Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis

Hang Jiang*, Yining Hua*, Doug Beeferman, Deb Roy | {hjian42, ninghua, dougb5, dkroy} @ mit.edu

Introduction
Processing the noisy and informal language of social media is challenging for traditional NLP tools because such messages are usually short in length and irregular in spelling and structure. Liu et al. (2018) introduced a tweet-based Tweebank V2 (TB2), including tokenization, part-of-speech (POS) tags, and Universal Dependencies, but there is no NER benchmark on TB2. Annotating named entities in TB2 allows researchers to not only train multi-task learning models but also study linguistic relationship between named entities and syntactic labels.

Contributions
➢ Create the Tweebank-NER benchmark
➢ Train and release the Twitter-Stanza pipeline.
➢ Compare Twitter-Stanza against existing models, showing simple neural architecture is effective and suitable for Tweet processing.
➢ Train Transformer-based models to establish a strong baseline on the Tweebank-NER benchmark.
➢ Release our data, models, and code, including Twitter-Stanza and Hugging Face BERTweet models.

Why do we need Tweebank-NER?
➢ Tweebank-NER is still challenging for current NER models (e.g. models pre-trained on WNUT17).
➢ It makes TB2 a complete dataset for multi-task learning.

Annotate Named Entities in Tweebank v2.0
➢ Follow CoNLL 2003 guidelines
➢ Use Qualtrics platform + Amazon Mechanical Turk
➢ Two-stage annotation
  ○ 3 annotators annotate Tweets
  ○ Tweets without consensus to be re-annotated by the first two authors
➢ Adopt token-level pairwise F1 score (70.7) calculated without the O label

Methods for NLP Modeling

Performance on Tweebank-NER
Main findings
➢ The best non-transformer model: Stanza NER model (TB2+W17)
➢ The best transformer model: HuggingFace-BERTweet (TB2+W17)
➢ TB2 and WNUT17 training sets boost the performance

Table 1: Annotated corpus statistics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets</td>
<td>1,639</td>
<td>710</td>
<td>1,201</td>
</tr>
<tr>
<td>Tokens</td>
<td>24,753</td>
<td>11,742</td>
<td>19,112</td>
</tr>
<tr>
<td>Avg. token per tweet</td>
<td>15.1</td>
<td>16.6</td>
<td>15.9</td>
</tr>
<tr>
<td>Annotated spans</td>
<td>979</td>
<td>425</td>
<td>750</td>
</tr>
<tr>
<td>Annotated tokens</td>
<td>1,484</td>
<td>675</td>
<td>1,183</td>
</tr>
<tr>
<td>Avg. token per span</td>
<td>1.5</td>
<td>1.6</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Table 2: Number of span annotations by entity type and Inter-annotator agreement scores in pairwise F1.

<table>
<thead>
<tr>
<th>Label</th>
<th>Quantity</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PER</td>
<td>777</td>
<td>84.6</td>
</tr>
<tr>
<td>LOC</td>
<td>317</td>
<td>74.4</td>
</tr>
<tr>
<td>ORG</td>
<td>541</td>
<td>71.9</td>
</tr>
<tr>
<td>MISC</td>
<td>519</td>
<td>50.9</td>
</tr>
<tr>
<td>Overall</td>
<td>2,154</td>
<td>70.7</td>
</tr>
</tbody>
</table>

Annotating Named Entities in Tweebank v2.0
➢ Follow CoNLL 2003 guidelines
➢ Use Qualtrics platform + Amazon Mechanical Turk
➢ Two-stage annotation
  ○ 3 annotators annotate Tweets
  ○ Tweets without consensus to be re-annotated by the first two authors
➢ Adopt token-level pairwise F1 score (70.7) calculated without the O label

Models
➢ Stanza
➢ Hugging Face (BERTweet + Token Classification)
➢ spaCy, FLAIR, spaCy-transformer

Questions
➢ How do Stanza models perform compared with other NLP frameworks on the core Tweet NLP tasks?
➢ How do transformer-based models perform compared with traditional models on these tasks?

Tokenization + Lemmatization
➢ Stanza (TB2) achieves the SOTA performance
➢ Combining TB2 + UD English-EWT hurt performance

Performance on Syntactic NLP Tasks

Methods for NLP Modeling

Performance on Tweebank-NER
Main findings
➢ The best non-transformer model: Stanza NER model (TB2+W17)
➢ The best transformer model: HuggingFace-BERTweet (TB2+W17)
➢ TB2 and WNUT17 training sets boost the performance

Table 5: Comparison among NER models trained on TB2 vs. WNUT17 on TB2 test in entity-level F1. "HgFace" stands for "HuggingFace".

<table>
<thead>
<tr>
<th>Systems</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>spaCy</td>
<td>52.20</td>
</tr>
<tr>
<td>FLAIR</td>
<td>62.12</td>
</tr>
<tr>
<td>HgFace-BERTweet</td>
<td>73.71</td>
</tr>
<tr>
<td>spaCy-BERTweet</td>
<td>73.79</td>
</tr>
<tr>
<td>Stanza</td>
<td>60.14</td>
</tr>
</tbody>
</table>

References