**Motivation**

1. Hashtags increasingly used to convey the actual message in tweets. Phrases and sentences turned into a hashtag.
2. Word with sentiment may trap inside a multi-word hashtag.
3. Noisy and compact nature of language leads to hashtags very difficult to segment; sometimes depends on context. e.g. #together; “to get her” or “together”?
4. Can we use carefully auto-segmented hashtags for training?

**Methods**

We formulate the problem as a binary classification; if character is at the beginning or inside of a word. Two classifiers used.

- Maximum Entropy Model (MaxEnt)
  - Good at handling high number of features and inter-dependencies
  - Used Maxent toolkit by Le Zhang at github.com/lzhang10/maxent

- Conditional Random Fields (CRFs)
  - Preferred for sequential problems
  - Used CRFsuite by Naoko Okazaki at www.chokkan.org/software/crfsuite/

**Datasets & Baseline**

Three Training Sets:
1. Tw-BOUN: 1M tweets via Search API
2. Tw-Stanford: 1.6M tweets by Sentiment140
3. Hashtags from SNAP Twitter data set

Two Manually Seg. Hashtag Sets:
1. Dev-BOUN (500h) + Test-BOUN (500h)
2. Tw-Stanford(1000h)+Test-Stanford(1268h)

**Results**

<table>
<thead>
<tr>
<th>System</th>
<th>Test-BOUN</th>
<th>Test-Stanford</th>
<th>F1 Acc.</th>
<th>F1 Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>74.0</td>
<td>69.3</td>
<td>63.0</td>
<td>64.3</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>63.6</td>
<td>57.8</td>
<td>70.1</td>
<td>68.2</td>
</tr>
<tr>
<td>SNAP n-grams</td>
<td>68.9</td>
<td>69.7</td>
<td>70.3</td>
<td>68.7</td>
</tr>
<tr>
<td>Word Breaker</td>
<td>84.4</td>
<td>86.2</td>
<td>84.6</td>
<td>83.6</td>
</tr>
</tbody>
</table>

**Conclusion**

- State-of-the-art URL breaker, MS Word Breaker not best at segmenting hashtags.
- MaxEnt better at handling high # of features with less training data.
- CRF requires less complex and much more training data to be competitive w/MaxEnt.
- Auto-segmented hashtags may not contain diverse training cases but still promising.

**Related Work**

- Most research on word segmentation
  - To segment Chinese and Arabic words
  - First, maximum matching + greedy approaches based on dictionary applied
  - Statistical approaches better at handling unknown and ambiguous cases
  - Viterbi algorithm used to choose the best possible segmentation

Research on hashtag segmentation
- No common test set in literature yet.
- Most didn’t report any result or share their test set.
- Many approached it as a part of bigger problem, not a stand-alone one.
- Bansal et al. achieved accuracy of 87.3 by re-ranker based approach.

**Hashtag Training Data**

Hypothesis: Enough number of n-gram occ. in tweets can show us how to segment hashtags.

**Features**

- Each character is a training instance with active features around it (cursor). We group features into two main types.
- Vocabulary-based Features: We use words themselves, their [len+logfreq] form, as well as their CMU word class forms.
- Orthography-based Features: Simple character normalization method applied before extracting features around cursor.

**Future Directions**

1. Local & global context-based features
2. Use language model (LM) alone
3. Use LM’s N-best segmentation as feature
4. Deep learning approach

**Highlighted References**