Tokenization: Returning to a long solved problem
A survey, Contrastive Experiment, Recommendations, and Toolkit

Rebecca Dridan & Stephan Oepen

Presented by: Valerio Basile
Alfa-informatica Reading Group 14/9/2012
Motivation

Breaking up natural language text [...] into distinct meaningful units (or tokens)

(Kaplan, 2005)

- Often combined with other string-level preprocessing.
- Example:
  
  I won't go! → I wo n't go !
Motivation
Common Conventions

*Penn TreeBank* style

- Punctuation into separate tokens
- Disambiguating straight quotes
- Separating contractions
  
  can't → ca n't

*Not* universally adopted
A Contrastive Experiment

An overview of current tokenization methods

<table>
<thead>
<tr>
<th>Tokenization Method</th>
<th>Differing Sentences</th>
<th>Levenshtein Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>tokenizer.sed</td>
<td>3264</td>
<td>11168</td>
</tr>
<tr>
<td>CoreNLP</td>
<td>1781</td>
<td>3717</td>
</tr>
<tr>
<td>C&amp;J parser</td>
<td>2597</td>
<td>4516</td>
</tr>
</tbody>
</table>

Total of 49,208 sentences and 1,173,750 gold-standard tokens in the PTB
A Contrastive Experiment

- Ambiguity of sentence-final period
  
  ... in the U.S. (extra period hallucinated)

  C&J: ... in the U.S. .

  CoreNLP: ... in the U.S.

  PTB sed script: ... in the U.S .

- Under-restricted punctuation rules, currencies, Irish names, hyphenation, quote disambiguation, ...
A Generalized Framework

REPP
(Regular Expression-based Pre-Processing)

• Cascade of rewriting rules
• Fixpoint iteration over groups of rules
• e.g. insert whitespace around punctuation marks before splitting tokens
A Generalized Framework

REPP operators

#  group formation
>
!  group invocation
!
:  substitution

:  token boundary detection
A Generalized Framework

REPP example

>wiki
#1
!([\^ ][\[])?,;:"'\])¬([\^ ]|$) \1\2\3
!(^|[^\^])¬([\[\{"'\}])([^\^]) \1\2\3
#
>1
:[[[:space:]]]+

- two rules stripping off prefix and suffix punctuation marks adjacent to whitespace
- rule sets organized as modules (e.g. wiki)
Characterization for Traceability

- Changes to the original text
- *Traceability* is required
token objects → original text
- Character position links
- Tokens as *stand-off* annotation
Characterization for Traceability

• Before processing, natural start and end character position

<table>
<thead>
<tr>
<th>I</th>
<th>w</th>
<th>o</th>
<th>n</th>
<th>'</th>
<th>t</th>
<th>g</th>
<th>o</th>
<th>!</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

• Character links are defined at the boundaries of matched spans

wo(n't) will

<table>
<thead>
<tr>
<th>wo</th>
<th>n't</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>will</th>
<th>n</th>
<th>'</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>
Evaluation

• REPP rules following the PTB conventions
• Initial difference in 1505 sentences (Levenshtein distance of 3543)
• Subsequent refinements lead to 603 different sentences (Levenshtein distance of 1389)
Discussion