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 addition of character level embedding by Liu et al.

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 formulation 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
- \( x_j = \frac{1}{N_j} \sum_{i=1}^{n} c_k * w_k \)

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

<table>
<thead>
<tr>
<th>Language</th>
<th>Perplexity LSTM-Word</th>
<th>Perplexity LSTM-CharCNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (ep=25)</td>
<td>97.6</td>
<td>92.3</td>
</tr>
<tr>
<td>Hindi (ep=5)</td>
<td>664.68</td>
<td>601.85</td>
</tr>
</tbody>
</table>

Table 1: Word Prediction Results

Important Result

<table>
<thead>
<tr>
<th>English Morphologically Similar Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>colorless</td>
</tr>
<tr>
<td>countless</td>
</tr>
<tr>
<td>careless</td>
</tr>
<tr>
<td>countless</td>
</tr>
<tr>
<td>clueless</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hindi Morphologically Similar Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>कास्की</td>
</tr>
<tr>
<td>दक्षिण</td>
</tr>
<tr>
<td>उत्तरकालिका</td>
</tr>
</tbody>
</table>

Figure 1: charWNN on Clearly Characters embeddings

Figure 2: Morphologically similar words in Hindi and Bengali

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

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.