BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2019
http://excel.fit.vutbr.cz

Exploring contextual information in neural machine translation

Josef Jon*

Abstract
This works explores means of utilizing extra-sentential context in neural machine translation (NMT). Traditionally, NMT systems translate one source sentence to one target sentence without any notion of surrounding text. This is clearly insufficient and different from how humans translate text. For many high resource language pairs, NMT systems output is nowadays indistinguishable from human translations under certain (strict) conditions. One of the conditions is that evaluators see the sentences separately. When evaluating whole documents, even the best NMT systems still fall short of human translations. This motivates the research of employing document level context in NMT, since there might not be much more space left to improve translations on sentence level, at least for high resource languages and domains. This work summarizes recent state-of-the art approaches, implements them, evaluates them both in terms of general translation quality and on specific context related phenomena and analyzes their shortcomings. Additionally, context phenomena test set for English to Czech translation was created to enable further comparison and analysis.
Keywords: Neural machine translation, context, discourse

Supplementary Material: *N/A*

*mailto:xjonjo00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Quality of state-of-the-art machine translation systems has improved vastly over the last few years, thanks to shifting the paradigm from phrase-based statistical machine translation to models based on complex artificial neural networks.

In 1986, Martin Kay [1] stated reasons why high quality machine translation is not possible - but that was before "The statistical revolution" [2], in times of rule-based systems and symbolic AI. Nowadays, there is almost no doubt that high quality machine translation is feasible - in some test scenarios, recent neural machine translation (NMT) systems are evaluated on par with or even better than human translators. However, challenges mentioned in Kay's statement, and many more, still hold true today, and they are not addressed even in the current state-of-the-art.

This work is focused on one of these challenges - utilizing discourse-level, cross-sentence context in NMT. Current systems usually only use one sentence as their input, which is clearly insufficient, as a single sentence may not contain enough information for a proper translation of itself. Exploiting the discourse addresses many interesting sub-problems, like adaptation to topic, genre, domain, or author's style, discourse consistency (e.g. lexical consistency – using the same translation for one entity throughout the whole document), coherence and cohesion, coreference resolution (e.g. cross-lingual pronoun disambiguation, also mentioned in Kay's paper).

However, utilizing context is more than solving each of the problems mentioned above separately, since discourse can contain information that is not contained in any of the sentences of the text alone. As stated by Kehler [3]: "The meaning of a discourse is greater than the sum of the meanings of its parts."

In this work, I implement some of the recent techniques of utilizing context in NMT and I evaluate them in terms of both general translation quality, and accuracy on translation of specific discourse phenomena. I try to analyze their shortcomings and design a system that mitigates them.

2. Discourse

This work focuses on translation of a text given a document context, or discourse. Eisenstein [4] characterizes discourse simply as ”multi-sentence linguistic phenomena“ in his recent NLP notes. Andrew Kehler [3] refers to discourse as ”collocated, related groups of sentences“. Kendall and Wickham [5] say that a discourse is a corpus of statements whose organization is regular and systematic. Broader definition of discourse is that it is the use of spoken or written language in context of society. For the rest of this work, we will assume that a discourse means multiple sentences that have some kind of connection between each other. An important thing to keep in mind is that the meaning communicated by the discourse is bigger than the sum of meanings of individual sentences. Discourse can contain information that none of the sentences contains by itself.

3. Current state-of-the-art in machine translation

In 2014, two papers papers with major impact on MT were released by Sutskever et al. [6] and Cho et al. [7]. The main differences compared to previously used approach, phrase based machine translation (PBSMT), were the following two. First, the NMT systems use continuous, distributed representation of words [8]. This means that words that appear in similar contexts are represented and processed similarly by the model, and that the representation is more semantic, in the spirit of J.R. Firth’s quote:

> You shall know a word by the company it keeps.

Second big shift from PBSMT is that the system uses only one model, based on encoder-decoder neural network, performing all the necessary operations, instead of combination of engineered models for each task.

Nowadays, one of two types of deep neural networks are used in practice. Almost simultaneously in 2014, Sutskever et al. [6] and Cho et al. [7] published papers concerning neural network based MT systems. The results of these systems led to big increase of popularity of this research topic. Both of the systems used encoder-decoder RNN networks with LSTM (Sutskever) or GRU (Cho) units, and were further improved by an attention mechanism [9]. Since summer of 2017, RNNs are being replaced with self-attention based models [10] which are more parallelizable, since they remove the need for sequential processing of the input sequence inside the network, and also usually offer superior translation quality.

4. Current approaches to using context in NMT

This work deals with employing extra-sentential context in NMT. Many publications about this topic emerged in the last two years. After Microsoft claimed reaching human parity in Chinese-English news translation [11], Läubli et al. [12] tried to analyze these claims and asses if they are true.

The translations were evaluated in terms of fluency and adequacy. The evaluators were shown a source sentence (in case of adequacy evaluation, fluency evaluators were only shown the two translations) and two translations, one produced by a human (professional translator) and one by Microsoft’s MT system. They were asked two questions:

> Which translation expresses the meaning of the source text more adequately? (adequacy)

and

> Which text is better English? (fluency)

The results did in fact confirm Microsoft’s claims - in terms of adequacy, the evaluators preferred MT in 50% of the sentences, did not have any preference in 9% and preferred the human translation in 41% of the cases. However, when the evaluators were asked to compare whole documents, the results changed drastically - only 32% of machine translated documents were preferred based on adequacy ratings. These results convincingly show the need for document level translation.

One of the earliest attempts in incorporating discourse into NMT is a work by Jean et al. [13]. The presented system utilizes a dual encoder RNN, with one encoder for a source sentence, as usual, and another auxiliary encoder for a context sentence. Attention mechanism for the contextual encoder also has source vector from the main attention as an input, besides the usual inputs (previous symbol, previous decoder state, annotation vector). The authors evaluated their model in terms of general translation quality (BLEU), as well as in more focused evaluation - pronoun prediction (RIBES). They observed improvements for both of the metrics while using small training data - ISWLT or WMT16 reduced to up to about 40%. However, when they trained on a larger corpus, the improvements vanished.

In [14], authors evaluate RNN and Transformer architectures with context windows of up to three previous source sentences and a next source sentence on the source side, and previous one or two target sentences on the target side. Context sentences were added either by concatenation (separated by a special token), or as an input for an additional encoder. They trained and evaluated their system on English-Italian IWSLT 2017 dataset, consisting of transcribed TED talks.

They observed drop in BLEU score when adding context to RNN via simple concatenation, probably because even though LSTMs have gating mechanisms and the network used attention, signal is still vanishing in long-range dependencies. When using multi-encoder architecture, BLEU increased for RNN. Other research suggests gains for RNNs even when using concatenation, but usually on OpenSubtitles dataset, where average sentence length is much shorter [15]. For a Transformer, where they ran only experiments using concatenation, the best combination was one previous and one following source sentence on the source side and one previous target sentence on the target side, yielding a 2 BLEU points gain over the baseline.

Paper by Voita et al. [16] utilizes dual encoder transformer, with some of the encoder layers weights shared and gated dual attention. The models was trained on OpenSubtitles corpus, and resulted in a slight improvement in BLEU, pronoun disambiguation and coreference resolution. Other approaches include memory networks Maruf and Haffari [17] or hierarchical RNNs. [18].

5. Experiments

To compare the effects of utilizing context in different domains and with different types of data, two publicly available datasets are used: Europarl [19] and OpenSubtitles2018 [1]. These datasets, which contain document boundaries, were split into train, development and test sets and standard preprocessing for machine translation was applied - tokenization, truecasing (both using Moses [20] scripts) and splitting into BPE segments using subword-nmt [21]. Preprocessed files were converted into formats suitable for the evaluated architectures using custom Python scripts. Marian[22], an efficient C++ NMT framework, was used to perform these experiments.

5.1 Evaluation

To evaluate systems in terms of BLEU scores [23], parts of training corpora were set aside to create development and test sets. The scores are computed using SacreBLEU [24].

For more targeted evaluation of inter-sentential phenomena, an approach used by Bawden et al. [15] was adopted. The authors created a manual contrastive test sets to quantify a machine translation system accuracy in translating coreference and coherence/cohesion phenomena. The set comprises of source sentences and both correct and incorrect translation. NMT model is used to score both translation in terms of cross-entropy, and choose the translation with higher probability. The test set is balanced so that any system without employing context scores 50%. In disambiguation part, which is used in this paper, there is one current source sentence, two possible previous source sentences and two possible translations - each one correct in one of the contexts. For example :

> Context 1:
>
> *We went to the cliffs to watch our favorite seal in the sea.*
>
> Context 2:
>
> *We went to his house, which was sealed by the police because of the crime investigation.*
>
> Source :
>
> *When we have seen the seal, we went back home.*

Now we have two pairs of target sentences for each context:
For Context 1:

> Incorrect: *Když jsme tu pečeť uviděli, šli jsme domů.*
> Correct: *Když jsme toho lachtana uviděli, šli jsme domů.*

and for Context 2:

> Incorrect: *Když jsme toho lachtana uviděli, šli jsme domů.*
> Correct: *Když jsme tu pečeť uviděli, šli jsme domů.*

For both contexts, correct and incorrect translation is scored by the model and the more probable one is chosen. Final accuracy is computed based on how many times the correct translation was preferred over the incorrect one. Since the test set contains both possible correct combinations paired with the wrong ones, a system without any knowledge of the previous context will always score 50% (as it will choose the

[1] http://opus.nlpl.eu/OpenSubtitles2018.php

same target sentence as more probable both times, the cross-entropy score will be the same for both contexts). For English-French, the original dataset was used [2], for English-Czech, relevant parts were translated and joined with newly created examples [3].

5.2 Models

Transformer and RNN network with GRU cells were used as a baseline. For the Transformer, the hyper-parameters were generaly the same as in the original paper [10] (transformer-base). For the RNN model, the architecture is similar to WMT2017 systems by University of Edinburgh [26].

Context size naming conventions Different configurations of input and output are labeled src$N$tgt$M$ to tgt$K$ in this paper, where N and M are counts of previous source and target sentences concatenated to the input, and K is how many previous context sequences are to be generated.

Thus, *src0tgt0 to tgt0* is a normal, vanilla NMT without any context influence, *src1tgt0 to tgt0* means that one previous source sentence is concatenated to the input, *src0tgt1 to tgt0* means that one previous source sentence is concatenated to the input and so on. For systems generating more than one target sentence, i.e. *src1tgt0 to tgt1*, the target side of the training data is preprocessed in the same way, this means that the model is learning to generate more than one sentence. For target context on source side, reference translation is used. I plan to perform realistic experiments with translations of the previous target sentences generated by the system itself in the complete master's thesis.

Concatenation The most straight-forward approach to employ extended context is to simply concatenate additional sentences to the input of the model. Therefore, in the first experiments, single encoder, single decoder model (RNN or Transformer), with context sentences concatenated with the source sentence, separated by a special token, was used.

For initial experiments, maximal sentence length in subwords was set to 80 for Europarl and 55 for OpenSubtitles, multiplied by number of source sentences (e.g. 160 (110) for src1tgt0 to tgt0), based on sentence length analysis of the dataset. For the Europarl baseline, this length turned out to be insufficient. This issue is further discussed later.

Multiple encoders Another way to integrate the context into an NMT model is via an additional encoder, usually with a same structure as the original one. Source sentence is fed into the original encoder, and context sentence into the additional one. Encoding runs independently for both of the encoders. Encoders can have either separate, or shared layers, i.e. weights of the neurons are the same for both encoders. The only difference in regard to the vanilla model is that source-target attention the decoder attends over both encoders. There are a several attention strategies for multiple encoders, for example hierarchical, serial or parallel attention, see [27], we will explore only serial attention, i.e. decoder first attends over one encoder and then, with state already updated by this attention, attends over the second one.

Context Encoder Inspired by [25], I implemented Transformer with context encoder in Marian [22]. This architecture also utilizes two encoders, yet there are a few differences in comparison with multiple encoder architecture described above. First, the encoders are not excactly the same - the context encoder has fewer self-attention layers (only one, while the source encoder has six). Second, the context encoder states are also attended over in the source encoder, and not only in decoder, in contrast to the previous approach. Also, the influence of context encoder is gated by a sigmoid gate. This should allow better usage of the context. Schematic overview is presented in Figure 1.

This architecture allows for a vanilla Transformer model to be pretrained on general data without document boundaries (which are usually much bigger). Then, weights of this model are frozen and the additional components (highlighted in red in the figure) are added . Their weights are then tuned on a smaller corpus with document level information. During inference, the system can either use the full model in case that the input has context information, or only the pretrained part, for single sentence translation.

In the experiments described in this paper, the models were trained on identical corpora in both phases (in the first one without the context level information), so it is expected that there is no observable gain from pretraining, other than quicker convergence of training in the second phase (much fewer parameters have to be learned). In my Master's thesis, I plan to perform realistic experiments with pretraining on big corpus without document level boundaries a then tune the context-aware model on small document-split corpus.

5.3 Results

As you can see in Table 1, adding context to RNN through concatenation hurts the translation performance. This is in line with observation made by Agrawal et al. in [14], where the authors obtained similar result

[2] https://github.com/rbawden/discourse-mt-test-sets
[3] https://github.com/cepin19/discourse-test-set

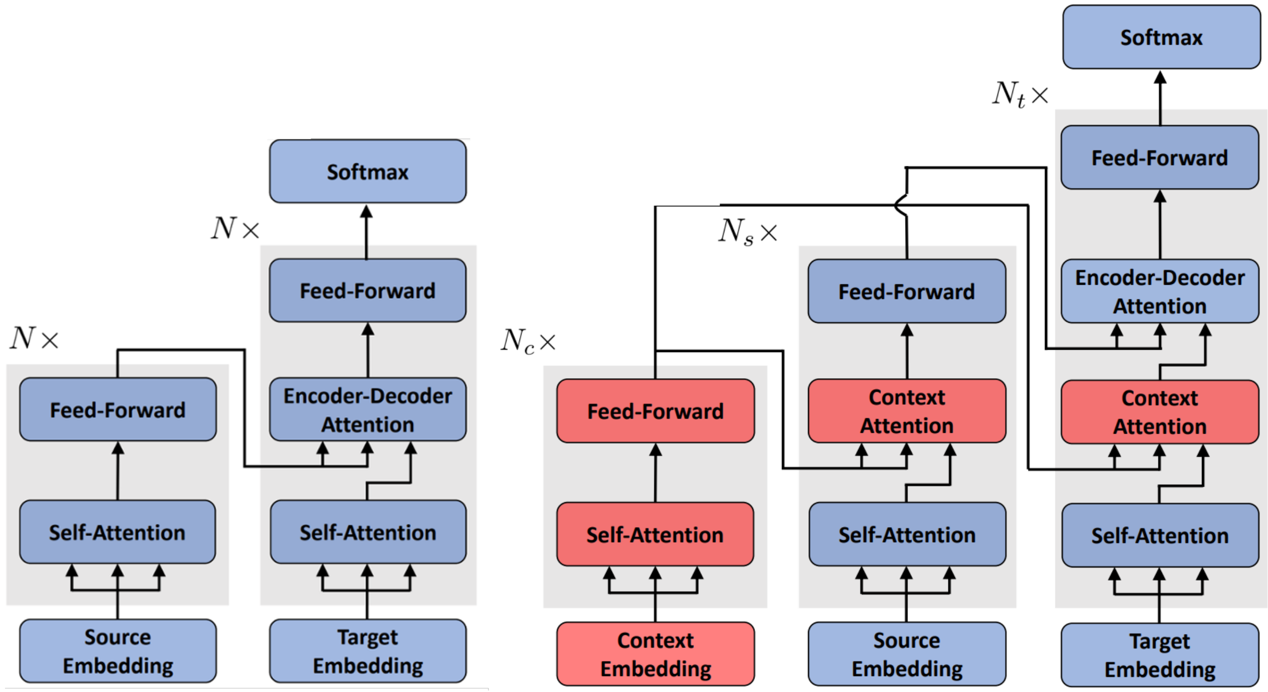

Figure 1. The original Transformer model (left) and the Transformer with context encoder (right), taken from [25]. The parts highlighted in red are the new additions to the model - context encoder, and attention over the context encoder states in both encoder and decoder.

on ISWLT 2017 English-Italian data set, consisting of transcribed TED Talks. For configurations src1tgt0 to tgt0 and *src1tgt0 to tgt1*, they observed BLEU drops of 1.8 and 2.8 respectively. Arguably because even though there are gating mechanisms employed, RNNs suffer from loss of signal in very long-range dependencies.

Context	src0tgt0 to tgt0	src1tgt0 to tgt0	src1tgt1 to tgt0
BLEU	**29.07**	28.88	27.82

Table 1. BLEU scores of concatenation experiments with RNN, English to Czech, Europarl, average of three runs, dev set

Context	src0tgt0 to tgt0	src1tgt0 to tgt0	src2tgt0 to tgt0
BLEU	27.77	27.93	**27.97**

Table 2. Results of concatenation experiments with Transformer, English to Czech, OpenSubtitles dev set

As shown in Tables 2, 3 and 4, the performance of Transformer model does not degrade when concatenating the input sequence, confirming the assumption that RNNs are worse equipped to deal with longer sequences. However, there are not big gains observed in BLEU score either. Only architecture that seems to obtain a significant improvement, at least in some scenarios, is concatenation.

Largest difference in BLEU scores was observed in *src1tgt1 to tgt0* concatenation on English-French dataset, i.e. base Transformer model with one previous source and one previous target sentence concatenated to the input sentence. The target context sentences in this scenario are taken from reference data - meaning the real result, using model-generated previous sentences, will probably perform significantly worse due to error propagation.

This configuration was not investigated further, since the dependence on previous target sentence is preventing parallelization of translation using batching, thus making this approach impractical. However, the evaluated configuration could be useful in post-editing scenario, where a human translator sequentially corrects translations made by an MT system, so the corrections made in previous sentences can be used to improve translation of the future sentences.

Discourse test set Some of the models were also evaluated in terms of accuracy on the discourse test sets described earlier, and the results are presented in *disambig* columns in corresponding tables. The only significant gains, again, were obtained by concatenation architecture, trained on OpenSubtitles. Dual encoder models, despite the slight gains in BLEU, do not seem to utilize context information too much - accuracy for different model checkpoints fluctuated between 48.5 – 51.5 % for a simple dual encoder model.

This suggests that the BLEU gains for dual encoder models are caused by something else than a successful context utilization. For dual encoder and

Context type	architecture	n	real context	random context	null context	$\Delta_{real,random}$	disambig
0,0 to 0	baseline	0.6	34.92	-	-	-	50%
0,0 to 0	baseline	opt	36.38	-	-	-	50%
1,0 to 0	concat	0.6	35.36	34.85	35.35	0.51	**62.3%**
1,0 to 0	concat	opt	**36.60**	35.97	36.35	0.63	**62.3%**
1,0 to 0	dual encoder	0.6	35.36	35.14	1.41	0.22	50.6%
1,0 to 0	dual encoder	opt	36.47	36.32	1.44	0.15	50.6%
1,0 to 0	dual encoder, shared	0.6	34.98	34.32	1.38	0.65	50.6%
1,0 to 0	dual encoder, shared	opt	36.33	35.64	1.39	0.69	50.6%
1,0 to 0	dual encoder, shared + tok	0.6	35.24	34.93	1.41	0.31	51.8%
1,0 to 0	dual encoder, shared + tok	opt	36.42	36.15	1.43	0.27	51.8%
1,0 to 0	context encoder	0.6	35.02	-	-	-	51.2%
1,1 to 0	concat	0.6	37.12	20.28	33.93	16.84	60%
1,1 to 0	concat	opt	**37.44**	20.92	35.14	16.52	60%

Table 3. Results for models trained on **English-French OpenSubtitles**, Transformer model, dev set. First column shows type of context - first number stands for previous source sentences added to input, second one for previous target sentences added to input and the third one is number of additional previous target sentences generated by the model, so for example 1,0 to 0 equals to model denoted *src1tgt0 to tgt0* elsewhere in the text. Second column is the model architecture, for dual encoder, *shared* means that all weights in the two encoders are shared, in *shared+tok* strategy, a special token is added to the start of the previous sentence - since all the layers are shared, the encoder would otherwise be unable to distinguish between context and source sentence and that may not be optimal. Third column is the length normalization coefficient - *opt* means an optimal value found by search over possibilities within a given range, see paragraph Length normalization. In columns number 4, 5, and 6, BLEU scores depending on whether real, random, or empty context sentences were used, are shown. Next column shows difference between real and random context BLEU scores. Finally, in the last column, accuracy on disambiguation part of contrastive discourse test set is presented.

Context type	architecture	len	n	real context	random context	null context	disambig
0,0 to 0	baseline	80	0.6	29.6	-	-	50%
0,0 to 0	baseline	160	0.6	30.3	-	-	50%
1,0 to 0	concat	160	0.6	30.3	30.3	29.4	51.8%
1,0 to 0	dual encoder	80	0.6	30.0	30.0	30.0	50.5%
1,0 to 0	CE	80	0.6	29.8	-	-	-
1,0 to 0	CE, +gate	80	0.6	29.9	-	-	-
1,0 to 0	CE, +gate, pretrain	80	0.6	30.0		-	-
1,0 to 0	CE, +gate, pretrain	160	0.6	**30.5**	30.4	0.1	**52.8**%
1,0 to 1, 1st sent	concat	160	0.6	30.0	-	-	-
1,0 to 1, 2nd sent	concat	160	0.6	29.8	-	-	-
1,0 to 1, 1st sent	concat	160	1.9	30.2	-	-	-
1,0 to 1, 2nd sent	concat	160	1.9	30.1	-	-	-
1,1 to 0	concat	240	0.6	29.97	-	-	-

Table 4. Results for models trained on **English-Czech Europarl**, Transformer model, dev set. For detailed description of the columns, see previous table. The additional *len* column shows maximal sentence length in subwords for training, see paragraph Maximum source sentence length for further discussion of this issue. *CE* denotes the context encoder model, *+gate* is the same model improved with sigmoid gate to filter the influence of context, *pretrain* means that the model was pretrained on the same corpus without context information. For 1,0 to 1 context type, the model is trained to generated not only the current target sentence, but also the previous one, separated by a special token. 2nd sentence score is obtained by striping off the first (previous) target sentence and calculating BLEU on dev set, whereas 1st sentence score is obtained by cutting off the second (current) target sentence and computing BLEU of the first target sentence on dev set that is shifted accordingly by one sentence.

context encoder configurations, the BLEU gains are observed probably mainly due to increased number of parameters in comparison to the baseline model - there are more attention layers and subsequent feedforward layers, which in theory should serve to incorporate the context information, but their main contribution in reality is presumably improving the representation of current source sentence.

Adversary context Several models were also evaluated with a random context as an input, instead of a real one. Quite surprisingly, the results were not much worse with the random context sentences, especially for Europarl corpus, as you can see in Tables 4 and 3. This, along with results on the discourse datasets, also shows that the models do not depend on context information too much. As mentioned in last paragraph, the BLEU gains over baseline for multi encoder models can be explained by increased number of parameters of the model.

For concatenation configuration, this is not true, model architectures are exactly the same regardless whether the context is used or not. However, on English-Czech Europarl corpus (see Table 4) an improvement over the baseline (29.6 BLEU) can be observed for concatenation system, even when random context is used (30.3 BLEU). Maximum source sentence length for training is set differently for baseline and concatenation models, which seems to be the issue.

Maximum source sentence length For Europarl, maximum length of the source sentence was set to 80 subwords for no context, multiplied by the number of context sentences for concatenation models. Since it is not probable that two exceedingly long sentences will follow each other, concatenation models had chance to train on these sentences, while the baseline model excluded them. I assumed it will not hurt the performance too much, based on a sentence length analysis, only 1.2 % of the source sentences were longer 80 subwords in English-Czech Europarl. As it turned out, this assumption was wrong.

When trained with maximum input length of 160, baseline model performs the same, or better, as other models, reaching BLEU score of 30.3 on English-French Europarl dev set. This does hold true for Europarl, on OpenSubtitles, I did not observe this problem and some limited gains can be seen. This suggests that there is more to gain using context on OpenSubtitles dataset than on Europarl, or different techniques need to be used for different datasets.

Length normalization Usually, in NMT, beam search is used to select the best sentence translation from hypotheses generated by the model. Beam search in NMT has two hyperparameters - beam size and length normalization constant n. Without length normalization, probabilities of each word along the beam are summed up and then the best overall score (log-likelihood) is chosen. Usually, this results in preference for shorter sentences, since less total tokens in the output will probably mean lower (better) score. To mitigate this issue, which often harms translation quality, we divide the final summed score by number of output tokens l to the power of n: l^n. However, n has to be chosen empirically since its optimal value varies from language to language, dataset to dataset, and model to model. Popular choice is 0.6, which is the default used in experiments in this paper, if not stated otherwise.

However, for some of the models, optimal n was determined by search in interval 0.4-3.0 (with step 0.1). Results for concatenation model trained on English-French OpenSubtitles are shown in Table 5. Optimal n was always much higher than 0.6, usually in range 1.5-2.5. Also, it is different for each model, so probably the most fair way to compare the models is to choose optimal n on the development set for each model and then compare the test set scores with these parameters. Beam size and length normalization value are not independent of each other - an ideal solution would be to run a grid search along these two parameters, which was not done due to computing restraints - beam size was set at 6 for all the experiments.

n	Real	Random	Empty
0.6	35.67	35.10	35.55
opt	36.60 (1.7)	35.97 (1.6)	36.35 (1.8)

Table 5. Effect of length normalization parameter on BLEU score, OpenSubtitles English-French, concatenation, src1tgt0 to tgt0, real, random and empty context. 0.6 was a default value in most of the experiments, values in parentheses are the optimal values for given input found by search in 0.4-3.0 with step of 0.1

6. Conclusions and future work

This paper summarizes current state-of-the-art of dealing with extra-sentential context in NMT. Some of the simpler architectures were evaluated and compared both in terms of general translation quality and evaluation focused on discourse phenomena. A compute-efficient architecture was implemented in a framework suitable for production systems. A hand made discourse test set for English to Czech translation was created. The experiments have shown that either current context-aware models are not very efficient at employing context, or there is a very little to gain in automated translation quality metrics by using context models, even though this varies by the dataset used. In the focused evaluation, some limited evidence of correct context usage was observed.

The results of experiments using random context sentences suggest that the context-aware models do not depend on context information too much - a recent paper by Jean and Cho [28] confirms this observation and the authors propose a model-independent modification of the cross-entropy loss function, which is aimed to make the model more sensitive to the context. Since this algorithm can be used with any neural network MT architecture, it is an interesting future research direction.

The best performing models in both metrics were the simplest ones - with context sentences concatenated to the input, separated by a special token, without any changes to the model architecture. These results seem to be in line with recent development in other fields of natural language processing, where big Transformer models, trained on huge datasets and many GPUs, usually outperform specialized models with an explicit problem knowledge programmed into them.

Acknowledgements

I would like to thank my supervisor Pavel Smrž and Martin Fajčík for their help, and also to Michal Hradiš for his detailed and encouraging feedback.

References

[1] Martin Kay. Machine translation will not work. In *24th Annual Meeting of the Association for Computational Linguistics*, 1986.

[2] Mark Johnson. How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 3–11. Association for Computational Linguistics, 2009.

[3] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[4] Jacob Eisenstein. *Natural Language Processing*. MIT Press, 2018.

[5] G. Kendall and G. Wickham. *Using Foucault's Methods*. Introducing Qualitative Methods series. SAGE Publications, 1999.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[7] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[11] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567, 2018.

[12] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? A case for document-level evaluation. *CoRR*, abs/1808.07048, 2018.

[13] Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*, 2017.

[14] Ruchit Agrawal, Marco Turchi, and Matteo Negri. Contextual handling in neural machine translation: Look behind, ahead and on both sides. 2018.

[15] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. *CoRR*, abs/1711.00513, 2017.

[16] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. *CoRR*, abs/1805.10163, 2018.

[17] Sameen Maruf and Gholamreza Haffari. Document context neural machine translation with memory networks. *CoRR*, abs/1711.03688, 2017.

[18] Longyue Wang, Zhaopeng Tu, Andy Way, and Liu Qun. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[19] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

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

[21] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015.

[22] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[24] Matt Post. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771, 2018.

[25] Jiacheng Zhang, Huanbo Luan, Maosong Sun, FeiFei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the transformer translation model with document-level context. *arXiv preprint arXiv:1810.03581*, 2018.

[26] Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. The university of edinburgh's neural mt systems for wmt17. *arXiv preprint arXiv:1708.00726*, 2017.

[27] Jindrich Libovický and Jindrich Helcl. Attention strategies for multi-source sequence-to-sequence learning. *CoRR*, abs/1704.06567, 2017.

[28] Sébastien Jean and Kyunghyun Cho. Context-aware learning for neural machine translation. *CoRR*, abs/1903.04715, 2019.