1 Which datasets are used in training and evaluation (size, language, etc.)?

Dataset used was provided by ACL WMT '14, which is a parallel corpora. It contains the following English-French parallel corpora: Europarl with 61M words, news commentary with 5.5M words, UN (United Nations) with 421M words and two crawled corporas with sizes of 90M and 272.5M words. Total words is 850M words which was reduced to final corpus of size 348M words. `news-test-2012` and `news-test-2013` parts of the corpora as validation data, this kind of protocol was suggested by the providors of the dataset. Testing was done on `news-text-2014` which consists of 3003 sentences not present in the training data. Training of the model was based on tokenized dataset, training data consisted of 30,000 most frequent words in each language, with no special preprocessing to it. Languages used were English, French, Spanish, German, Czech, Russian, Hindi

2 Which neural model architectures are evaluated in the paper?

Most of the proposed neural machine translation models belong to a family of encoder-decoders. The baseline model used in comparision is an RNN Encoder–Decoder with LSTM cell called RNNencdec. The proposed model by Bahdanau et al. contains bidirectional RNN as an encoder, attention layer, and RNN as decoder called RNNsearch from here on out. The research team trained two models from each architecture, one trained with sentences up to 30 words and one up to 50 words. RNNencdec contains 1000 hidden units in both encoder and decoder part. The RNNsearch has 2000 hidden units in encoder as it is bidirectional that comes up to 1000 hidden units per neural network, in the decoder part it has 1000 hidden layers. Both models use multilayered networkd with a single maxout hidden layer to compute the conditional probability of each target word.

3 Briefly summarize the results and the used metric. Did the neural attention improve the results compared to the previously existing models?

3.1 Quantitative results

Comparision of qualitative results was is based on performance measured in BLEU[1] score. BLEU score was used because it is shown to correlate with human judgement during translation from various languages to english. To interpret the BLEU scores metric i studied Google's documentation[2] and scale presentation attached in there[3]. This helps to summarize the results further. In page 5 of the whitepaper it is shown that RNNsearch-50 (new model with sentence length 50 words) holds steady ∼27 BLEU score for sentences up to 60 words, while other RNNsearch-30 starts to diminish after approx. 20 words. The both variations of RNNencdec peak at ∼22 BLEU before 20 words and start to diminish.

To interpret BLUE scale i attached here related description from Google:

[1] `https://aclanthology.org/P02-1040.pdf`

[2] `https://cloud.google.com/translate/automl/docs/evaluate`

[3] `https://www.cs.cmu.edu/%7Ealavie/Presentations/MT-Evaluation-MT-Summit-Tutorial-19Sep11.pdf`

BLEU score	Interpration
< 10 10 - 19	Almost useless Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30-40	Understandable to good translation

Both models and variations were compared not only eachother but to Moses, which is a conventional phrase-based translation system. In training Moses uses a separate monolingual corpus consisting of 418M words *in addition* to the parallel corpora used by RNNencdec and RNNsearch models. The performance of RNNsearch is on level with Moses' scores, and with no additional monolingual corpora used this is consideret significant achievement.

3.2 Qualitative results

By the result graphs attached in the paper page 6, it is clear to see visibile strong attention weighting between related words. The RNNsearch-50 was also much better than the novel version in translating long sentences and containing the information needed.