1.1 Citation – 1 point per paper

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
Cite as: **arXiv:1609.08144** [cs.CL]
(or **arXiv:1609.08144v2** [cs.CL] for this version)
https://www.arxiv-vanity.com/papers/1609.08144/

1.2 Task – 2 points per paper

Google's Neural Machine Translation system is an end-to-end learning approach for automated translation, which attempts to address issues like robustness (input sentences with rare words), computational expenses in training and translation inference on large data sets and the weakness of traditional methods. This is done by a model consisting of deep LSTM network with 8 encoder and 8 decoder layers using residual and attention connections. And employs low-precision arithmetic during inference computations and beam search techniques for normalization procedures to increase the translation quality.

1.3 Data – 2 points per paper

The datasets used were publicly available corpora WMT'14 En->Fr (training set includes 36M sentence pairs) and En->De (training set includes 5M sentence pairs) for GNMT models with word, character, word-piece based vocabularies. In addition to this, the GNMT was tested on Google's translation production corpora (which is 2-3 decimal order of magnitude bigger than WMT corpora).
Development set = newstest2012+ newstest2013 Test set = newstest2014

1.4 Approach – 5 points per paper

In the implementation, the model consists of encode (bidirectional LSTM at bottom layer), decode (RNN+softmax layer) and attention networks. the recurrent networks are deep stack (better then shallow) Long Short-Term Memory (LSTM) RNNs. Our LSTM RNNs have 8 layers, with residual connections between layers to encourage gradient flow. For parallelism (model and data), they connect the attention from the bottom layer of the decoder network to the top layer of the encoder network and train n model replicas concurrently using a Downpour SGD algorithm. To improve inference time, they employ low-precision arithmetic for inference, which is further accelerated by special hardware (Google's Tensor Processing Unit, or TPU). To effectively deal with rare words, they use sub-word units (also known as "wordpieces") for inputs and outputs in our system. Using wordpieces gives a good balance between the flexibility of single characters and the efficiency of full words for decoding and sidesteps the need for special treatment of unknown words. Their beam search technique includes a length normalization procedure to deal efficiently with the problem of comparing hypotheses of different lengths during

decoding, and a coverage penalty to encourage the model to translate all the provided input.

1.5 Evaluation – 5 points per paper

The model GNMT was evaluated using BLEU score matrix. As it doesn't capture the quality of translation fully, side by side (S*S) evaluation was carried on where the human raters evaluate and compare the quality of two translations presented side by side for a given source sentence. And it ranges from 0(completely non sense translation) to 6(perfect translation).

The models in experiments are word-based, character-based, mixed word-based or several wordpiece models with varying vocabulary sizes (BLEU increases with increase in vocabulary size and decreasing speed for training). The above step resulted in 38.95 BLEU score with WMP-32K(En->Fr). The second step, RL training which was used to fine-tune sentence BLEU after normal MLE training increased the score by 1. Outcome of ensembling 8 RL-refined models in the next step was 41.16 BLEU.

Finally, to better understand the quality of model and the effect of RL refinement, 4-way side-by-side human evaluation and BLEU scores were compared on phrase based statistical translation and NMT with or without RL refined models. All the GNMT models are wordpiece models, without model ensembling, and use a shared source and target vocabulary with 32K wordpieces. The results show that our model reduces translation errors by more than 60% compared to the PBMT model on these major pairs of languages. And the results between GNMT and human cases are indistinguishable.