Document-Context Language Models

Jacob Eisenstein

In the RNN language model of Mikolov (2010), we have

$$\mathbf{s}_n \leftarrow f(\mathbf{1}_{w_n}, \mathbf{s}_{n-1}) \quad (1)$$

$$w_{n+1} \sim \text{SoftMax}(\boldsymbol{V}\mathbf{s}_n), \quad (2)$$

where $f(\mathbf{1}_{w_n}, \mathbf{s}_{n-1}) = \sigma(\boldsymbol{U}[\mathbf{1}_{w_n}^\top, \mathbf{s}_{n-1}^\top]^\top)$; fancier RNNs such as LSTM and GRU use more complex gating systems in this function, but the basic idea is the same.

We can extend this model to incorporate a document-context vector $\mathbf{h}_{i-1}$, by specifying

$$w_{n+1} \sim \text{SoftMax}(\boldsymbol{V}_s\mathbf{s}_n + \boldsymbol{V}_d\mathbf{h}_{i-1}), \quad (3)$$

where the vector $\mathbf{h}_{i-1}$ summarizes the document context at sentence $i-1$. There are several options for computing this vector, but it should depend on (a) the words in the sentence $\mathbf{w}_{i-1}$, and (b) the previous document context $\mathbf{h}_{i-2}$. Therefore, we propose another RNN-style model,

$$\mathbf{w}_i \leftarrow g(\mathbf{w}_i) \quad (4)$$

$$\mathbf{h}_i \leftarrow f(\mathbf{w}_i, \mathbf{h}_{i-1}), \quad (5)$$

where the function f again defines an RNN, LSTM, GRU, etc. The function $g()$ might indicate a convolutional neural network, or could also indicate some kind of left-to-right structure.

The first key empirical question (helpfully raised by Lingpeng) is whether this model is any better than the original RNN, or a sentence sentence-sensitive ver

Having established the basic DCLM framework, let's now consider some extensions, before

References

Tomas Mikolov. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.