Character Word Embedding for NLP tasks in Indian Languages

Anirban Majumdar, Amit Kumar
Department of Computer Science IIT Kanpur

Advisor: Dr. Amitabh Mukherjee
Department of Computer Science IIT Kanpur

Abstract

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 morphologically rich task such as POS tagging

Previous Work

- Learning Character-level Representations and Using charWNN to extract intraword information by Santos et al.
- Enhanced Word embedding by average addtion of character level embedding by Liu et al.

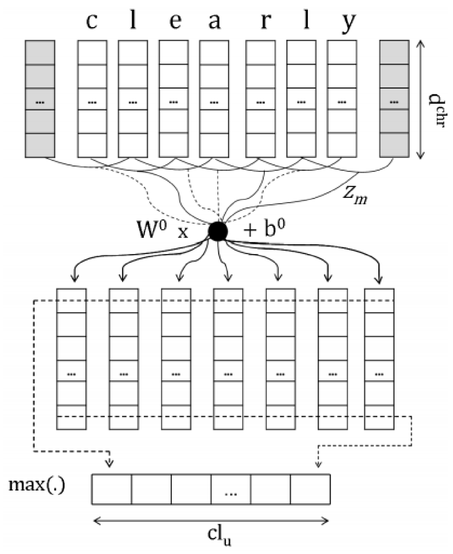

Figure 1: charWNN on Clearly Characters embeddings

Our Methodology

- Wikipedia english corpus (16 million words, Vocab Size: 70k)
- Training data for POS tagger : wikipedia hindi corpus (200 MB)
- Wiki Extractor for cleaning up the corpus github.com/bwbaugh/wikipedia-extractor

Mathematical Section

Mathematical formultation of our approach is explained below:

- Words are segmented in to prefix + root + suffix and heavy weightage is given to suffix and prefixes character embedding
- x_j represent the character level word embedding

$$x_j = \frac{1}{N_j} \sum_{i=1}^{n} c_k * w_k$$

Important Result

colorless	Financiallly	Electricity	Publication	Teaching	Principal
colourless	Functionally	Eccentricity	Perturbation	Touching	Pictorial
cordless	Semantically	Intercity	Precaution	Teasing	Municipal
careless	Originally	Elasticity	Recombination	Searching	Political
countless	Fantastically	Plasticity	Proclaimation	Catching	Semiemerical
clueless	Traditionally	Apostolicity	Partition	Threading	Pectoral

Table 2: Table Showing Morphologically Similar Words (English)

ভারতীয়	প্রাপ্ত	প্রযোগ	আলোচনাকালে	আমদানিকে	সরলভাবে
ভরতীয়	পর্যাপ্ত	প্ররোগ	শুনানিকালে	দোকানদারকে	মিলিতভাবে
জাতীয়	সমাপ্ত	প্রিয়োগ	বর্তমানকালে	অন্যদিকে	দলনিরপেক্ষভাবে
উত্তরীয়	ব্যাপ্ত	পুনর্প্রয়োগ	খননকালে	আফ্রিকানকে	জাতীয়ভাবে
কাকতীয়	পুনঃপ্রাপ্ত	স্তরীরোগ	অপহরণকালে	আরেকজনকে	অবিকৃতভাবে
কার্তীয়	ব্যাপ্ত	নাট্যপ্রয়োগ	নির্ভরশীলতাকে	মহাদলকে	বাড়াবে

Figure 2: Morphologically similar words in Hindi and Bengali

Roadmap Followed

- Getting character embedding for English and Hindi
- Generating the character level word embedding
- Segmenting words in to root and affixes
- Finding word similarity using word level embedding and character level embedding
- Concatinating to obtain Character enhanced word embedding
- Comparing LSTM result on word embedding and charCNN embedding
- Trying to use this embedding for POS Tagging

Results

Perplexity	LSTM-Word	LSTM-CharCNN
English(ep=25)	97.6	92.3
Hindi(ep=5)	664.68	601.85

Table 1: Word Prediction Results

Conclusion

We have shown that character level word embedding are very useful for capturing the morphological information of words. But joint embedding of words is not that good for general language model.As a future work we want to further analyse these model and want to test them on other general NLP tasks.

Additional Information

- All the work done in the project is independent of language so can be extended to other languages
- Similar approach could be extend to arabic languages because of their morphological richness
- Character embeddings obtained are not task specific so could be used for other task

References

[1] Cicero D. Santos and Bianca Zadrozny.
Learning character-level representations for part-of-speech tagging.
In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826. JMLR Workshop and Conference Proceedings, 2014.

[2] Zhiyuan Liu Maosong Sun Huanbo Luan Xinxiong Chen, Lei Xu.
Joint learning of character and word embeddings.
2015.

[3] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush.
Character-aware neural language models.
CoRR, abs/1508.06615, 2015.

Acknowledgements

We thank Dr. Amitabh Mukherjee for refining our ideas and constantly guiding us throughout the project. We acknowledge Kundan Kumar for helping us in LSTM CNN implementation.