. pages 392–395.

Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Copenhagen, Denmark.

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. 2017b. 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, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017c. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Demonstrations at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Berlin, Germany.

Nakatani Shuyo. 2010. Language detection library for java.

Miloš Stanojevic and Khalil Simaan. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. pages 414–419.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2016a. Context gates for neural machine translation. *arXiv preprint arXiv:1608.06043* .

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016b. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811* .

Andrejs Vasiļjevs, Raivis Skadiņš, and Jörg Tiedemann. 2012. Letsmt!: a cloud-based platform for do-it-yourself machine translation. In *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, pages 43–48.

Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. 2016. Recurrent highway networks. *arXiv preprint arXiv:1607.03474* .