Spoken sentence embedding from character by jointly learning Character-level Compositional word model and RNN sentence encoder

Geonmin Kim[1], Hwaran Lee[1], Jaemyung Yu[2], Soo-Young Lee[1]
[1]Dept. of Electrical Engineering, Korea Advanced Institute of Science and Technology
[2]School of Computing, Korea Advanced Institute of Science and Technology

E-mail: {gmkim90, hwaran.lee, jaemyung, sylee}@kaist.ac.kr

Abstract:
The advent of distributional representation of sentence improves performance of many tasks in natural language processing. However, there is lack of approaches to deal with distributional representation for spoken sentences, which have more irregularities such as partial word and out of vocabulary words compared to written sentences. This irregularities make many rare words in word dictionary, and the number of spoken utterances is usually not enough to train each irregular words. Therefore, estimation of embedding of those words becomes poor, resulting in degraded quality of sentence embedding.

We hypothesize that employing character-level word composition model, which estimates embedding of word by non-linear composition of character embedding, helps to estimate embedding of irregular spoken words which share many sub-words with other regular words. Composition of those sub-words can be learned from several word examples.

Based on our hypothesis, we propose an end-to-end sentence embedding model from characters composed of two neural architectures: a Character-level word Compositional model, and a sentence encoder. For the word compositional model, we employ Convolution Neural Network and Highway network, in order to capture composition of sub-words to build words. For the sentence encoder, we employ Bi-directional Long Short Term Memory Recurrent Neural Network encoder to make fixed size vector from variable length sentences. Those two architectures are jointly trained to optimize the task.

The proposed model is evaluated on Switchboard Dialogue Act classification task. Several different input units (word, composition word, character), and sentence encoding methods (last hidden, mean hidden) are compared. In conclusion, building sentence embedding with compositional word model achieves the best result as 72.09% accuracy on test set, which are relatively 8.20% and 1.17% improved than when sentence encoder sequentially process word sequence, and character sequence. Also it is relatively 1.54% improved from current state of the art result that used additional dialogue history information.

Keywords: spoken utterance representation, word composition model, RNN sentence encoder, last hidden encoding, mean hidden encoding, dialogue act

Acknowledgement

This work was supported by the ICT R& D program of MSIP / IITP. [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning]