A Quick Summary: Finding Function in Form Compositional Character Models for Open Vocabulary Word Representation

Original Paper: https://arxiv.org/pdf/1508.02096.pdf

8 March 2019

1 Ideas:

(a) Word vectors should not have an independence assumption, in particular in morphologically rich languages. (e.g. write and writing)

(b) Similar forms is neither necessary nor sufficient for similarity in function (e.g. course vs coarse, batter and bitter).

(c) The Compositional Character to Word (C2W) model presented in this paper passes the character embeddings of a word into a bidirectional LSTM to obtain a word embedding. These embeddings are used for downstream tasks (like language modelling).

(d) The intention is to get the LSTM parameters to encode idiosyncratic lexical and regular morphological knowledge.

2 Model:

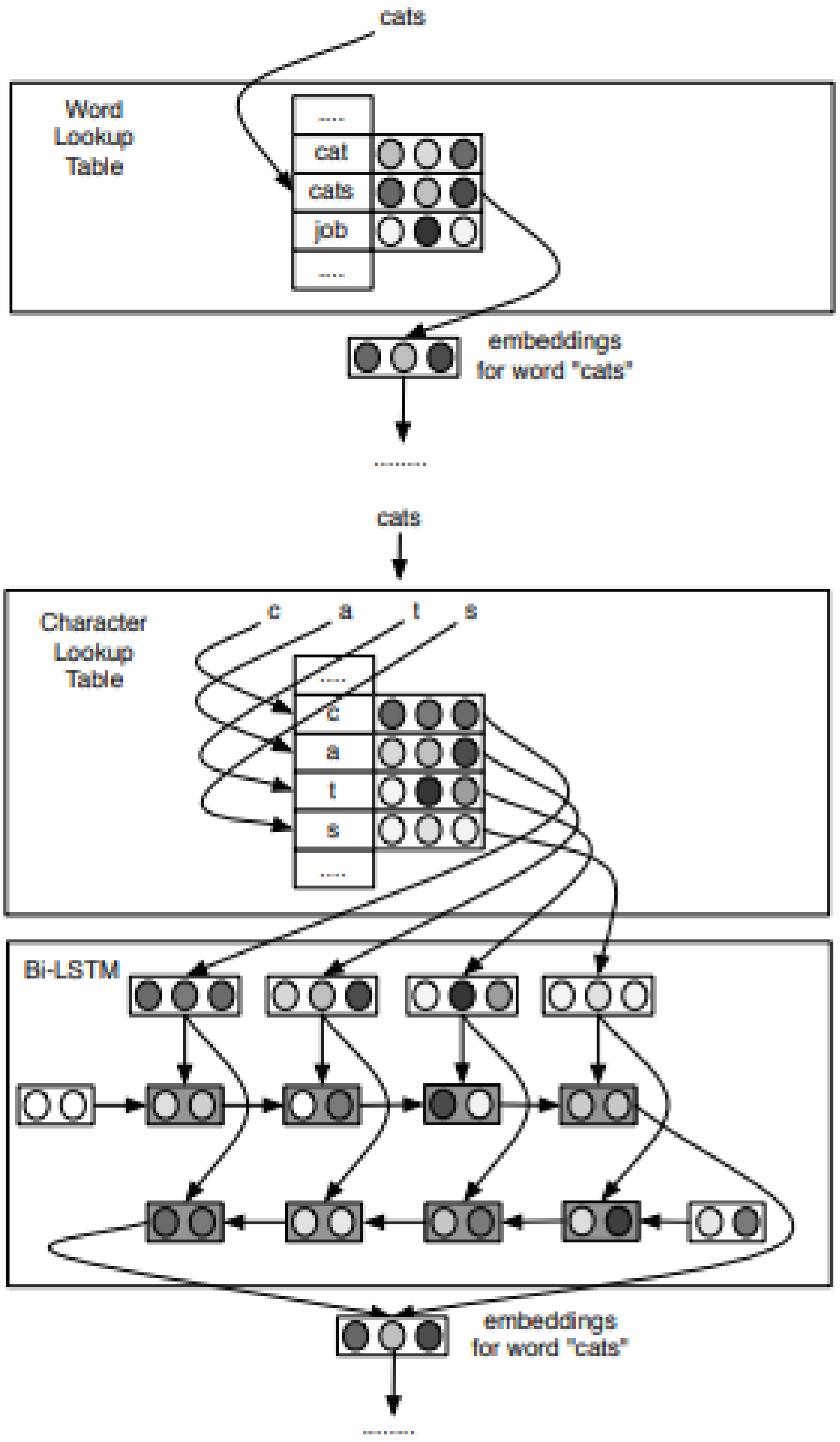

Figure 1: How words are encoded as embeddings (from original paper)

3 Results:

(a) For the task of language modeling, the model achieved a lower perplexity for the Fusional and Agglutinative languages studied as compared to a 5-gram and Word model, with significant improvements for Turkish.

   i Fusional languages can be understood to be languages where a single inflectional morpheme is used to denote multiple grammatical, syntactic, or semantic features.

   ii Agglutinative languages are characterised by having words that may contain different morphemes to determine their meanings, but all of these morphemes (including stems and affixes) remain, in every aspect, unchanged after their unions.

4 Notes:

(a) Seems interesting - I think the marriage of formal (orthographic) and contextual information is a step in the right direction for word representations.