Sentence Segmentation

COMP61332: Text Mining
Week 1
Riza Batista-Navarro

The Task

Determination of boundaries between sentences

Sentences used in subsequent NLP tasks

Is it enough to detect the **full stop**?

- Could be an end-of-sentence (EOS) marker
- Or an end of abbreviation marker
- Or both?

Examples

John loves Mary.

John loves Mary. She does not love John.

Mary loves the Dr. but the Dr. loves Eve.

Mr. Jones loves Mary.

Mr. Jones loves Mary… she loves him too.

Mr. Jones loves Mary, but she doesn’t love him…

You can search at www.google.com

You can search at www.google.com. They’re great.

Challenges

Variation in **delimiters** (also known as end-of-sentence or EOS characters)

Typical : “**.**”, “**!**”, “**?**”

How about: “**;**”, “**,**” and “**—**”

Challenging example

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 rabbithole under the hedge.

(Carroll, Alice in Wonderland)

More challenging examples

“You showed me,” he said, “the way to go”.

“You showed me,” he said, “the way to go.”

... known as “text mining.” (AmEng usage)

...known as “text mining”. (BrEng usage)

Two incidents were reported by Warren St. Station Supervisor Fred Jones.

Two incidents were reported by Warren St. Station Supervisor Fred Jones announced the delays.

Domain dependence

Protein name starting with a lowercase character might confuse normal sentence segmentation approaches.

Results support the observations that ERK activity is required for phospholipid hydrolysis independently of cPLA2 translocation. cPLA2-mediated AA release must be preceded by translocation of the enzyme to its membrane substrate.

(from a biomedical article)

Approaches

Regular expressions (Patterns)

Dictionaries (e.g., abbreviation lists)

Hand-crafted rules (e.g., to check whether the word following an EOS delimiter starts with an uppercase character)

Statistical and ML approaches

Hybrid approaches

Examples of useful rules or features

First character after potential EOS char

- Should be uppercase? Problematic for some languages, e.g. German
- Permissible chars after potential EOS, e.g. lowercase characters?

Abbreviations

- titles not likely to occur at EOS (e.g., Dr. Jones)
- company indicators could occur at EOS (e.g., MySocialMedia Inc.)

OpenNLP Sentence Detection

Sentence Detection

The OpenNLP Sentence Detector can detect that a punctuation character marks the end of a sentence or not. In this sense a sentence is defined as the longest white space trimmed character sequence between two punctuation marks. The first and last sentence make an exception to this rule. The first non whitespace character is assumed to be the begin of a sentence, and the last non whitespace character is assumed to be a sentence end. The sample text below should be segmented into its sentences.

```
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. Mr. Vinken is
chairman of Elsevier N.V., the Dutch publishing group. Rudolph Agnew, 55 years
old and former chairman of Consolidated Gold Fields PLC, was named a director of this
British industrial conglomerate.
```

After detecting the sentence boundaries each sentence is written in its own line.

```
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC,
    was named a director of this British industrial conglomerate.
```

spaCy

Statistical sentence segmenter V3.0

NEEDS MODEL

The `SentenceRecognizer` is a simple statistical component that only provides sentence boundaries. Along with being faster and smaller than the parser, its primary advantage is that it's easier to train because it only requires annotated sentence boundaries rather than full dependency parses. spaCy's trained pipelines include both a parser and a trained sentence segmenter, which is disabled by default. If you only need sentence boundaries and no parser, you can use the `exclude` or `disable` argument on `spacy.load`
sentence recognize

Rule-based pipeline component

The `Sentencizer` component is a pipeline component that splits sentences on punctuation like `.`, `!` or `?`. You can plug it into your pipeline if you only need sentence boundaries without dependency parses.