Empirical robustness study of state of the art seq2seq neural network for machine translation

Raffaele Della Pietra

Optimization Class Project. MIPT

`https://github.com/rafdp/typo_ml`

Introduction

During the past years the machine translation architecture has become a popular tool for more than just translation from one language to another. It gained many uses, such as paraphrasing within a single language (autoencoders), questionnaire-like interaction, chat-bots and others.
With such a vast field of uses, an obvious problem arises: are the current state of the art solutions for machine translation robust to various misspellings?
In this work I tried to assess the dependence of quality of translation on the type and amount of spelling mistakes that may arise while interacting with a human being.

Corpus

The Latvian - Russian corpus from OPUS [1] was used. It contains about 380k paired utterances, which were divided in a train corpus (345k) and an evaluation corpus (38k). The same evaluation corpus was replicated for every type of misspelling emission configuration. To prepare the corpus for machine translation, normalization was applied, removing rare characters, punctuation and casing.

Misspelling types

The different types of misspellings involved were as follows: character deletion, character doubling, word glueing, word splitting, random character insertion, random wrong character and consecutive character swap. No language-specific misspellings such as keyboard-related typos were considered, as the research was supposed language independent.

Evaluation setup

To achieve statistical stability in all the evaluation stages, a quite large subcorpus (38k) was replicated for each configuration of misspellings. The configurations are as follows:

Experiment	Type of typos	Amount
cln	No typos	
del	deletion	40%
dbl	doubling	40%
gl	glueing	40%
spt	splitting	40%
ins	insertion	40%
rnd	random wrong	40%
swp	consecutive swapping	40%
shf1	all combined	1% each
shf2	all combined	2% each
shf5	all combined	5% each
...	...	...
shf50	all combined	50% each

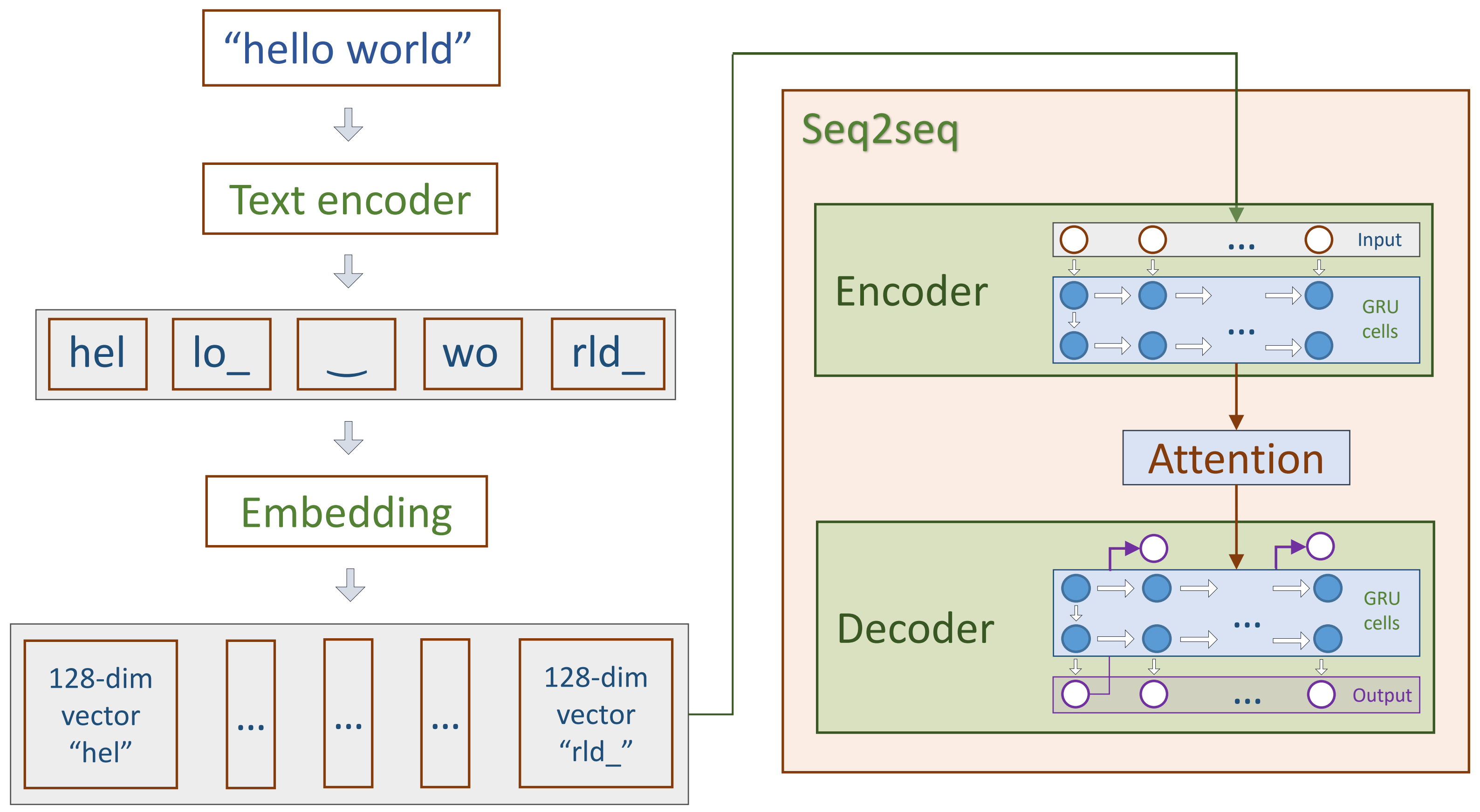
Text encoding

For source encoding, two methods were used: frequent subword, called wordpiece, and character encoding. For target encoding, wordpiece encoding was used both times, as it is faster to train and there is no need for character encoding, as there are no manipulations done on the output.

Neural network model

A sequence to sequence model [2] with Luong attention [3] with concatenation attention score is used.

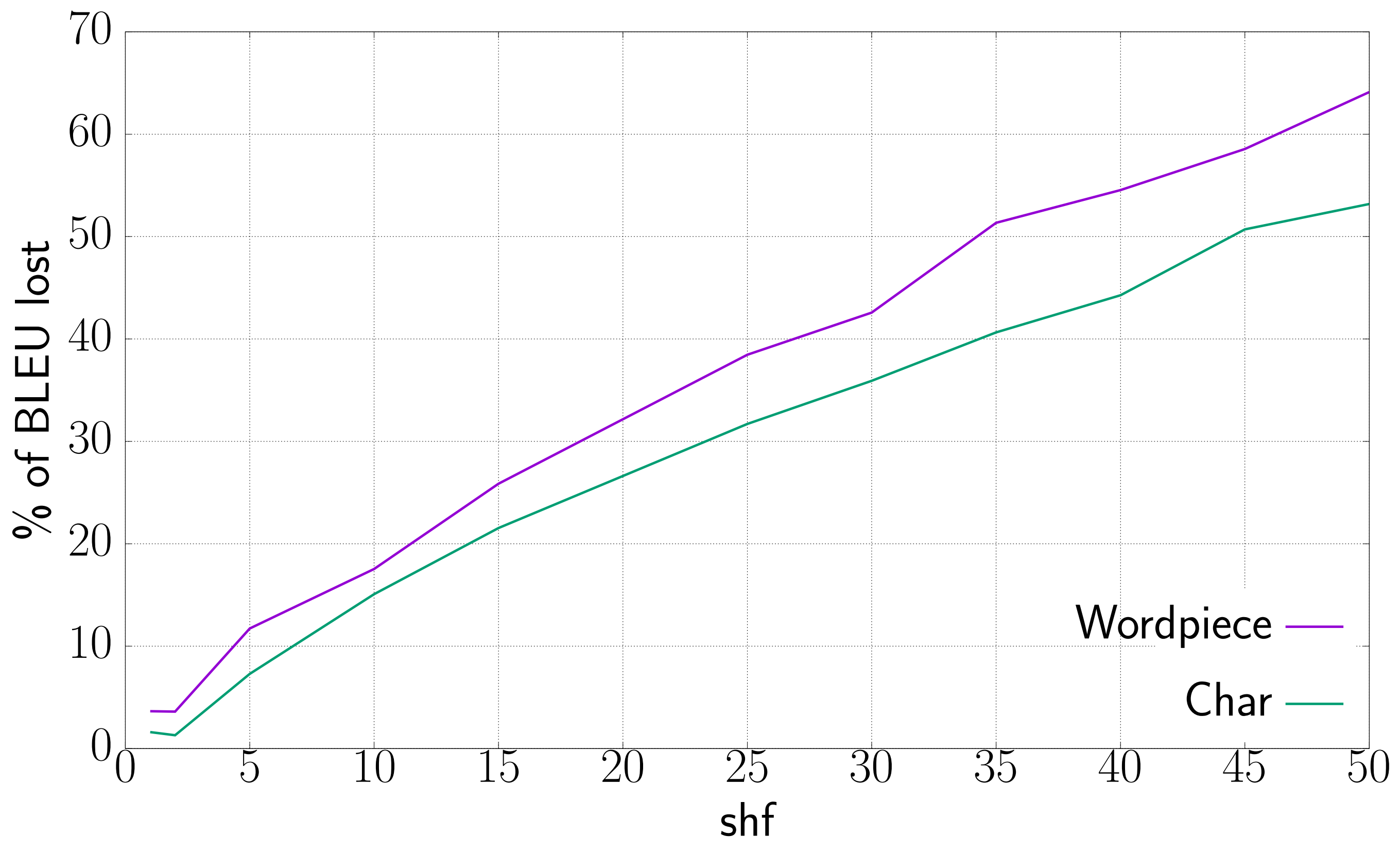
- 2 layer unidirectional GRU cells
- 9 epochs training
- hidden cell size 128
- embedding size 128
- fixed learning rate 0.001 with Adam optimizer



Results

Scores are the percentage of BLEU [4] lost in comparison to cln.

Experiment	Wordpiece	Character
del	10.25	9.83
dbl	8.21	3.18
gl	19.71	19.72
spt	10.63	8.83
ins	7.09	6.31
rnd	12.83	12.23
swp	12.01	9.02



Conclusion

As we can see, the seq2seq network is quite robust to any single type of misspelling. For a 40% emission rate the loss of quality is less than 20%. The character encoding performs generally better than wordpiece, especially in doubling, which is understandable, as the wordpiece configuration might become completely different as a result of doubling of any letter, while the character model is not affected in such a way. For mixed emission, we can find the ratio of loss of BLEU to misspelling emission to be about 1.25 for wordpiece and 1.07 for character, which gives an estimation on how the two models perform in comparison. Considering the amount of misspelling in mixed emissions (each type with stated percentage), the results are comparable to single misspelling emission.

References

[1] `http://opus.nlpl.eu`

[2] Sequence to Sequence Learning with Neural Networks
Ilya Sutskever, Oriol Vinyals, Quoc V. Le, 2014

[3] Effective Approaches to Attention-based Neural Machine Translation,
Minh-Thang Luong, Hieu Pham, Christopher D. Manning, 2015

[4] Bleu: a Method for Automatic Evaluation of Machine Translation,
Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, 2001