NLP for Geography

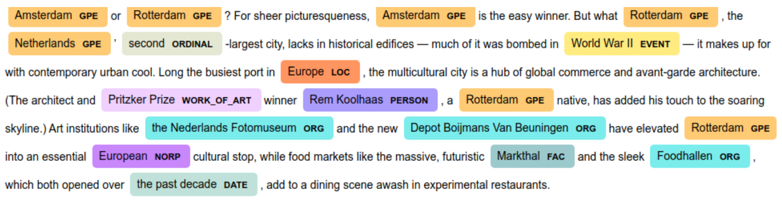
Geo-text data mining

Gosse Bouma

Information Science
Groningen University

May 2023

Overview



Amsterdam GPE or Rotterdam GPE ? For sheer picturesqueness, Amsterdam GPE is the easy winner. But what Rotterdam GPE , the Netherlands GPE ' second ORDINAL -largest city, lacks in historical edifices — much of it was bombed in World War II EVENT — it makes up for with contemporary urban cool. Long the busiest port in Europe LOC , the multicultural city is a hub of global commerce and avant-garde architecture. (The architect and Pritzker Prize WORK_OF_ART winner Rem Koolhaas PERSON , a Rotterdam GPE native, has added his touch to the soaring skyline.) Art institutions like the Nederlands Fotomuseum ORG and the new Depot Boijmans Van Beuningen ORG have elevated Rotterdam GPE into an essential European NORP cultural stop, while food markets like the massive, futuristic Markthal FAC and the sleek Foodhallen ORG , which both opened over the past decade DATE , add to a dining scene awash in experimental restaurants.

- Introduction to NLP
- Named Entity Detection and Classification
- Named Entity Linking and Geocoding
- Information Extraction with linguistic patterns
- Using Large Language Models (ChatGPT)

Natural Language Processing

Natural Language Processing

- **Tools** and **applications** for **analyzing** (and generating) **text** and speech
- Very detailed:
  - Models for recognizing the various meanings of the word Python (*word sense disambiguation*)
- Very general:
  - Large Language Models (ChatGPT) as generic building blocks that can be fine-tuned or prompted for specific tasks

Challenges

- *Multilinguality* (English, Dutch, Spanish, Japanese, Hebrew, etc.)
- Register *Variation* (newspaper vs. fiction vs. social media)
- *Ambiguity*: Lexcial (Python), structural (syntactic), named entities (Groningen)

Natural Language Processing

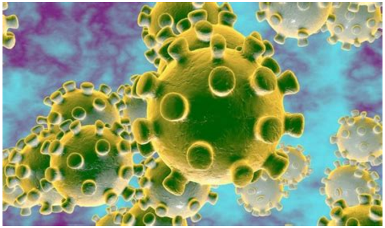
Natural Language Processing

- **Tools** and **applications** for **analyzing** (and generating) **text** and speech
- Very detailed:
  - Models for recognizing the various meanings of the word Python (*word sense disambiguation*)
- Very general:
  - Large Language Models (ChatGPT) as generic building blocks that can be fine-tuned or prompted for specific tasks

Challenges

- *Multilinguality* (English, Dutch, Spanish, Japanese, Hebrew, etc.)
- Register *Variation* (newspaper vs. fiction vs. social media)
- *Ambiguity*: Lexcial (Python), structural (syntactic), named entities (Groningen)

Word Senses: Corona



Word Senses: Python

Word Senses and Embeddings

Word embeddings reflect the *most frequent sense* of a word

```
fasttext nn cc.en.300.bin
```

Python		**Pythons**	
python	0.749	Python	0.641
Pythonic	0.726	pythons	0.619
Python.	0.713	Constrictors	0.565
Perl	0.707	Snakes	0.524
Python-like	0.706	python	0.519
Python3	0.683	Rattlesnakes	0.477
Python-based	0.664	Pythonesque	0.474
Python2	0.659	Monty	0.465
Numpy	0.653	Pythonidae	0.464
Pythons	0.641	Iguanas	0.464

Word Senses and Embeddings

```
fasttext nn cc.en.300.bin
```

Word embeddings reflect the *most frequent sense* of a word

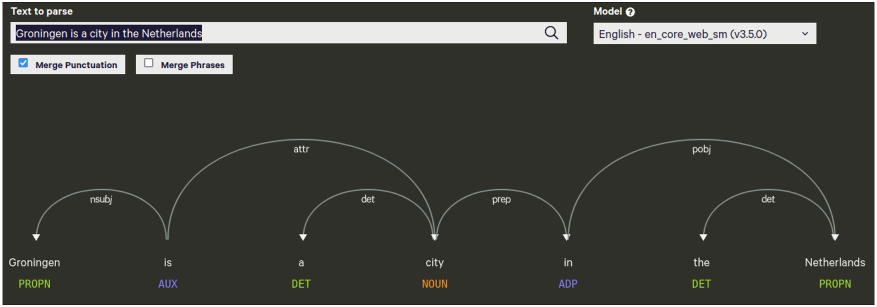
corona		**Corona**	
coronas	0.700	Coronita	0.607
coronae	0.590	Tecate	0.570
Corona	0.531	Hermosa	0.567
aurora	0.493	Coronas	0.567
halo	0.490	Redondo	0.563
chromosphere	0.489	Cerveza	0.557
nanoflares	0.482	Coronado	0.536
filamentary	0.468	corona	0.531
aureole	0.464	Estrella	0.527
halo-like	0.458	Laguna	0.520

Syntactic Ambiguity

One morning I shot an elephant in my pajamas.
How he got in my pajamas I'll never know." Groucho Marx

Syntactic Ambiguity

One morning I shot an elephant in my pajamas.
How he got in my pajamas I'll never know." Groucho Marx

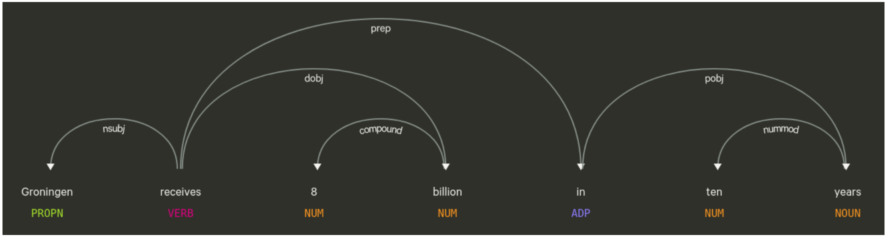


Text to parse: Groningen is a city in the Netherlands

Model: English - en_core_web_sm (v3.5.0)

[✓ Merge Punctuation] Merge Phrases

a city in the Netherlands



receives [8 billion] in ten years

NLP Pipeline

To analyze the linguistic structure of a text involves one or more of the following steps:

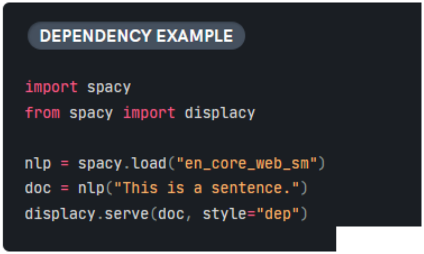
- **Preprocessing**: Sentence splitting and tokenization
- **Lexical Analysis**: Lemmatization and Part-of-Speech tagging
- **Syntactc Analysis**: Phrase Structure analysis or Dependency Analysis
- **Semantic Interpretation**: Logical analysis, coreference resolution, word sense disambiguation
- **Discourse Interpretation**: Rhetorical and logical relations between sentences

Natural Language Processing Toolkit

spacy.io

- Python toolkit for analyzing natural language
- Sentence Splitting: segment a text into sentences
- Tokenization: segment a string into a list of tokens
- Lemmatization: label tokens with their lemma (words $\rightarrow$ word, were $\rightarrow$ be)
- Part-of-Speech: label tokens with Part-of-Speech (VERB, NOUN, DET, PROPN, etc.)
- Syntax: syntactic dependency relations between words

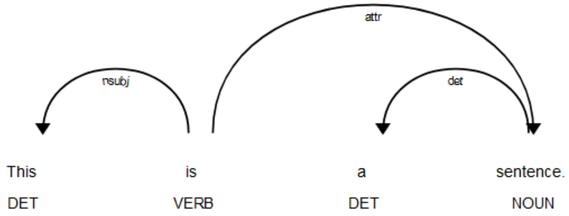
Spacy



DEPENDENCY EXAMPLE

```
import spacy
from spacy import displacy

nlp = spacy.load("en_core_web_sm")
doc = nlp("This is a sentence.")
displacy.serve(doc, style="dep")
```



demo : `https://explosion.ai/demos/displacy`

Spacy introduction

```
import spacy
# this loads the model for analysing English text
nlp = spacy.load("en_core_web_sm")

question = nlp('What is the eye color of a siamese cat?')
for word in question :
    print(word.text, word.lemma_, word.pos_)
```

```
What PRON what
is AUX be
the DET the
eye NOUN eye
color NOUN color
of ADP of
a DET a
siamese ADJ siamese
cat NOUN cat
? PUNCT ?
```

Spacy introduction

```
import spacy
# this loads the model for analysing English text
nlp = spacy.load("en_core_web_sm")

question = nlp('What is the eye color of a siamese cat?')
for word in question :
    print(word.text, word.lemma_, word.pos_)
```

```
What PRON what
is AUX be
the DET the
eye NOUN eye
color NOUN color
of ADP of
a DET a
siamese ADJ siamese
cat NOUN cat
? PUNCT ?
```

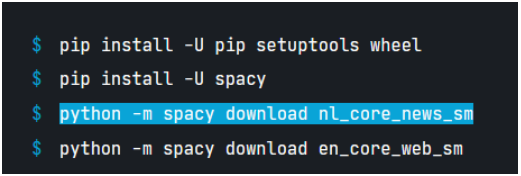
Syntactic Pattern Matching

Finding Phrases

- Once a text is analysed, we can search for phrases that match a syntactic patterns:
  - Adjective-noun combinations (*'largest city, new model, artificial intelligence'*)
  - Subject-verb-object combinations (*'google-buy-company, koolhaas-win-prize'*)

More Spacy

- Installation: `https://spacy.io/usage`



```
$ pip install -U pip setuptools wheel
$ pip install -U spacy
$ python -m spacy download nl_core_news_sm
$ python -m spacy download en_core_web_sm
```

- Tutorial: `https://spacy.io/usage/spacy-101`