Neural Machine Translation by Jointly Learning to Align and Translate
by Bahdanau, Dzmitry et al.

Bahdanau et al. introduce a new approach for machine translation using an attention model. The model is a bidirectional-LSTM that outputs 2 hidden states per input location. The approach obtains different vector representations at each time step, instead of the then-common approach of obtaining fixed vector representations. An issue with older approaches using fixed length vectors was that it wasn't possible to assume all pertinent information was encoded into the one vector. Now the decoder depends on the attention vector, and as the title suggests, the weights are learned jointly within the architecture.

The paper discusses using a BGRU network of 1000 hidden units. The models were trained for 5 days, and outputs were generated via beam search. Attention can be expensive, but it improves performance, particularly with longer sentences (as discussed in the week prior). When words were translated from English to French (used parallel corpora by ACT WMT '14), the attention matrix showed positive results in alignment between the languages by word order. Their RNNsearch model outperformed the basic RNN encoder-decoder models proposed by Cho et al. in 2014.