Latent Part-of-Speech Sequences for Neural Machine Translation

Xuewen Yang[1], Yingru Liu[1], Dongliang Xie[2], Xin Wang[1], and Niranjan Balasubramanian[1,3]

[1]Stony Brook University
[2]Beijing University of Posts and Telecommunications
[1]{xuewen.yang,yingru.liu,x.wang}@stonybrook.edu
[2]xiedl@bupt.edu.cn
[3]niranjan@cs.stonybrook.edu

1 Appendix

1.1 Implementation and Training Details

We implement all Transformer-based models using `Fairseq` [1] Pytorch framework.

For all translation tasks, we choose the *base* configuration of Transformer with $d_{model} = 512$. During training, we choose Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$. The initial learning rate is 0.0002 with 4000 warm-up steps. The learning rate is scheduled with the same rule as in (Vaswani et al., 2017). Each batch on one GPU contains roughly 2000 tokens for IWSLT tasks and 800 tokens for the WMT En-De task. We train IWSLT tasks using two 1080Ti GPUs and train WMT task using 8 K80 GPUs. The hyperparameter λ is set to 0.2. For inference, we use beam search with beam size 5 to generate candidates.

1.2 Dataset Details

We evaluate our model on two small translation datasets - IWSLT'14 German-English (De-En) and English-French (En-Fr) (Cettolo et al., 2015) and a much bigger one - WMT'14 English-German (En-De).

IWSLT'14 En-De/En-Fr We use the datasets extracted from IWSLT 2014 machine translation evaluation campaign (Cettolo et al., 2015), which consists of 153K/220K training sentence pairs for En-De/En-Fr tasks. For En-De, we use 7K data split from the training set as the validation set and use the concatenation of dev2010, tst2010, tst2011 and tst2012 as the test set, which is widely used in prior studies (Huang et al., 2018; He et al., 2018; Bahdanau et al., 2017; Ranzato et al., 2016). For En-Fr, the tst2014 is taken as the validation set and tst2015 is used as the test set, which is the same with prior studies (Denkowski and Neubig, 2017; Cheng et al., 2018). We also lowercase the sentences of En-De and En-Fr following general practice. Before encoding sentences using sub-word types based on byte-pair encoding (Sennrich et al., 2016), which is a common practice in NMT, we parse POS tag sequences of the sentences using Stanford Parser (Chen and Manning, 2014). The POS tag sequences produce POS vocabulary of size 32 for both English and French and 32 for German. Sentences are then encoded using sub-word types. To make the lengths of POS tag sequences equal to their corresponding sub-word sentences, if several sub-words belong to the same word, they are given the same POS tag. For IWSLT'14 En-De dataset, we build a English sub-word vocabulary of size 6632 and a German sub-word vocabulary of size 8848. For En-Fr dataset, we build a English sub-word vocabulary of size 7172 and a French sub-word vocabulary of size 8740.

WMT'14 English-German (En-De)

We use the same dataset as (Vaswani et al., 2017), which consists of 4.5M sentence pairs. We use the concatenation of newstest2012 and newstest2013 as the validation set and newstest2014 as the test set. Sentences are encoded using byte-pair encoding with a shared vocabulary of about 40K sub-word tokens. The method to generate POS tag sequences is the same, except that we merge some POS tags of similar meaning to one and get a POS tag vocabulary of size 16 for both German and English. This operation reduces computational cost, and gives us a bigger batch for training.

[1]https://github.com/pytorch/fairseq

References

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations 2017*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2015. Report on the 11 th iwslt evaluation campaign , iwslt 2014. In *Proceedings of IWSLT 2014*.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1756–1766.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems 31*, pages 7944–7954.

Po-Sen Huang, Chong Wang, Sitao Huang, Dengyong Zhou, and Li Deng. 2018. Towards neural phrase-based machine translation. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

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*, pages 1715–1725.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.