Exploring contextual information in neural machine translation

Josef Jon, submission 54
FIT VUT
xjonjo00@stud.fit.vutbr.cz

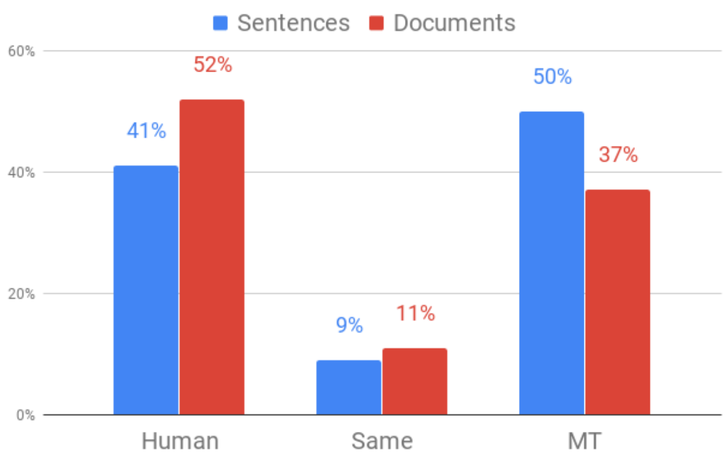
Excel@FIT 2019

1. Introduction

In the past few years, a huge progress has been made in human language machine translation (MT). Nowadays, for language pairs and domains with large amount of high quality parallel training data (millions of sentences), neural network based systems can, in some cases, produce output nearly indistinguishable from a human translation. However, many issues remain unresolved. One of these problems is that most of the MT systems assume that sentences in the source text are independent – the sentences are processed one by one, without sharing any information between the translations. This assumption is false – often, important context for translation of a sentence is located outside of the sentence itself. In the last two years, several techniques of utilizing context in neural machine translation (NMT) were presented. This paper tries to analyze the recent work and compare it in terms of how well the context is used and how it affects the translation quality.

2. Why?

The one to one sentence paradigm is used only as simplification of the engineering side of machine translation. It is clear that human translators do not forget everything they know about the text before translating a new sentence. To somehow measure effects of context knowledge on translation adequacy empirically, Laubli et al.[1] compared Microsoft's MT system and human translators in two scenarios. In the first one, the evaluators were shown a source sentence, MT system translation and human translation, and they were instructed to choose which one they prefer. In the second scenario, they were shown full documents. Results show that in the sentence level evaluation, the MT system performed on par with human translation. However, when the evaluators saw the whole document, human translations gained a lot of ground and were evaluated as superior.



Laubli et al.[1]: Translation adequacy rating for Chinese-English translation, shows percentage of evaluators preferring a translation made by NMT, human, or neither of them.

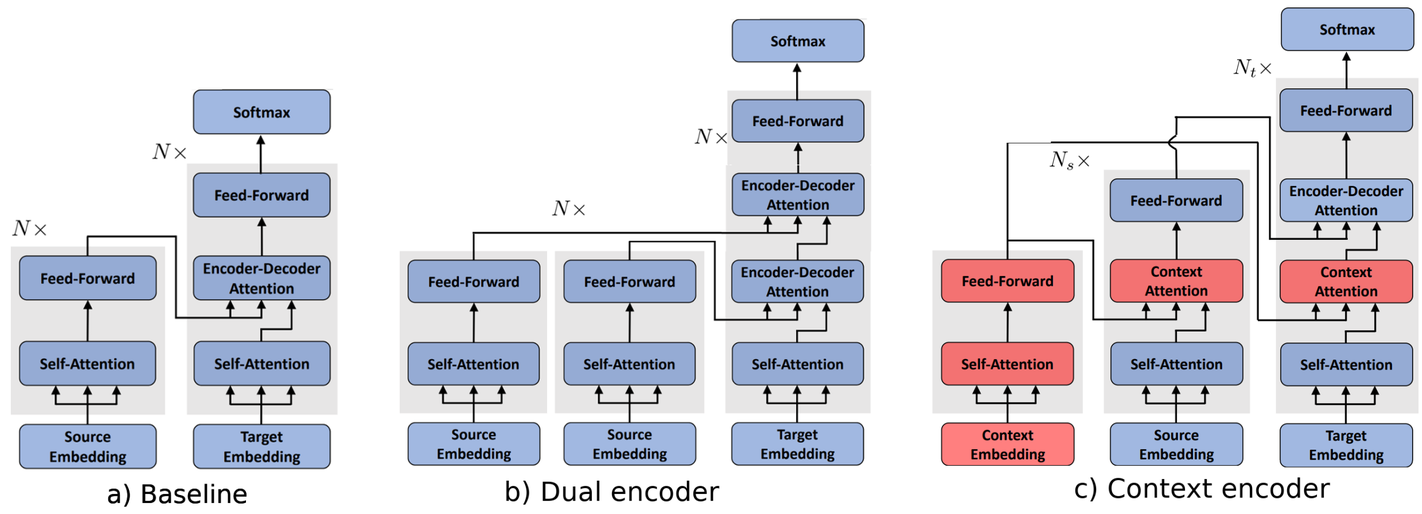
3. Test set

Context 1:	We went to the cliffs to watch our favorite seal in the sea.
Context 2:	His house was sealed by the police because of the crime investigation.
Source:	When we have seen the **seal**, we went back home.
Translation 1:	Když jsme toho **lachtana** uviděli, šli jsme domů.
Translation 2:	Když jsme tu **pečeť** uviděli, šli jsme domů.

4. Models

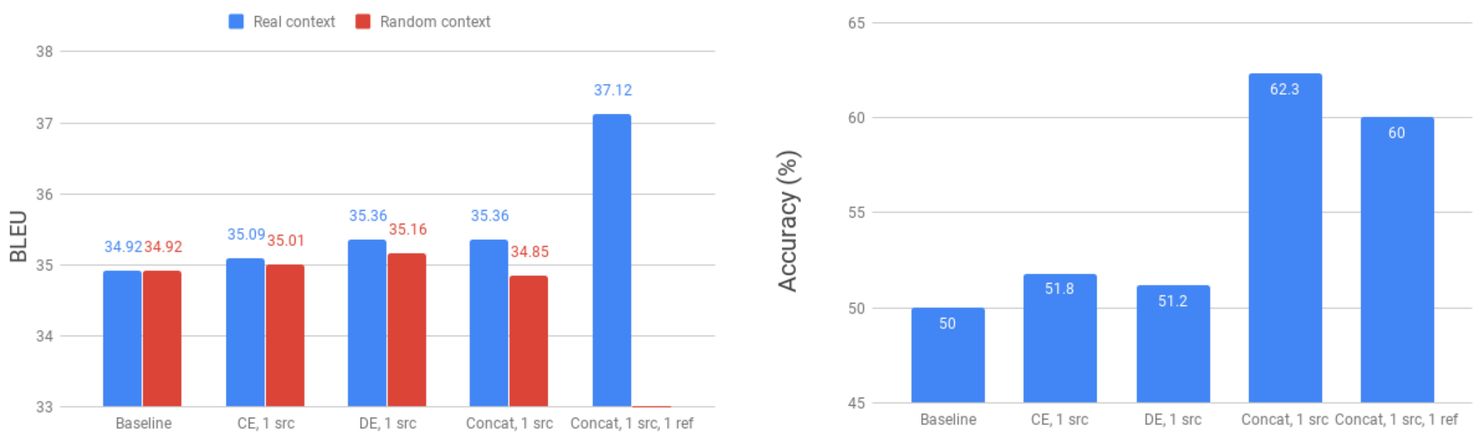
Nowadays, two network architectures are generally used in NMT: Recurrent neural networks (RNN) with LSTM or GRU units, and more recently, the **Transformer**[2]. Both follow the encoder-decoder general structure: encoder creates a vector representation of the sentence in source language, while the decoder utilizes this representation to generate a translation in the target language. There are several possible approaches of employing additional context in an NMT model:

- **Concatenation** of input sentences
- Using an **additional network** to create a representation of the context
- **Dual encoder** – two identical encoder running parallelly for source and context sentence, states of both encoders are used in the decoder
- **Context encoder**[3] – combination of 2) and 3), two encoders running sequentially, first the context is encoded, context representation is used in computations of ordinary encoder



5. Evaluation

- Most common MT quality metric - BLEU
  - Based on overlap of token n-grams in an MT system generated translation and human-made reference translation (says how similar is the MT output to a human reference translation)
  - Many issues with BLEU, but usually correlates well with human evaluation
  - Compare translation quality with real and random context - maybe the observed gains in BLEU are only because the network us bugger, and not due to context utilization
- More focused analysis – specialized test set to measure accuracy on discourse phenomena (translation **disambiguation based on a previous source sentence**)
  - For **English to French**, test set made by Bawden et al.[4] was used, for **English to Czech**, part of this test set was translated and more examples were created
  - Two different previous source sentences, current source sentence and two possible translations, each correct for one of the previous sentences



Opensubtitles, English to French, BLEU scores on dev set and accuracy in discourse phenomena translation

6. Conclusions

1. Recent evaluation campaigns suggest that there may not be much more room for improvement in single sentence translation in high-resource language pairs and domains -> **document level translation**
2. The simplest approach – concatenation of input sentences – seems to work the best
3. For more complicated models, some gains in BLEU score can be seen, however, these gains are probably not due to correct context utilization
4. It's not easy to measure how well the model uses context by common translation quality metrics -> specialized test set and metrics
5. Future research: context-aware learning[5], hierarchical attention networks, different attention mechanisms for multiple encoder architectures

7. References

[1] Samuel Läubli et al. Has machine translation achieved human parity? A case for document-level evaluation. *CoRR*, abs/1808.07048, 2018.

[2] Ashish Vaswani et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[3] Jiacheng Zhang et al. Improving the transformer translation model with document-level context. *arXiv preprint arXiv:1810.03581*, 2018.

[4] Rachel Bawden et al. Evaluating discourse phenomena in neural machine translation. *CoRR*, abs/1711.00513, 2017.

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