We approach part-of-speech tagging for informal, online conversational text using large-scale unsupervised word clustering and new lexical features. Our system achieves state-of-the-art tagging results on both Twitter and IRC data. Additionally, we contribute the first POS annotations guidelines for such text and release a new dataset of English language tweets annotated using these guidelines.

**Model**

Discriminative sequence model (MEMM) with L1/L2 regularization

**Tagger Features**

- Hierarchical word clusters via Brown clustering (Brown et al., 1992) on a sample of 56M tweets
- Surrounding words/clusters
- Current and previous tags
- Tag dict. constructed from WSJ, Brown corpora
- Tag dict. entries projected to Metaphone encodings
- Name lists from Freebase, Moby Words, Names Corpus
- Emoticon, hashtag, @mention, URL patterns

**Results**

Our tagger achieves state-of-the-art results in POS tagging for each dataset:

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Oct27Test</th>
<th>DAILY47</th>
<th>NPSCHATTest</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>91.60</td>
<td>92.80</td>
<td>91.19</td>
</tr>
<tr>
<td>with clusters; without tags, namelists</td>
<td>91.15</td>
<td>92.38</td>
<td>90.66</td>
</tr>
<tr>
<td>without clusters; with tags, namelists</td>
<td>89.81</td>
<td>90.81</td>
<td>90.00</td>
</tr>
<tr>
<td>only clusters (and transitions)</td>
<td>89.50</td>
<td>90.54</td>
<td>89.55</td>
</tr>
<tr>
<td>without clusters, tagdicts, namelists</td>
<td>86.86</td>
<td>88.30</td>
<td>86.26</td>
</tr>
<tr>
<td>Gimpel et al. (2011) version 0.2</td>
<td>88.89</td>
<td>89.17</td>
<td>88.50</td>
</tr>
<tr>
<td>Inter-annotator agreement (Gimpel et al., 2011)</td>
<td>92.2</td>
<td>92.0</td>
<td>93.2</td>
</tr>
</tbody>
</table>

**Tagger Accuracy**

This work: 93.4 ± 0.3

Forsyth (2007): 90.8

**Accuracy on NPSCHATTest corpus**

Tagger: Accuracy

This work: 90.0 ± 0.5

Forsyth (2007): 85.3

**Accuracy on RITTER corpus**

Tagger: Accuracy

This work: 90.5 ± 0.5

Forsyth (2011): 85.3

Tagger, tokenizer, and data all released at:

www.ark.cs.cmu.edu/TweetNLP

**Software & Data Release**

- Improved emoticon detector and tweet tokenizer
- Newly annotated evaluation set, fixes to previous annotations