НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Sentence segmentation

Francis M. Tyers

ftyers@hse.ru
https://www.hse.ru/org/persons/209454856

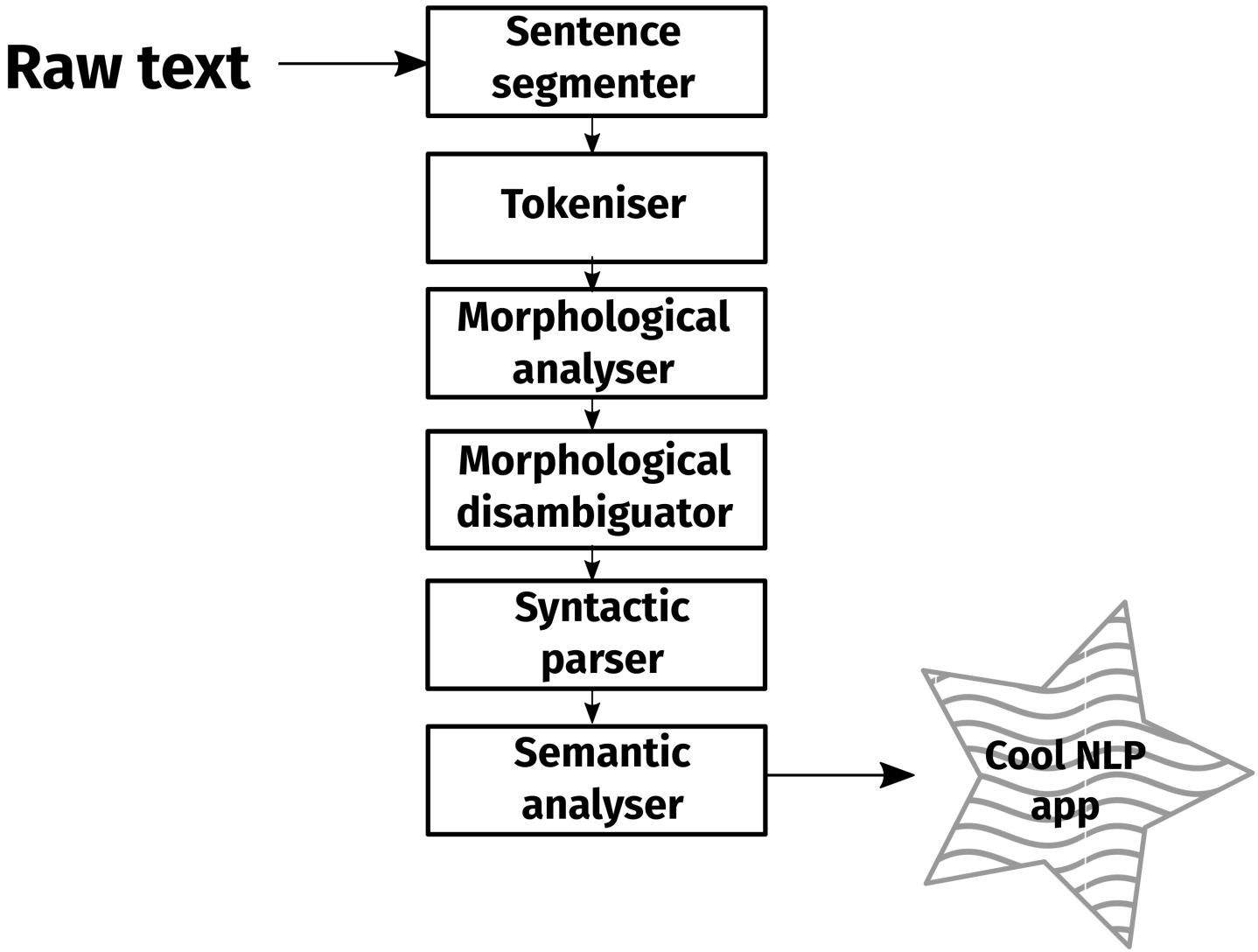
Национальный исследовательский университет
«Высшая школа экономики» (Москва)

1 октября 2018 г.

Introduction

- Raw text
- Processing
- Corner cases

Pipeline


Raw text
Sentence segmenter
Tokeniser
Morphological analyser
Morphological disambiguator
Syntactic parser
Semantic analyser
Cool NLP app

Raw text

- **Ideal world:** Text nicely split into sentences
- **Reality:**
  - Paragraphs
  - Ambiguous punctuation
  - Newlines in the middle of sentences

Why?

- OCR
- Extracted formatted text, e.g. `pdftotext`
- Speech or social media data

Как отмечает историк В. Д. Есаков, «Длительное отсутствие научного руководителя, вызванное проведением экспедиции в страны Средиземноморья, в которой Вавилов пробыл с июня 1926 по август 1927 г., привело к определённому росту бюрократических тенденций в руководстве институтом, росту центробежных устремлений, к критике избранных исследовательских направлений, к упрёкам в отрыве от практики. Встревоженный этими нежелательными в деятельности научного учреждения проявлениями Н. И. Вавилов ставит вопрос об отходе от руководства институтом». 24 ноября 1927 г. он пишет Н. П. Горбунову об отставке: «Ряд событий, имевших место в 1927 году, частью во время моего отсутствия, частью же во время моего пребывания в Ленинграде, заставил меня сильно задуматься над целесообразностью моего пребывания на посту директора Института прикладной ботаники… По внутреннему, глубокому убеждению я не могу считать обвинение в отсутствии руководства правильным. Я принадлежу к числу работников, которые знают наши оба учреждения с самого начала их основания (Отдел прикладной ботаники с 1908 г.). Самый большой плюс нашего объединённого учреждения, по моему убеждению, его исключительная научная спаянность, в большей части работников… Эта спаянность позволила быстро и широко развить работу в области прикладной ботаники… Наша научная коллегия, несмотря на десятки научных работников, которые она включает, представляет спаянное целое, и мы очень редко расходимся в определении направлений работы и развития нашего учреждения. Словом, по внутреннему убеждению обвинений в отсутствии руководства я совершенно принять не могу».

Как отмечает историк В. Д. Есаков, «Длительное отсутствие научного руководителя, вызванное проведением экспедиции в страны Средиземноморья, в которой Вавилов пробыл с июня 1926 по август 1927 г., привело к определённому росту бюрократических тенденций в руководстве институтом, росту центробежных устремлений, к критике избранных исследовательских направлений, к упрёкам в отрыве от практики. Встревоженный этими нежелательными в деятельности научного учреждения проявлениями Н. И. Вавилов ставит вопрос об отходе от руководства институтом». 24 ноября 1927 г. он пишет Н. П. Горбунову об отставке: «Ряд событий, имевших место в 1927 году, частью во время моего отсутствия, частью же во время моего пребывания в Ленинграде, заставил меня сильно задуматься над целесообразностью моего пребывания на посту директора Института прикладной ботаники… По внутреннему, глубокому убеждению я не могу считать обвинение в отсутствии руководства правильным. Я принадлежу к числу работников, которые знают наши оба учреждения с самого начала их основания (Отдел прикладной ботаники с 1908 г.). Самый большой плюс нашего объединённого учреждения, по моему убеждению, его исключительная научная спаянность, в большей части работников… Эта спаянность позволила быстро и широко развить работу в области прикладной ботаники… Наша научная коллегия, несмотря на десятки научных работников, которые она включает, представляет спаянное целое, и мы очень редко расходимся в определении направлений работы и развития нашего учреждения. Словом, по внутреннему убеждению обвинений в отсутствии руководства я совершенно принять не могу».

Два каменных будды,
Нагие, стоят у дороги;

Их ветр овевает
И хлещут дожди и метели.

Завидуй! Не знают
Они человечьей разлуки.

Task

Sentence segmentation, or:

- Sentence boundary disambiguation
- Sentence boundary detection
- Sentence tokenisation

Why?

- Bad sentence segmentation hurts downstream applications

Some applications

- Vital for creating translation memories:
  - If the memory contains paragraphs, essentially all matches will be *fuzzy*
- Also applies to parallel corpora.
  - Good segmentation and tokenisation can increase BLEU score substantially

What is a «sentence»?

- Usually thought of as a predicate + arguments + any coordinated elements

But how about:

- Titles
- Lists
- Reported speech:
  - “I don’t know, she said”
  - She said, “I don’t know”
  - She said: “I don’t know”.
  - “I don’t know”. She said it so softly I could barely hear.

Some examples:

- Amnesty chief spokesman Mike Blakemore said: ”It’s unbelievable that no one can account for 200,000 assault rifles.
- On your marks, get set: ready... steady... go!

Approaches

Replacement rules:

- Replace
- Most often with regular expressions

Or as a binary classifier:

- For each full stop in the sentence,
  - Output EOS or not EOS

A simple solution

```
import sys

line = sys.stdin.readline()
while line != '':
        for c in '?!.':
            line = line.replace(c, c + '\n')
        sys.stdout.write(line)
        line = sys.stdin.readline()
```

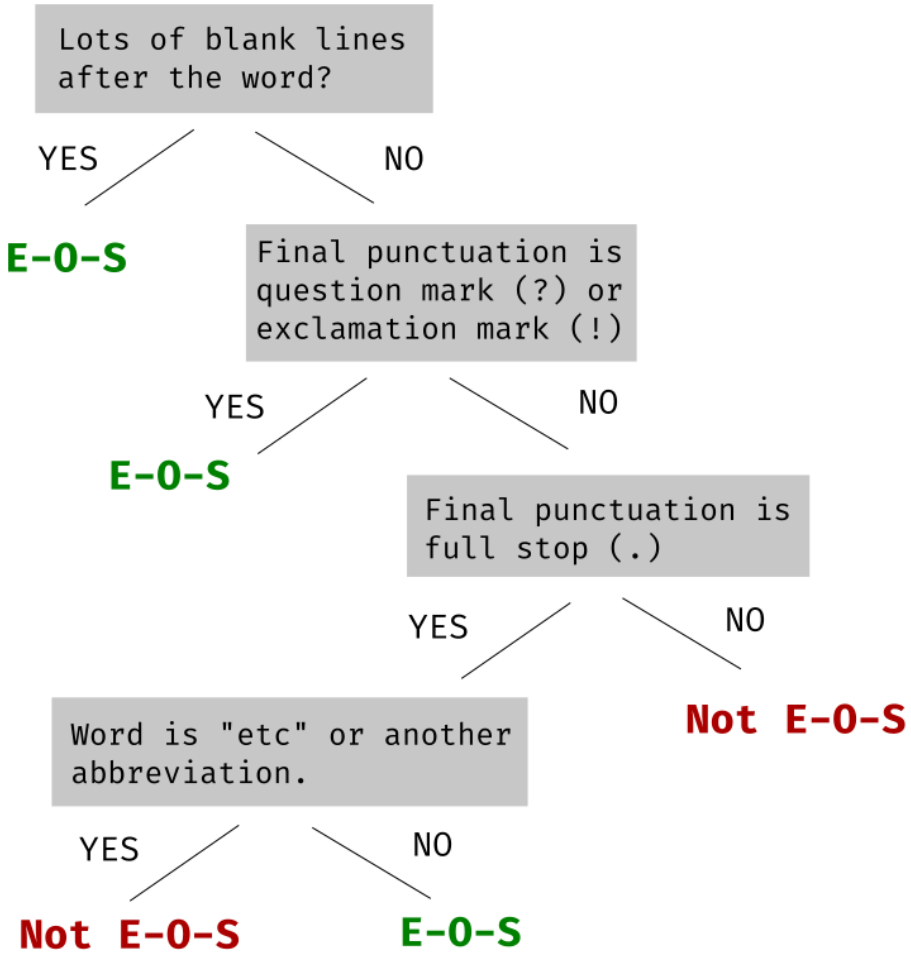
Abbreviations

Different kinds:

- Unambiguous: NATO, kg, cm
- Less ambiguous: Mr., Dr.
- Ambiguous: г., etc.,

A decision tree

How to determine if a word is the end of a sentence:



Implementation

A decision tree can be implemented simply as if–then–else:

```
line = sys.stdin.readline()
while line != '':
    for token in line.split('␣'):
        if token[-1] in '!?':
            sys.stdout.write(token + '\n')
        elif token[-1] == '.':
            if token in ['etc.', 'i.e.', 'e.g.']:
                sys.stdout.write(token + '␣')
            else:
                sys.stdout.write(token + '\n')
        else:
            sys.stdout.write(token + '␣')
    line = sys.stdin.readline()
```

Other features

Word features:

- The word with “.” (Upper, lower, ALLCAPS, [0-9])
- The word after “.” (Upper, lower, ALLCAPS, [0-9])

Numeric features:

- Length of the word with “.”
- Probability that the word with “.” appears at the EOS
- Probability that the word after “.” appears at the BOS

Machine learning

As this is a binary classification task, we can also use machine learning:

- MaxEnt / Logistic regression
- SVM
- Neural networks, etc.

Generating training data

Existing segmented corpus:

- Just concatenate all of the sentences together.

Edge cases

- Sentence without spaces between:
  - *Hello world.Today is Tuesday.Mr. Smith went to the store and bought 1,000.That is a lot.*
- Non sentence boundary with next word capitalised
  - *I work for the U.S. Government in Virginia.*
- A.M. / P.M. as non sentence boundary and sentence boundary
  - *At 5 a.m. Mr. Smith went to the bank. He left the bank at 6 P.M. Mr. Smith then went to the store.*

From
`https://github.com/diasks2/pragmatic_segmenter`

Practical

- Download a dump of Wikipedia and compare two sentence segmenters

Why Wikipedia:

- Text is free to use/download and redistribute
- Challenging text, some tricky edge cases
- Available in many languages

Reading list

- Mikheev, A. (1994) “Periods, Capitalized Words, etc.” *Computational Linguistics* 16(4)
- Kiss, T. and Strunk, J. (2006) “Unsupervised Multilingual Sentence Boundary Detection”. *Computational Linguistics* 32(4)