**Social Media NLP:** domain adaptation and annotated datasets

**Universal Dependencies (UD):** adaptable to different genres and languages

**Our work:** UD v2 on English Social Media

- Annotation: Tweebank v2 (4x larger than v1)
- Pipeline: Distillation for fast/accurate parsing

**Annotation**

- Twitter-specific constructions that are not covered by UD guidelines (cf. Sanguinetti et al. 2017 for Italian)

**Pipeline**

- overcome noise in the annotation
- accurate parsing without sacrificing speed

---

**Annotation Guidelines**

**Tokenization**

Tradeoff between preservation of original tweet content and respecting the UD guidelines.

**Part-Of-Speech**

Conform to UD guidelines in most cases. Use syntactic head’s POS for abbreviations.

**Dependencies**

Identify non-syntactic tokens (see above Fig.)

- *discourse* for *sentiment emoticon*, topical hashtag, and truncated word
- *list* for referential URL conforming UD
- Retweet construction is treated as a whole

**Tweebank v2**

**Data source:** Tweebank v1 + Feb to Jul 2016 Twitter Stream

**Statistics:**

- 18 people involved
- 3,550 annotated tweets
- 4.5 times larger than v1
- POS agreement: 96.6
- Dep. agreement: 88.8 (U) / 84.3 (L)

**Disagreements:**

- POS for named entities
- Syntactically ambiguous tweets
- See our paper for more details

**Twitter-specific Constructions**

- **URL**
- **Ellipsis**
- **Listing of entities**
- **Parataxis sentences**
- **Phrasal abbreviations**
- **Retweet**
- **@-mention (reply)**
- **Hashtag**
- **Truncated words**

<table>
<thead>
<tr>
<th>Annotation</th>
<th>POS tagger</th>
<th>Parser</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Our contribution</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Our contribution</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Our contribution</td>
</tr>
<tr>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Our contribution</td>
</tr>
<tr>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Our contribution</td>
</tr>
<tr>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Our contribution</td>
</tr>
<tr>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Our contribution</td>
</tr>
</tbody>
</table>

**Tokenizer**

- **Tweet tokenization:** contextual dependent and requires adaptation
- **Statistical modeling vs rule-based model**
- **We propose to use biLSTM for tokenization and it performs better**

**POS tagger**

- **We consider the existing POS tags**
- **Rich feature-based** (Owoputi et al., 2013) vs neural tagger (Ma and Hovy, 2016)

**Parser**

- **Annotation:** noisy, complicates the parser training
- **Overcome the noise with ensemble**
- **Ensemble is slow. We do distillation and it’s fast and accurate**

**System** | **F1** | **System** | **Acc.** | **System** | **LAS** | **Kt/s**
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford CoreNLP</td>
<td>97.3</td>
<td>Stanford CoreNLP</td>
<td>90.6</td>
<td>Kong et al. (2014)</td>
<td>76.9</td>
<td>0.3</td>
</tr>
<tr>
<td>Twokenizer</td>
<td>94.6</td>
<td>Owoputi et al., 2013</td>
<td>94.6</td>
<td>Dozat et al. (2017)</td>
<td>77.7</td>
<td>1.7</td>
</tr>
<tr>
<td>Ours biLSTM</td>
<td>98.3</td>
<td>Ma and Hovy, 2016</td>
<td>92.5</td>
<td>Ballesteros et al. (2015)</td>
<td>75.7</td>
<td>2.3</td>
</tr>
<tr>
<td>Ensemble</td>
<td>79.4</td>
<td>Distillation</td>
<td>77.9</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Pipeline Evaluation**

- Tokenization: 98.3, POS tagging: 93.3, UD parsing: 74.0

---

Yijia Liu¹ · Yi Zhu² · Wanxiang Che¹ · Bing Qin¹ · Nathan Schneider³ · Noah A. Smith⁴

¹Harbin Institute of Technology · ²University of Cambridge · ³Georgetown University · ⁴University of Washington

Dataset @ http://tiny.cc/0juzty · Software @ http://tiny.cc/fluzyt