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

Tokenization Method	Differing Sentences	Levenshtein Distance
tokenizer.sed	3264	11168
CoreNLP	1781	3717
C&J parser	2597	4516

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

I				w		o		n		'		t				g		o		!	
0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11

- Character links are defined at the boundaries of matched spans

```
!wo(n't)    will␣\1
```

wo			n't		
2		4	4		7

will			n	'	t
2		4	4		7

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