Named Entity Recognition in Tweets
- An Experimental Study

-- Presented By Haoyu Chen
Basic Info

Conference: Empirical Methods in Natural Language Processing (EMNLP)
Authors’ institution: University of Washington, Seattle, WA
Time: 2011
Google Citation: 305

Accepted Rate -
24% at 2011 EMNLP
Outline

- Background
- Challenges
- Methodology & Evaluation
- Contributions
Review -- Named Entity Recognition (NER)

Jim bought 300 shares of Acme Corp. in 2006.

\[
\text{Jim}_{\text{Person}} \text{ bought 300 shares of } \text{Acme Corp.}_{\text{Organization}} \text{ in } \text{2006}_{\text{Time}}.
\]
Review -- BIO Encoding for NER

For example, we need name **person** and **date** entity, then define new tags.

**Begin:**  **B-PERS, B-DATE**

-- Beginning of a mention of a person/date

**Inside:**  **I-PERS, I-DATE**

-- Inside of a mention of a person/date

**Outside:**  **O**

-- outside of any mention of a named entity

<POS Tagging with restricted Tagset>
Challenges in Tweets

Fresh Big Data

8,842 Tweets sent in 1 second

762,835,646 Tweets sent today

Source: http://www.internetlivestats.com/twitter-statistics/
Challenges in Tweets

- Fresh Big Data
  => Out Of Vocabulary (OOV), More Types of Entities
- Informal and Noisy Text
  => OOV, Uninformative Capitalization

<table>
<thead>
<tr>
<th></th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Hobbit has FINALLY started filming! I cannot wait!</td>
</tr>
<tr>
<td>2</td>
<td>Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for $250</td>
</tr>
<tr>
<td>3</td>
<td>Government confirms blast n nuclear plants n japan...don’t knw wht s gona happen nw...</td>
</tr>
</tbody>
</table>

Table 1: Examples of noisy text in tweets.
Challenges in Tweets

- Fresh Big Data
  => Out Of Vocabulary (OOV) , More Types of Entities
- Informal and Noisy Text
  => OOV, Uninformative Capitalization
- 140 characters Limit
  => Lack of Context

KKTNY in 45min........
Methodology -- System Flow Chart

Shallow Syntax in Tweets

- POS
- Shallow Parsing (Chunking)
- Capitalization

Classifying Named Entities

Segmenting Named Entities
Traditional POS Tagging

- **Data Set, Baseline**
  - Brown Corpus, 0.90
  - Tweets, 0.76

- **Data Set, State-Of-Art**
  - WSJ, 0.97
  - Tweets, 0.80
Improvements on POS

- Apply hierarchical clustering (Brown Clustering) on 52 million tweets to capture lexical variations.

Improvements on POS

- Apply hierarchical clustering (Brown Clustering) on 52 million tweets to capture lexical variations.
- Conditional Random Fields is used to get the help of the context.

Simple Example: I will be there toma.

I will be there [tomorrow Adv]. ✓
I will be there [tomato noun]. X
Improvements on POS

- Apply hierarchical clustering (Brown Clustering) to capture lexical variations.
- Conditional Random Fields is used to get the help of context.

- Add new tags, such as urls, #hashtags, @usernames, and retweets. (100% accuracy)
- To overcome difference in style and vocabulary, they manually annotated 800 tweets as in-domain training data.
- Incorporate out-domain training data, such as IRC.
### Evaluation on POS (4-fold validation)

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Error Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline (NN)</td>
<td>0.189</td>
<td>-</td>
</tr>
<tr>
<td>Word’s Most Frequent Tag</td>
<td>0.760</td>
<td>-</td>
</tr>
<tr>
<td>Stanford POS Tagger</td>
<td>0.801</td>
<td>-</td>
</tr>
<tr>
<td>T-POS(PTB)</td>
<td>0.813</td>
<td>6%</td>
</tr>
<tr>
<td>T-POS(Twitter)</td>
<td>0.853</td>
<td>26%</td>
</tr>
<tr>
<td>T-POS(IRC + PTB)</td>
<td>0.869</td>
<td>34%</td>
</tr>
<tr>
<td>T-POS(IRC + Twitter)</td>
<td>0.870</td>
<td>35%</td>
</tr>
<tr>
<td>T-POS(PTB + Twitter)</td>
<td>0.873</td>
<td>36%</td>
</tr>
<tr>
<td>T-POS(PTB + IRC + Twitter)</td>
<td>0.883</td>
<td>41%</td>
</tr>
</tbody>
</table>

Table 2: POS tagging performance on tweets. By training on in-domain labeled data, in addition to annotated IRC chat data, we obtain a 41% reduction in error over the Stanford POS tagger.
## Evaluation--Shallow Parsing (Chunking)

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Error Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline (B-NP)</td>
<td>0.266</td>
<td>-</td>
</tr>
<tr>
<td>OpenNLP</td>
<td>0.839</td>
<td>-</td>
</tr>
<tr>
<td>T-CHUNK (CoNLL)</td>
<td>0.854</td>
<td>9%</td>
</tr>
<tr>
<td>T-CHUNK (Twitter)</td>
<td>0.867</td>
<td>17%</td>
</tr>
<tr>
<td>T-CHUNK (CoNLL + Twitter)</td>
<td>0.875</td>
<td>22%</td>
</tr>
</tbody>
</table>

Table 4: Token-Level accuracy at shallow parsing tweets. We compare against the OpenNLP chunker as a baseline.
Methodology -- System Flow Chart

- POS
- Shallow Parsing (Chunking)
- Capitalization

Classifying Named Entities
Segmenting Named Entities

Named Entity Recognition
Segmenting Named Entities

- Larger annotated dataset to effectively learn a model of named entities - Randomly sampled 2400 tweets.
- BOI encoding for representing segmentation
- Dictionaries included a set of type lists gathered from Freebase
- Use the result of shallow syntax as input

Note: Freebase is an online collection of structured data and Google's Knowledge Graph is powered in part by it
Evaluation--Segmenting Named Entities

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F₁</th>
<th>F₁ inc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford NER</td>
<td>0.62</td>
<td>0.35</td>
<td>0.44</td>
<td>-</td>
</tr>
<tr>
<td>T-SEG(None)</td>
<td>0.71</td>
<td>0.57</td>
<td>0.63</td>
<td>43%</td>
</tr>
<tr>
<td>T-SEG(T-POS)</td>
<td>0.70</td>
<td>0.60</td>
<td>0.65</td>
<td>48%</td>
</tr>
<tr>
<td>T-SEG(T-POS, T-CHUNK)</td>
<td>0.71</td>
<td>0.61</td>
<td>0.66</td>
<td>50%</td>
</tr>
<tr>
<td>T-SEG(All Features)</td>
<td>0.73</td>
<td>0.61</td>
<td>0.67</td>
<td>52%</td>
</tr>
</tbody>
</table>
Classifying Named Entities

- Distinctive entities => Big training dataset?
- Lack of context => Out-domain knowledge

**Solution:** Leverage large lists of entities gathered from Freebase as a source of distant supervision

**Benefit:** Allow use of large amount of unlabeled data in learning
Classifying Named Entities

- 30% of entities on Twitter are out of Freebase dictionary while 35% of entities on Twitter has multiple meaning types in Freebase.
- **Enhanced Solution:** Topic Model is to discover the hidden thematic structure in docs, and Latent Dirichlet allocation (LDA) is one of the most used topic model.
- **Basic Idea:** Every doc is made of multiple topics. The words in the documents are generated from those multiple topics.
Evaluation--Classifying Named Entities

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority Baseline</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>Freebase Baseline</td>
<td>0.85</td>
<td>0.24</td>
<td>0.38</td>
</tr>
<tr>
<td>Supervised Baseline</td>
<td>0.45</td>
<td>0.44</td>
<td>0.45</td>
</tr>
<tr>
<td>DL-Cotrain</td>
<td>0.54</td>
<td>0.51</td>
<td>0.53</td>
</tr>
<tr>
<td>LabeledLDA</td>
<td>0.72</td>
<td>0.60</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 8: Named Entity Classification performance on the 10 types. Assumes segmentation is given as in (Collins and Singer, 1999), and (Elsner et al., 2009).
Contributions

- Design and implement a complete system for Named Entity Recognition (NER) in Tweets. By optimizing each steps of NER system, it shows a substantially improvement on performance.

- It introduces a new approach to classify named entity by applying distant supervision with Topic Models (), which is able to train large amount of unlabeled dataset.
Thank you!

Any Questions?
Application Demo