A Beyond Error Propagation in Neural Machine Translation: Characteristics of Language Also Matter (Supplemental Material)

A.1 NMT Datasets

The translation datasets we used in our experiments are from five different translation tasks. The details are in the following descriptions.

1) IWSLT 2014 German-English (De-En) (Cettolo et al., 2014) translation task. The dataset contains about $153k$ parallel training sentences, and $6.7k$ sentences for both validation and test set. 2) WMT 2014 English-German (En-De) translation task. The dataset contains about $4.5M$ training pairs[1], $6k$ validation set and $3k$ test set. 3) WMT 2017 English-Chinese (En-Zh) translation task[2]. There are nearly $24M$ sentences in the training set, $2k$ for both validation and test. 4) ASPEC English-Japanese (En-Jp) (Nakazawa et al., 2016) translation, this corpus contains $1.5M$ training samples, nearly $1.8k$ for validation and test set. 5) IWSLT 2014 English-Turkish (En-Tr) (Cettolo et al., 2014) translation dataset, which contains about $350k$ training pairs, $16k$ valid pairs and $7.4k$ test pairs.

For the first three translations, the sentences are preprocessed using byte-pair encoding (Sennrich et al., 2016) into sub-words, while for the En-Jp translation, the sentences are on the word level. For the EN-Tr translation, the dataset is separately processed into morphological segmentation by using Zemberek[3].

A.2 NMT Models

Transformer Model The generation model we used is Transformer (Vaswani et al., 2017), which is based on the self-attention architecture. We use *transofmer_small* setting for De-En and En-Tr, *transformer_base_v1* for En-De and En-Jp, *transformer_big* for En-Zh (Vaswani et al., 2018). For the right-to-left model, we simply reverse the target language sentence as our training data. For example, for De-En translation, we first reverse the target English sentence, and then align the original source German sentence together with reversed English sentence as pair data for training. The models are optimized through Adam as used in the original paper (Vaswani et al., 2017). During decoding phase, we generate the translation sentence by simply greedy search.

RNN Model We also conduct experiments on RNN based models. The RNN models we adopted in Section 5 are GRU based single-layer models, which contain a bidirectional GRU encoder and a unidirectional GRU decoder. For the De-En translation task, the GRU model is a relatively small model, for which the embedding size and hidden size are both set as 256. For the summarization task, the embedding size of the GRU model is 512 and the hidden size is 1024. The models are trained by Adadelta with learning rate 1.0.

A.3 Dependency Parsing Results

The dependency parsing results for English corpus and Japanese corpus are provided here, the examples are as follows.

English parsing For English parsing we use Stanford Parser[4] together with NLTK (Bird et al., 2009):

CASE: *"and the great indicator of that , of course , is language loss"*

PARSING:

[((u'loss', u'NN'), u'cc', (u'and', u'CC')), ((u'loss', u'NN'), u'nsubj', (u'indicator', u'NN')), ((u'indicator', u'NN'), u'det', (u'the', u'DT')), ((u'indicator', u'NN'), u'amod', (u'great', u'JJ')), ((u'indicator', u'NN'), u'prep', (u'of', u'IN')), ((u'of', u'IN'), u'pobj', (u'that', u'DT')), ((u'that', u'DT'), u'prep', (u'of', u'IN')), ((u'of', u'IN'), u'pobj', (u'course', u'NN')), ((u'loss', u'NN'), u'cop', (u'is', u'VBZ')), ((u'loss', u'NN'), u'nn', (u'language', u'NN'))]

Then we can count the dependency words in one tuple that are both from the left half or the right half, e.g., *'indicator'* depends on *'the'* and both belong to the left half.

Japanese Parsing For Japanese parsing we use J.DepP[5]:

CASE: "これらの要素と予精度の特性について明した。"

PARSING:

** 0 1D@0.908514* これら 名,代名,一般,*,*,*,これら,コレラ,コレラ*B@0.000000* の

[1] https://nlp.stanford.edu/projects/nmt/

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

[3] https://github.com/orhanf/zemberekMorphTR

[4] https://nlp.stanford.edu/software/lex-parser.shtml

[5] http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

助,体化,*,*,*,*,の,ノ,ノI@0.000000
* 1 4D@0.000000 要素名,一般,*,*,*,*,要素,ヨウソ,ヨ ソB@0.999910 と助,格助,一般,*,*,*,と,ト,トI@0.000000
* 2 3D@0.993463 予 名,サ 接,*,*,*,*,予,ヨソ ク,ヨソ クB@0.999645 精度名,一般,*,*,*,*,精度,セイド,セイドI@0.028107 の助,体化,*,*,*,*,の,ノ,ノI@0.000000
* 3 4D@0.000000 特性名,一般,*,*,*,*,特性,トクセイ,トクセイB@0.999907 について助,格助,,*,*,*,について,ニツイテ,ニツイテI@0.000000
* 4 -1D@0.000000 明名,サ接,*,*,*,*,明,セツメイ,セツメイB@0.999984 し,自立,*,*,サスル,用形,する,シ,シI@0.014534 た助,*,*,*,特殊 タ,基本形,た,タ,タI@0.000878 。 号,句点,*,*,*,*,。,。,。I@0.001575

The two ids at the begging of each line shows the dependency words. In this case, the last token is “。” with id 4, we simply remove this token when counting the number.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign. In *Proceedings of the 11th IWSLT*.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus.

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, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, François Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.