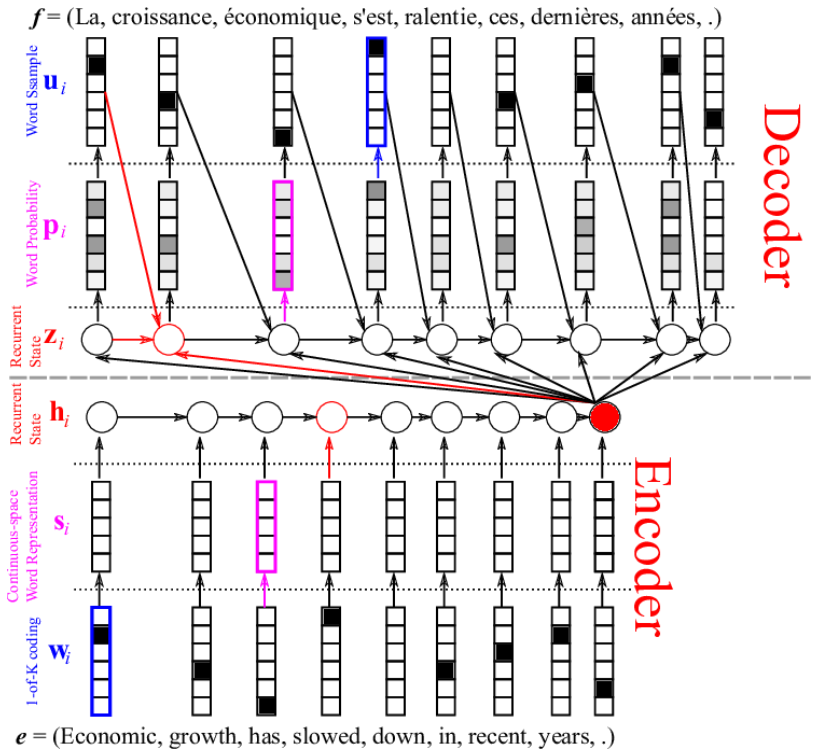
MT Summer Term 2021 Ex13: Seq2Seq NMT with RNN and Attention

1. In your own words, please explain the following pictorial representation of an RNN-based seq2seq NMT system (source: Bahdanau et al. ICLR 2015):



Please make sure you cover word embeddings, contextualised embeddings, the hidden states of the encoder and decoder, the context variable etc. Where is the bottleneck here?

Can you relate the following equations to this:

NMT: encoder-decoder seq2seq RNN

- $\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x})$ where $\mathbf{x}$ is source sequence and $\mathbf{y}$ the target sequence

Encoder:

- Input sentence $\mathbf{x}$ as sequence of input vectors $(x_1, \cdots, x_{T_x})$
- RNN with hidden state $h_t = f(x_t + h_{t-1})$
- Vector/sequence of hidden states $c = q(\{h_1, \cdots, h_{T_x}\})$
- f, q non-linearities
- c context vector
- Sutskever et al. (2014) LSTM as f and $q(\{h_1, \cdots, h_{T_x}\}) = h_{T_x} = c$

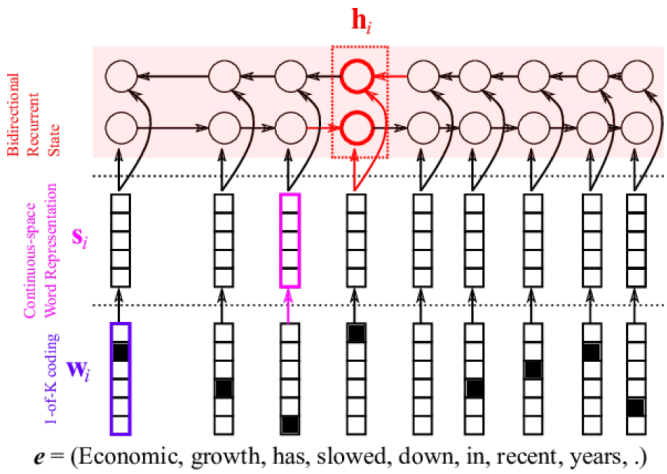
Decoder:

– Predict next word $y_{t'}$, given context vector c and previously generated sequence $\{y_1, \cdots, y_{t'-1}\}$

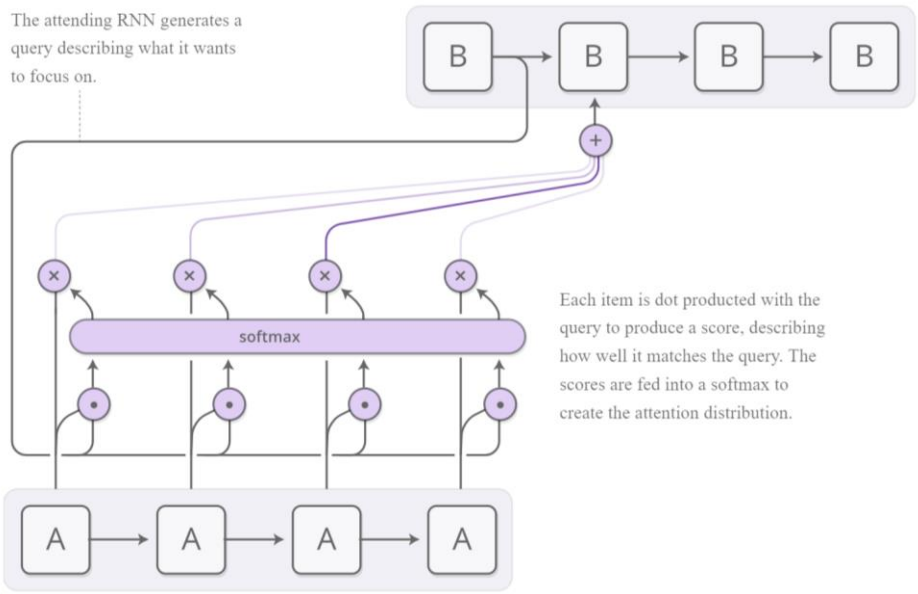
– and $p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t | \{y_1, \cdots, y_{t'-1}\}, c)$

– where $\mathbf{y} = \{y_1, \cdots, y_T\}$ and $p(y_t | \{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$

2. What is the advantage of using a bidirectional RNN in the encoder as depicted below (source: Bahdanau et al. ICLR 2015):



3. In your own words, please explain how attention works in RNN-based seq2seq NMT (Olah & Carter, "Attention and Augmented Recurrent Neural Networks", Distill, 2016):



and please relate the following formulas to this picture:

- $p(y_i | y_1, \cdots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$
- where s_i is a decoder RNN hidden state: $s_i = f(s_{i-1}, y_{i-1}, c_i)$

- c_i is a different (!) c_i for each target word y_i
- encoder RNN maps input to seq. of annotations h_i: $h_i = \left[\overrightarrow{h_i}\,;\,\overleftarrow{h_i}\right]^{\mathrm{T}}$ (bi-directional RNN)
- context vector $c_i = \sum_{j=1}^{T_x} \alpha_{i,j} h_j$
- where $\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T_x} \exp(e_{i,k})}$ and $e_{i,j} = a(s_{i-1}, h_j)$

4. In your own words, please explain (source: Bahdanau et al. ICLR 2015):

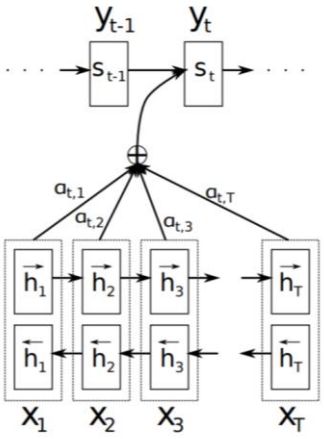
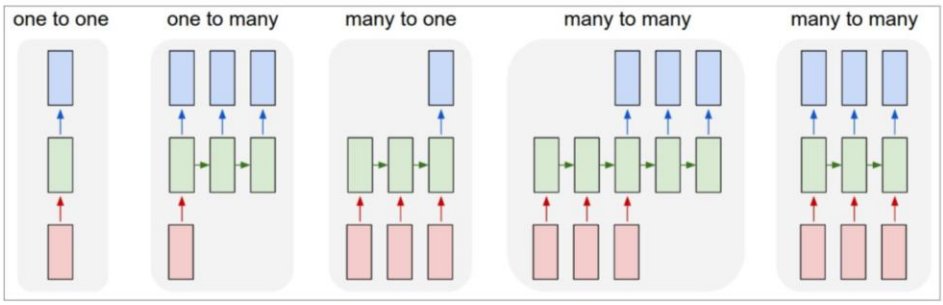


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

5. In your own words, explain the slogan "embed, encode, attend and decode" in seq2seq RNN based encoder-decoder systems. In what sense were they the "Swiss Army Knife" of NLP? Can you find NLP examples of the following (http://karpathy.github.io/2015/05/21/rnn-effectiveness/):



6. Please read the Bahdanau et al. ICLR 2015 paper.

7. In your own words, please explain byte-pair encoding and what problem it solves.