Character Word Embedding for NLP tasks in Indian languages

Anirban Majumder, Amit Kumar

4 October 2015

1 Abstract

Part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.Many popular software exist in this domain like Standford POS Tagger.

2 Related Work

Work on learning of representations for NLP has focused exclusively on the word level. However, information about word morphology is very crucial for various NLP tasks which has rich syntactic requirements. For that, one should also take into account the characters of the words which encaptures the various syntactic features of the word. Usually, when a task needs morphological or word shape information, the knowledge is included as handcrafted features (Collobert,2011). Recently a paper by Santos et al. 2014[1] has taken a convolutional approach to obtain the characterlevel word embedding from the individual character embedding using CNN and showed that it improves the results for tasks such as POS tagging.

3 Proposal

Most POS tagger typically, use stochastic methods combined with linguistic resources to achieve reasonably good results. In the recent time Word Embeddings have been used as unsupervised approach to achieve results comparable to that of supervised methods which use handcrafted features.But information about word morphology and shape is normally ignored when learning word representations.Character level

embedding can capture the intra-word information specially when dealing with morphologically rich languages.So we propose to use neural network that learns character-level representation of words and associate them with usual word representations to perform POS tagging. We want to extend that approach for Hindi.[2]

4 References

References

[1] Cicero D. Santos and Bianca Zadrozny. "Learning Character-level Representations for Part-of-Speech Tagging". In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. Ed. by Tony Jebara and Eric P. Xing. JMLR Workshop and Conference Proceedings, 2014, pp. 1818–1826. URL: http://jmlr.org/proceedings/papers/v32/santos14.pdf.

[2] Zhiyuan Liu Maosong Sun Huanbo Luan Xinxiong Chen Lei Xu. "Joint Learning of Character and Word Embeddings". In: (2015).