НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Sentence segmentation

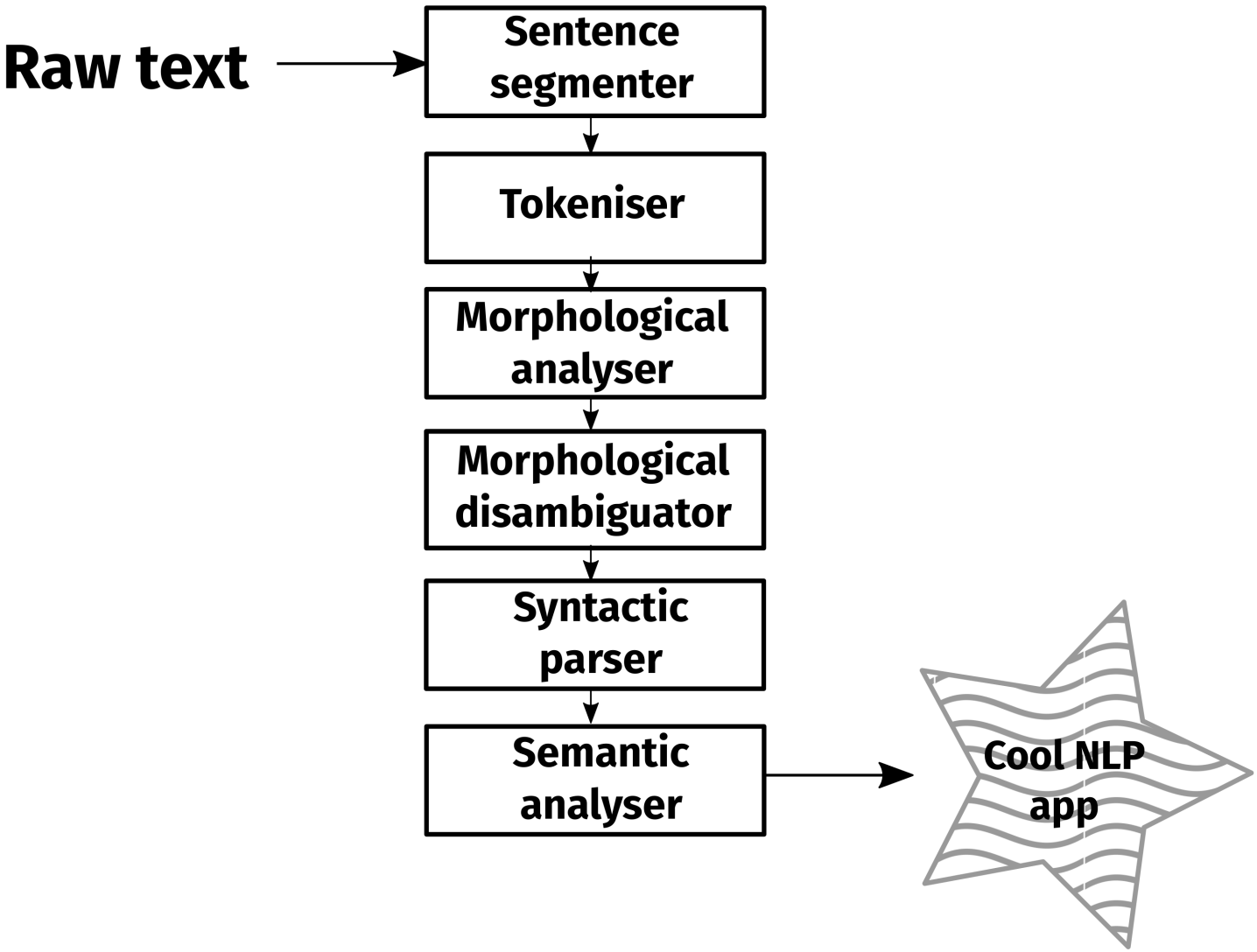
Francis M. Tyers

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

Национальный исследовательский университет
«Высшая школа экономики» (Москва)

11 октября 2018 г.

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`
- Pagination
- Speech or social media data

Wikipedia:

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».

Wikipedia:

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».

PDF:

Here it is: the *Citizen Kane* of Italian post-apocalyptic rip-offs. I know that's not really saying much, but amidst many counterparts, *2019: After the Fall of New York* stands inches above its radioactive cinematic kin. Many similar spaghetti sci-fi epics seem to be filmed exclusively at a deserted rock quarry, but *2019* is an explosion of colorful characters, varied locations and a script that finds inspiration in multiple post-nuke movies beyond *Mad Max* films. It's a familiar story—but without pausing for thought, genre-hopping director Sergio Martino knows how to keep the ridiculous action moving.

In the aftermath of war, two governments are battling for control of the United States. There's the Pan-American Confederacy, who have the fancier computers and control panels, and the Euracs, who have mean soldiers in black that ride white horses and hunt contaminated humans. Michael Sopkiw (*Blastfighter*) stars as Snake

```
Here it is: the
Citizen Kane
of Italian post-apocalyptic rip-offs.
I know that’s not really saying much, but amidst many counterparts,

the Fall of New York

stands inches above its radioactive

cinematic kin. Many sintilar spaghetti sci-fi epics seem to be filmed
exclusively at a deserted rock quarry,but 2019 is an explosion of col
ourful characters, varied locations and a script that finds inspiration
in multiple post-nuke movies beyond Mad Max films. It’s a familiar
story-but without pausing for thought, genre-hopping director
Sergio Martino knows how to keep the ridiculous action moving.
In the aftermath of war, two governments are battling for con
trol of the United States. There’s the Pan-American Confederacy,
who have the fancier computers and control panels, and the Euracs,
who have mean soldiers in black that ride white horses and hunt
contantinated humans. Michael Sopkiw (Blastjighter) stars as Snake
```

Output of `pdftotext`:

- Split words: *col—ourful*
- Newlines in sentences: *the Fall of New York …stands inches above …*

```
I know that’s not really saying much, but amidst many counterparts,
the Fall of New York
stands inches above its radioactive
cinematic kin. Many sintilar spaghetti sci-fi epics seem to be filmed
exclusively at a deserted rock quarry,but 2019 is an explosion of col
ourful characters, varied locations and a script that finds inspiration
in multiple post-nuke movies beyond Mad Max films. It’s a familiar
story-but without pausing for thought, genre-hopping director
Sergio Martino knows how to keep the ridiculous action moving.
In the aftermath of war, two governments are battling for con
trol of the United States. There’s the Pan-American Confederacy,
who have the fancier computers and control panels, and the Euracs,
who have mean soldiers in black that ride white horses and hunt
contantinated humans.
```

Output of `pdftotext`:

- Split words: *col—ourful*
- Newlines in sentences: *the Fall of New York …stands inches above …*

Два каменных будды,
Нагие, стоят у дороги;

Их ветр овевает
И хлещут дожди и метели.

Завидуй! Не знают
Они человечьей разлуки.

How many sentences ?

Task

Sentence segmentation, or:

- Sentence boundary disambiguation
- Sentence boundary detection
- Sentence tokenisation

Why?

- Most downstream tasks work on a sentence level
  - Parsing, information extraction, …
- Bad sentence segmentation hurts downstream applications

Some applications

Aside from use in downstream tasks,

- Vital for creating translation memories:
  - If the memory contains paragraphs, essentially all matches will be *fuzzy*[1]
- Also applies to parallel corpora.
  - Good segmentation and tokenisation can increase BLEU score substantially

[1] A *fuzzy* match in a translation memory is a non-complete sentence match.

What is a «sentence»?

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

But how about:

- Titles (e.g. section titles)
- Lists (bulleted or numbered)
- 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

```
1  import sys
2
3  line = sys.stdin.readline()
4  while line != '':
5          for c in '?!.':
6              line = line.replace(c, c + '\n')
7          sys.stdout.write(line)
8          line = sys.stdin.readline()
```

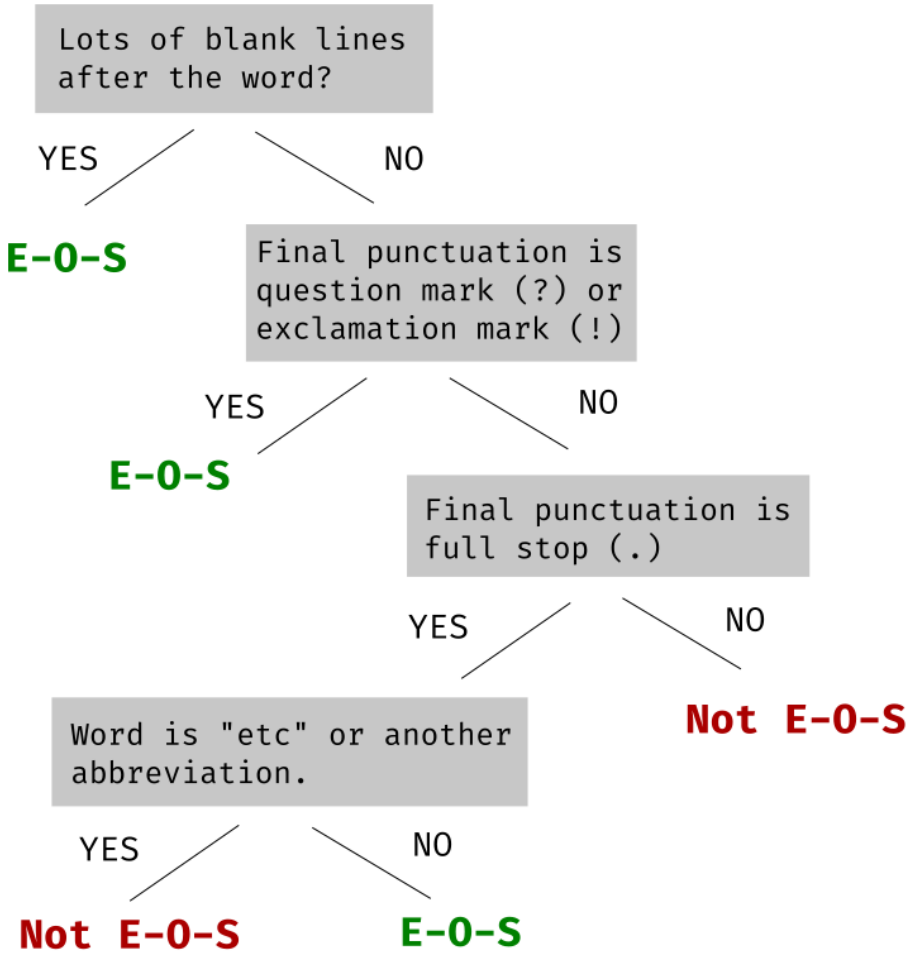
Abbreviations

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:

```
1  line = sys.stdin.readline()
2  while line != '':
3      for token in line.split('␣'):
4          if token[-1] in '!?':
5              sys.stdout.write(token + '\n')
6          elif token[-1] == '.':
7              if token in ['etc.', 'i.e.', 'e.g.']:
8                  sys.stdout.write(token + '␣')
9              else:
10                 sys.stdout.write(token + '\n')
11         else:
12             sys.stdout.write(token + '␣')
13     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

Important for coverage:

- 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

For the 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

Two examples from the journal *Computational Linguistics*:

- Mikheev, A. (1994) “Periods, Capitalized Words, etc.” *Computational Linguistics* 16(4)
  - Deals with both sentence boundary detection, capitalised word disambiguation and abbreviations.
- Kiss, T. and Strunk, J. (2006) “Unsupervised Multilingual Sentence Boundary Detection”. *Computational Linguistics* 32(4)
  - Unsupervised approach relying on detection of abbreviations.