Annotating Italian Social Media Texts in Universal Dependencies

M. Sanguinetti*, C. Bosco*, A. Mazzei*, A. Lavelli**, F. Tamburini***

*Dipartimento di Informatica, Università di Torino
**Fondazione Bruno Kessler, Trento
***FICLIT, Università di Bologna
PoSTWITA-UD:
a collection of Italian texts from Twitter annotated according to the Universal Dependencies format

Goals:

• create a treebank of social media texts
• contribute to the wider debate about social media text processing and analysis
Dataset

Developed by processing and further enriching the **PoSTWITA** corpus: the dataset used for the EVALITA 2016 task on Part-of-Speech tagging of social media

**Composition:**

- dev set: 6,438 tweets (114,967 tokens)
- test set: 300 tweets (4,759 tokens)
Treebank Preprocessing

PoSTWITA

PoSTWITA-UD

all Internet-specific tags converted into SYM

multi-word tokens re-splittered

lemmas and morphological features added using Anita*

*(Tambruni and Melandri, 2012)
• train:
  UD_Italian v.2 (11,699 sentences)

• test:
  1) PoSTWITA test set (300 tweets)
     a) with lemmas and language-specific tags (-LF)
     b) with morphological features (-F)
  2) UD_Italian test set (489 sentences) (-UD)

• evaluation metric: LAS $F_1$
Parsing Experiments

- **train:**
  - UD_Italian v.2 (11,699 sentences)

- **test:**
  - 1) PoSTWITA test set (300 tweets)
    - a) with lemmas and language-specific tags (-LF)
    - b) with morphological features (-F)
  - 2) UD_Italian test set (489 sentences) (-UD)

- evaluation metric: \( LAS \ F_1 \)

### Parser Results

<table>
<thead>
<tr>
<th>Parser</th>
<th>-LX</th>
<th>-F</th>
<th>-UD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATE graph-based</td>
<td>62.53</td>
<td>67.05</td>
<td>91.26</td>
</tr>
<tr>
<td>MATE transition-based</td>
<td>64.92</td>
<td>66.65</td>
<td>91.44</td>
</tr>
<tr>
<td>RBG full</td>
<td>64.36</td>
<td>67.07</td>
<td>90.16</td>
</tr>
</tbody>
</table>
Challenging aspects of Twitter language and its analysis:

• continuous shift from written to spoken language, and *vice versa*
• hashtags, mentions/replies, emoticons/emojis, and other conventions of computer-mediated communication
• unconventional, even unintelligible, elements (e.g. unknown/mispelled words)
Annotation Guidelines

How we dealt with them in annotation:

• (if syntactically integrated) assigning them their corresponding syntactic role
How we dealt with them in annotation:

- extending the already existing relations with specific subtypes
  either new
Annotation Guidelines

How we dealt with them in annotation:

• extending the already existing relations with specific subtypes
  ... or mutated from other treebanks
Future Work

Short-term goals:

• Complete annotation and guidelines (first release expected: November 2017 – UD v.2.1)
• Extend parsing experiments using the resource as training set
• Use the resource for Sentiment Analysis applications

Long-term goals:

• Enrich the resource with texts from other social media
• Open this work to a multilingual comparison
Thank you!