Introduction to Computational Linguistics

Tokenization and Sentence Boundary Detection

Jan-Philipp Söhn

jp.soehn@uni-tuebingen.de

December 12th, 2007

Incremental Linguistic Analysis

- tokenization
- morphological analysis (lemmatization)
- part-of-speech tagging
- named-entity recognition
- partial chunk parsing
- full syntactic parsing
- semantic and discourse processing

Tokenization: Motivation

- Robust NLP
- Processing of large corpora
- Preprocessing step for other applications

Preprocessing the Text: Tokenization

Tokenization refers to the annotation step of dividing the input text into units called *tokens*.

Each token consists of either:

- a morpho-syntactic word
- a punctuation mark or a special character (e.g. &, @, %)
- a number

Why is Tokenization Non-trivial?

- Disambiguation of punctuation
  E.g. period can occur inside cardinal numbers, after ordinals, after abbreviations, at end of sentences

- Recognition of complex words
  - compounds, e.g. *bank transfer fee, US-company*
  - mergers, e.g. clitization in French *t'aime* or English *England's*
  - multiwords, e.g. complex prepositions (*provided that*, *in spite of*)

Tokenization for Japanese

Japanese: the ultimate nightmare for tokenization!

Just take a look: e.g. `http://www.yomiuri.co.jp/index-j.htm`

What is so hard ?

- cannot rely on blank spaces and punctuation
- combination of two writing systems: kanji (Chinese characters) and hiragana (mostly used for marking grammatical endings)
- E.g. WATASHI-wa ("first party", meaning: I); large cap part is in Kanji and remaining part is in hiragana.

Deterministic Tokenization

- If the output never contains alternative segmentations for any part of the input, the tokenizer is called **deterministic**.
- Deterministic tokenization is commonly seen as an independent preprocessing step unambiguously producing items for subsequent morphological analysis.
- Deterministic tokenization is commonly implemented as a cascade of finite-state transducers.

Sentence Segmentation

Task:

Determining how a text should be divided into sentences for further processing.

Terminology:

- sentence boundary detection
- sentence boundary disambiguation
- sentence boundary recognition

Sentences I

O. Dittrich:
Ein Satz ist eine modulatorisch abgeschlossene Lautung, wodurch der Hörende veranlasst wird, eine vom Sprechenden als richtig anerkennbare, relativ abgeschlossene apperzeptive (beziehende) Gliederung eines Bedeutungstatbestandes zu versuchen.

D. Jespersen:
eine (relativ) vollständige und unabhängige menschliche Äußerung, deren Vollständigkeit und Unabhängigkeit sich in ihrem Alleinstehen zeigt, d.h. darin, daß sie für sich allein geäußert wird.

Sentences II

A. Meillet:
eine Gemeinsamkeit von Artikulationen, die untereinander durch gewisse grammatische Beziehungen verbunden sind, grammatisch von keiner anderen Gesamtheit abhängen und sich selbst genügen.

W. Meyer-Lübke:
ein Wort oder eine Gruppe von Wörtern, die in der gesprochenen Sprache als Ganzes erscheinen, die sich als eine Mitteilung eines Sprechenden an einen anderen darstellen.

Sentences III

A. Nehring:
der sprachliche Ausdruck für eine vom Sprechenden jeweils hergestellte Ordnung einer gegebenen Mannigfaltigkeit von Sachverhalten.

W. Porzig:
ein Bedeutungsgefüge von derjenigen Form, durch die (in der betreffenden Sprache) Sachverhalte als abgeschlossene gemeint werden.

A. Stöhr:
eine mehrfache Benennung desselben Geschehnisses durch logisch gleichwertige Satzglieder.

Sentences IV

H. Paul:
der sprachliche Ausdruck, das Symbol dafür, daß sich die Verbindung mehrerer Vorstellungen oder Vorstellungsgruppen in der Seele des Sprechenden vollzogen hat, und das Mittel dazu, die nämliche Verbindung der nämlichen Vorstellungen in der Seele des Hörenden zu erzeugen. Jede engere Definition des Begriffes Satz muß als unzulänglich zurückgewiesen werden.

L. Bloomfield
an independent linguistic form, not included by virtue of any grammatical construction in any larger linguistic form.

Sentences in Real Life 1

There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.

Sentences in Real Life 2

The holes certainly were rough–"Just right for a lot of vagabonds like us," said Bigwig–but the exhausted and those who wander in strange country are not particular about their quarters.

1. Two high-ranking positions were filled Friday by Penn St. University President Graham Spanier.
2. Two high-ranking positions were filled Friday by Penn St. University President Graham Spanier announced the appointments.

Problems with Sentence Segmentation

- Strict punctuation rules might exist, adherence varies
- Different punctuation marks or characters
- Task: disambiguate all punctuation marks that denote sentence boundaries:
  periods, question marks, exclamation point, semicolons, dashes, commas
- Use can vary with text types

Contextual Factors

- Case distinctions (no nominative at the end of a sentence)
- Part of speech (no adjective at the end of a sentence)
- Word length (function words as *and*/*as* are short)
- Lexical endings (to exclude abbreviations)
- Prefixes and suffixes before and after the punctuation mark
- Abbreviation classes