A Quick Summary: Enriching Word Vectors with Subword Information

Original Paper: http://aclweb.org/anthology/Q17-1010

1 March 2019

1 Ideas:

(a) Word embeddings allow us to obtain a continuous representation of words.

(b) The model introduced in this paper is the Subword Model, which builds on the continuous skip-gram model. This new model learns embeddings of n-grams of each word, with the embedding of that word evenually being the sum of the n-grams that comprise it.

2 Explanations:

(a) With such a representation, we would be able to derive certain interesting structures in the language, for example:

`King - man + woman = Queen`

The objective function is given by:

$$\sum_{t=1}^{T} \Big(\sum_{c \in C_t} log(1 + e^{-s(w_t, w_c)}) + \sum_{n \in N_{t,c}} log(1 + e^{s(w_t, n)}) \Big)$$

Think of s as a scoring function, which returns higher values for words that are more likely to appear in the context of the target word. So we want to maximize s in the first term and minimize it in the second term.

(b) For example, the word "where" will be represented by the character n-grams

`<wh, whe, her, ere, re>, <where>`

and the scoring function will be given by:

$$s(w, c) = \sum_{g \in G_w} z_g^T v_c$$

where G_w is the set of the n-grams, and z_g is the embedding of the n-gram g.

3 Results:

(a) Performs slightly higher than the baselines that do not take into account substrings of individual words.

4 Notes:

(a) This does not seem particularly reasonable - why should there be any similarity between the "her" in the word "where" and the word "ether".

(b) I'm not that confident that this might lead to deeper insights.