Natural Language Processing

Aims Senegal 2021

Course Logistics

- Instructor: Dr. Elvis Ndah
- Affiliation: University of Ghent and Anju Software (Belgium)
- Contact (email): elvis.ndah@gmail.com

- Others
  - Slides will be distributed before each lecture
  - Full course materials on github

Text books

1. Speech and Language Processing 3rd ed, Jurafsky and Martin. https://web.stanford.edu/~jurafsky/slp3/
2. Natural Language Processing, Jacob Eisenstein. https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
3. Foundations of Statistical Natural Language Processing, Chris Manning and Hinrich Schultze. MIT Press. Cambridge, MA: May 1999.

How to get the best out of the course

1. Be proactive
   - Read the required ducoments
   - Start early with assignments

2. Communicate
   - Participate in discussions
   - Sttend classes

3. There is no stupid question

The focus of this course

1. Understanding of the modern techniques for NLP
   - Start with the basics
   - Methods in modern NLP: RNN, encoder-decoder, transformers
2. A High level understanding of human languages and the difficulties in understanding and reproducing them.
3. Understanding of the ability to build systems for major problems in NLP
   1. Word meaning
   2. Language models
   3. Machine translation
   4. Question answering

Course grading

- Final Course Project
  - Details: TBD

Prerequisites

Recommended knowledge

- Python programming
- Probability and statistics
- Deep learning (Machine learning)

Lecture 1:
Introduction to NLP

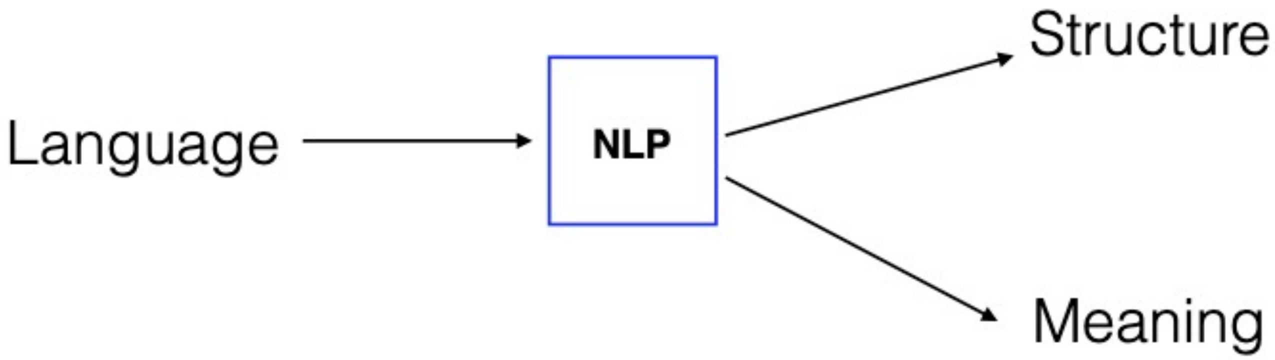
Overview

1. What is NLP?
2. Why work on NLP?
3. Why is NLP hard?
4. NLP application (task)
5. NLP pipelines

What is NLP?

Goal:

- Develop methods for processing, analyzing and understanding the structure and meaning of all natural (human) language.



Language

NLP

Structure

Meaning

- It concerns with the interaction between natural languages and computing devices.

Why work on NLP?

Main purpose:

- Build systems that help humans communicate and interact with each ather and devices.
- They systems are useful to help
  - Automatically manage and summarize text
  - Machine Translation: communicate without language barrier
  - Model and analyse properties of language
  - Speech recognition

What’s special about natural language?

Lingustic analysis

- Phonology: sounds that make up language
- Morphology: internal struxcture of words
- Syntax: structure of phrases, how words modify one another
- Semantics: meaning of language in the world
- Discourse: relations between cluases and sentences

Why is NLP hard?

- Language is encoded via continuous signals (Sounds, Gestures (sign language, Image)

- Languages are ambiguous (ambiguity)
  - Syntactic (grammatical)
  - Semantic
  - Lexical

- Languages are complex (variability)
- Understanding requires vast knowledge
- Human input is scarce

Why is NLP hard – Syntactic ambiguity

What is grammar?

This is one of the most fundamental question to answer before working on high level NLP task such as language modelling

- Grammar formalisms
  - A precise way to define and describe the structure of sentences

- Specific grammars
  - Implementations (in a particular formalism) for a particular language (English, French, Chinese,…)

Why is NLP hard – Syntactic ambiguity

Syntactic or grammertical ambiguity: *two or more possible meanings within a single sentence.*

> *Finally, a computer that understands you like your mother”*
>
> *(Ad , 1985)*

1. The computer understands you as well as your mother understands you.
2. The computer understands that you like your mother.
3. The computer understands you as well as it understands your mother.

Why is NLP hard – Semantic ambiguity

Semantic ambiguity: *occurs when a word, phrase or sentence, taken out of context, has more than one interpretation.*

“We saw her duck”

- The word *“her duck”* can can refer either to
  - the person’s bird - the noun *“duck”* modified by the possessive pronoun “*her”*
  - a motion she made - the verb *“duck”*, subject of the objective pronoun *“her”*, object of the verb *“saw”*

Why is NLP hard – Lexical ambiguity

Lexical ambiguity: *two or more possible meanings within a single word*

Finally, a computer that understands your lie cured mother”

- The word *lie* can have multiple meanings in the and will not change the context of the sentence.
- The ambiguity is on what actually cured mother
  - *lie:* an intentionally false statement
  - *lie:* spice or home made remedy

Why is NLP hard – Variability

- There are many different ways to express the same meaning in language.
    - *Man united ends up 6 points*
    - *Man united climbs 6 points*
    - *Man united gains 6 points*
    - *All top teams surged*

- Key computational challenge in NLP is to compute similarity of the above phrases.

Why is NLP hard? – Solution

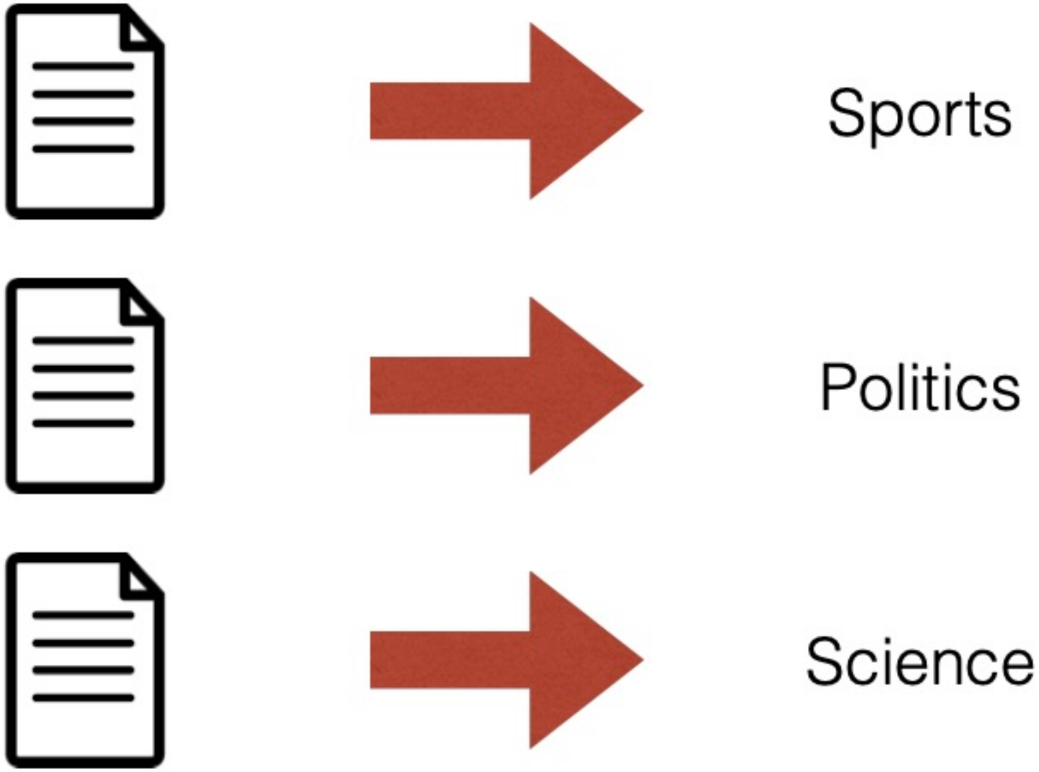
Solution:

- Incorporate linguistic knowledge
- Learn from human input, when available
- Automatically learn structure

NLP applications

- Text categorization
- Information Extraction
- Searching
- Question Answering
- Virtual assistants
- Machine translation
- Summarization
- Reading comprehension

NLP applications – text categorization



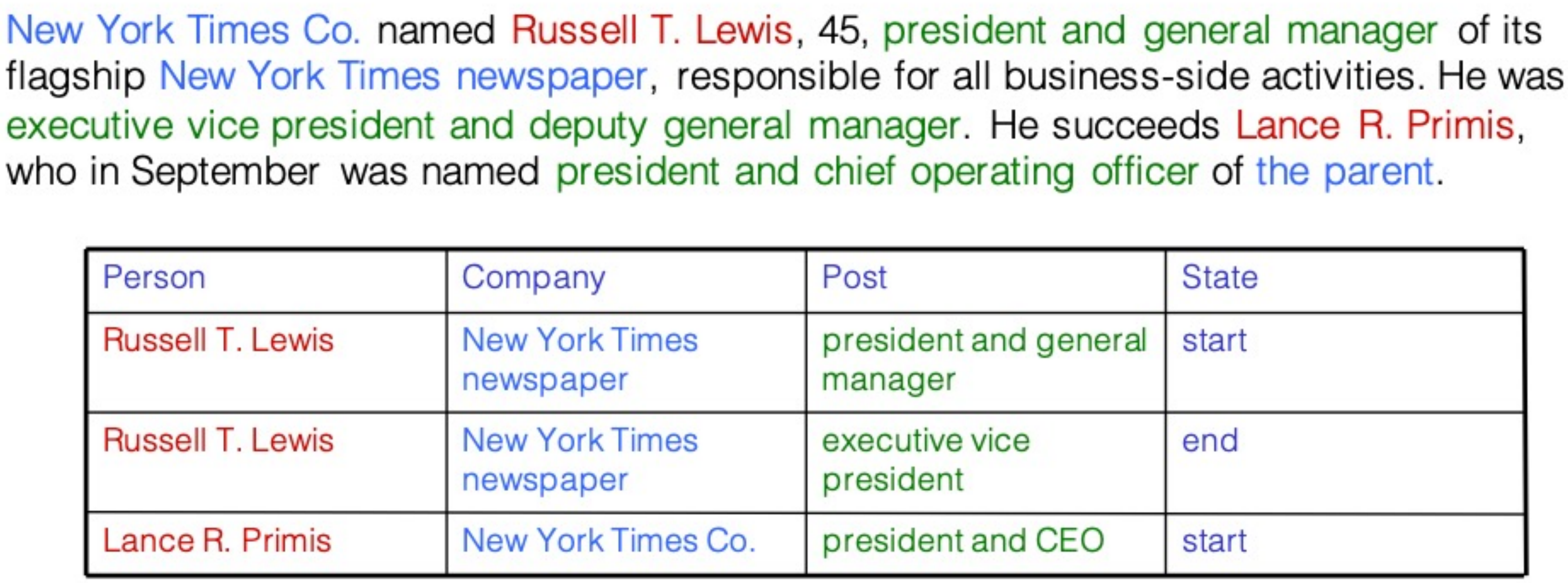
offensive

Not offensive

neutral

NLP applications – information extraction

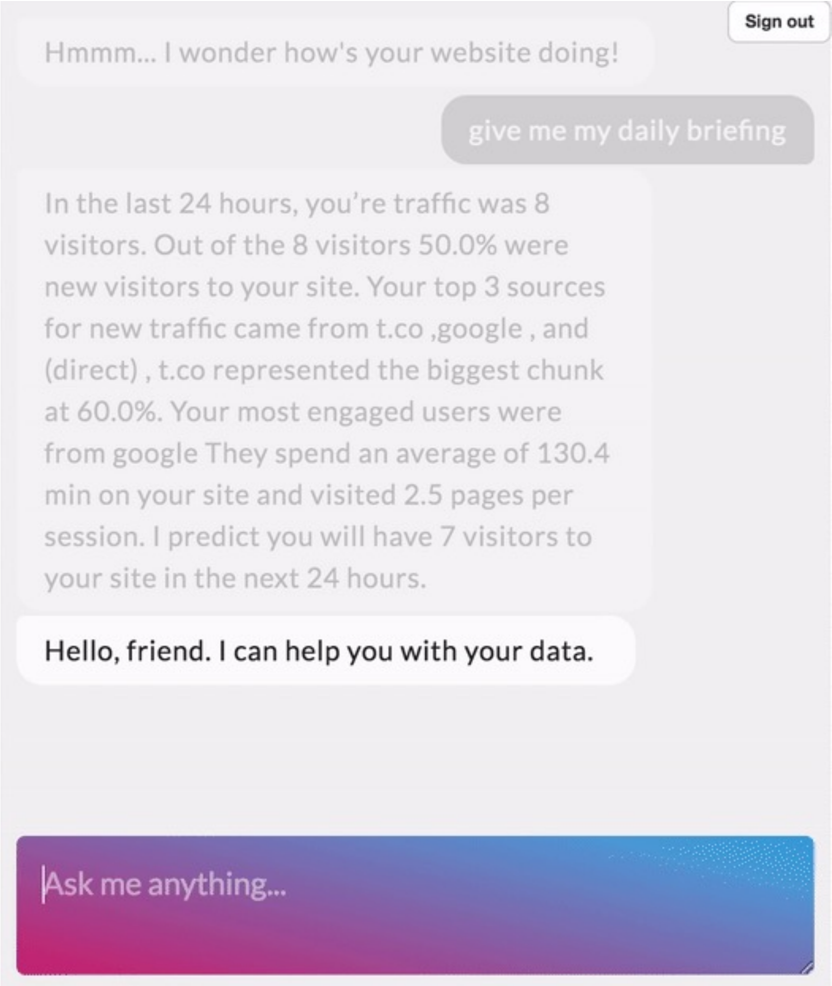
The task of **Information Extraction (IE)** involves extracting meaningful information from unstructured text data and presenting it in a structured format.



New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

NLP applications – Question answering



Hmmm... I wonder how's your website doing!

give me my daily briefing

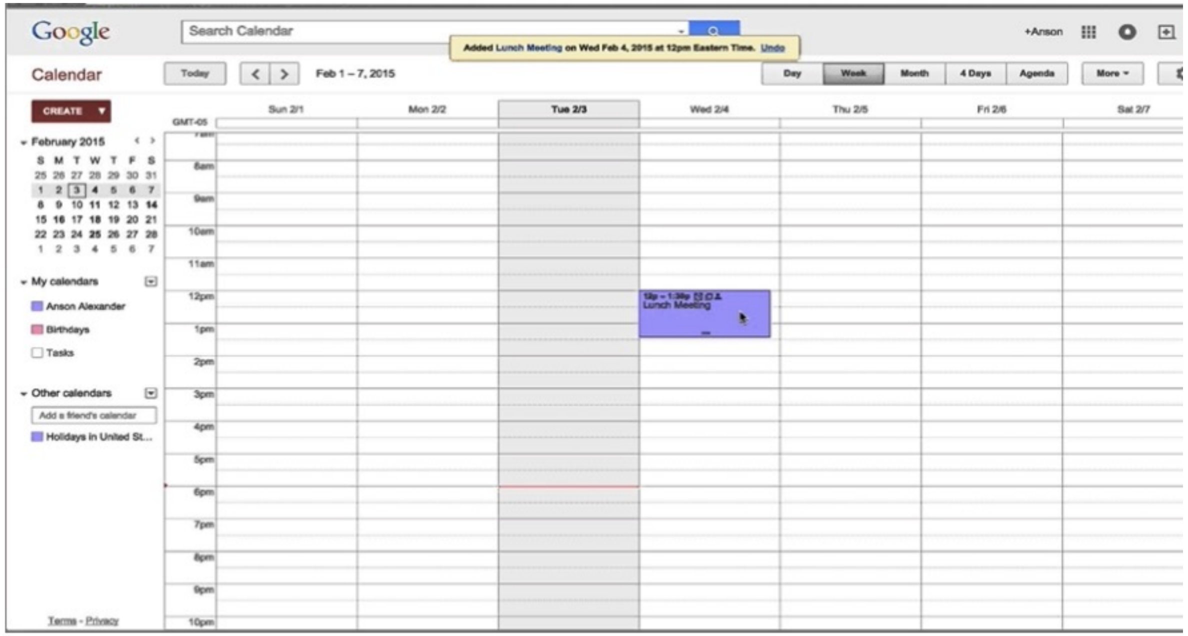
In the last 24 hours, you're traffic was 8 visitors. Out of the 8 visitors 50.0% were new visitors to your site. Your top 3 sources for new traffic came from t.co ,google , and (direct) , t.co represented the biggest chunk at 60.0%. Your most engaged users were from google They spend an average of 130.4 min on your site and visited 2.5 pages per session. I predict you will have 7 visitors to your site in the next 24 hours.

Hello, friend. I can help you with your data.

https://towardsdatascience.com/nlp-building-a-question-answering-model-ed0529a68c54

NLP applications – virtual assistants

Move all my Wednesday meetings in April with John to 5pm



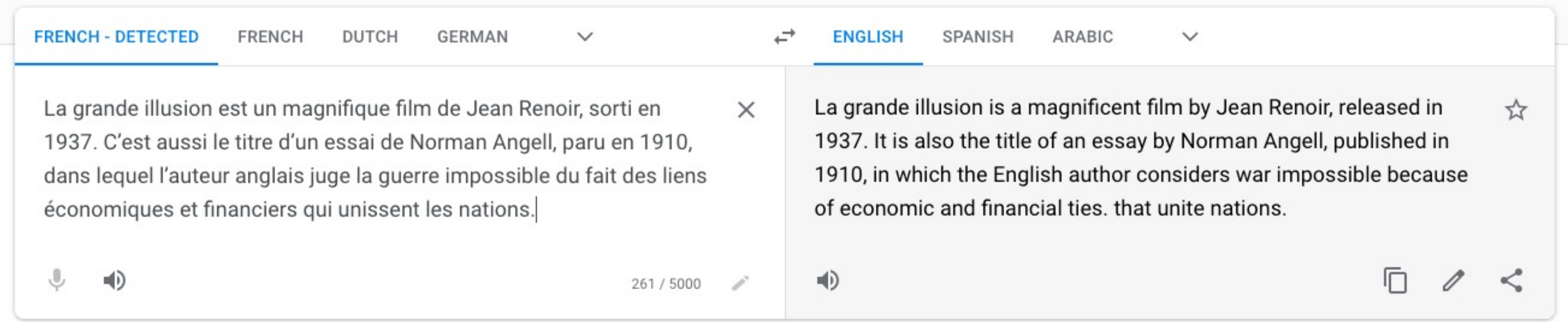
Added Lunch Meeting on Wed Feb 4, 2015 at 12pm Eastern Time. Undo

https://www.rogerpeele.com/how-to-choose-the-right-virtual-assistant-for-your-company-in-sydney/

NLP applications – Machine translation

Automatically translate from one natural language to another

- Rule-based Machine Translation (RBMT): 1970s-1990s
- Statistical Machine Translation (SMT): 1990s-2010s
- **Neural Machine Translation (NMT): 2014-**



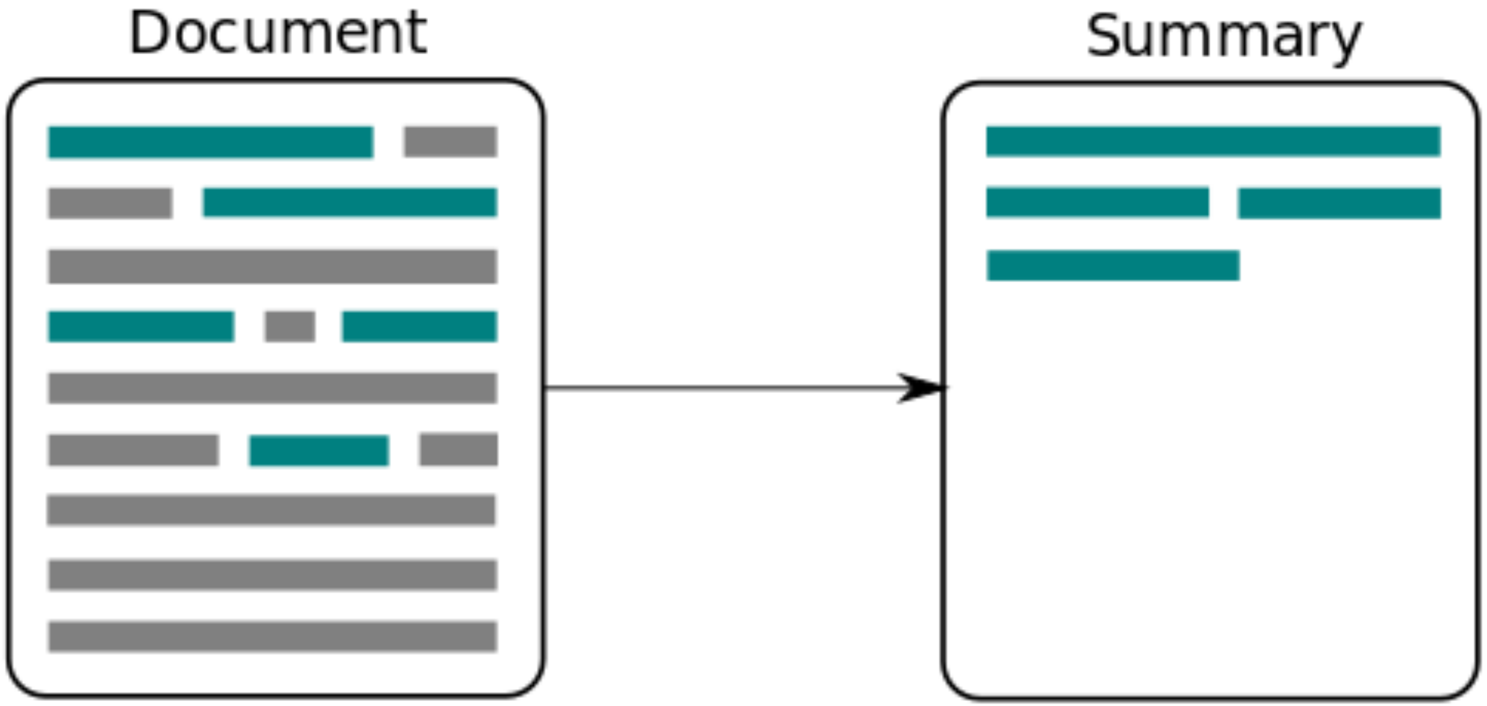
La grande illusion est un magnifique film de Jean Renoir, sorti en 1937. C'est aussi le titre d'un essai de Norman Angell, paru en 1910, dans lequel l'auteur anglais juge la guerre impossible du fait des liens économiques et financiers qui unissent les nations.

La grande illusion is a magnificent film by Jean Renoir, released in 1937. It is also the title of an essay by Norman Angell, published in 1910, in which the English author considers war impossible because of economic and financial ties. that unite nations.

- How do the words map?
- How does the grammar map?
- Does the output language read fluently?

NLP applications – Summarization

Task of automatically shortening long pieces of text or document.



https://hackernoon.com/summarization-with-wine-reviews-using-spacy-b49f18399577

NLP applications – Reading comprehension

More than a decade ago, Carl Lewis stood on the threshold of what was to become the greatest athletics career in history. He had just broken two of the legendary Jesse Owens' college records, but never believed he would become a corporate icon, the focus of hundreds of millions of dollars in advertising. His sport was still nominally amateur.

Eighteen Olympic and World Championship gold medals and 21 world records later, Lewis has become the richest man in the history of track and field – a multi-millionaire.

- Who is Carl Lewis?
- Did Carl Lewis break any world records? (and how do you know that?)
- Is Carl Lewis wealthy? What about Jesse Owens?

The NLP Pipeline

Each step of the pipeline generate a representation for the preceding

- **Tokenizer and segmentation:** to identify words and sentences boundaries
    - Text normalization and vocabulary creation
- **Morphological analyser** (POS tagger)**:** identify the structure of words
- **Word sense disambiguation:** identify the meaning of words
- **Syntactic/semantic parser:** obtain the structure and meaning of sentences
- **discourse model:** keep track of the various entities and events mentioned