NATURAL LANGUAGE PROCESSING

A PRACTICAL OVERVIEW OF THE POSSIBILITIES FOR INFORMATION RETRIEVAL

Christophe SERVAN, PhD – Chief Scientist

Qwant

CHRISTOPHE SERVAN, PHD

WEB SEARCH ENGINE

Web Search engines:

- Google
- Baïdu
- Yahoo
- Yandex
- Bing
- Naver
- Seznam
- Qwant

Meta-Web Search Engine:

- Ecosia
- DuckDuckGo
- …

QWANT: THE EUROPEAN SEARCH ENGINE THAT RESPECTS YOUR PRIVACY

No third-party cookie

Hashed/salted IP address

No search history

Own server farm in EU, algorithms & index

HTTPS

Host of additional services

QWANT IN NUMBERS

- Launched in 2013 (V4 in summer 2018)
- 150 employees with rapid growth
- Located in France (Paris, Rouen, Nice, Ajaccio, Epinal), Germany and Italy
- Investors: Axel Springer, CDC, EIB
- Sole European search engine with its own index and algorithms
- More than 80 million connections per month
- In the Top 50 sites in Europe (Source: similarweb.com)

Qwant

ACTUAL STATE OF NLP IN IR

POPULAR ENGINES FOR IR

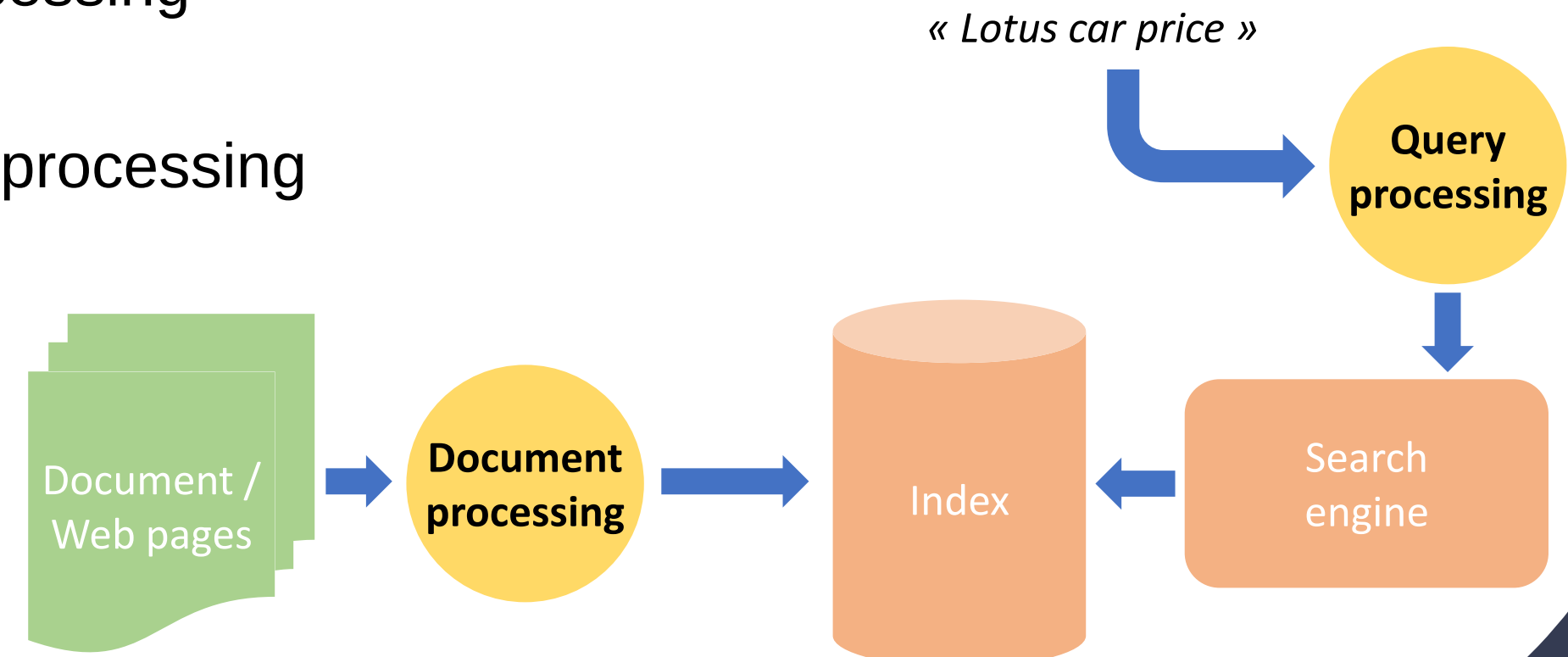
- LEMUR (INDRI)
- Lucene
- Nutch
- SolR
- Compass
- Elasticsearch

… and many other…

NATURAL LANGUAGE PROCESSING (NLP) IN IR

Classical view

- Query processing
- Document processing



NLP FOR IR: PRACTICAL STATE OF THE ART

- Query processing
- Features
- Document processing

NLP FOR IR: PRACTICAL STATE OF THE ART

- Query processing
  - Stemming
  - Stop words
  - Suggestions
  - Entailment
- Document processing
- Features

NLP FOR IR: PRACTICAL STATE OF THE ART

- Query processing
  - Stemming
  - Stop words
  - Suggestions
  - Entailment
- Document processing
  - Stemming
  - Stop words
- Features

NLP FOR IR: PRACTICAL STATE OF THE ART

- Query processing
  - Stemming
  - Stop words
  - Suggestions
  - Entailment
- Document processing
  - Stemming
  - Stop words
- Features
  - TF-IDF
  - BM25

NLP FOR IR: PRACTICAL STATE OF THE ART

- Query processing
  - Stemming
  - Stop words
  - Suggestions
  - Entailment
- Document processing
  - Stemming
  - Stop words

- Features
  - TF-IDF
  - BM25

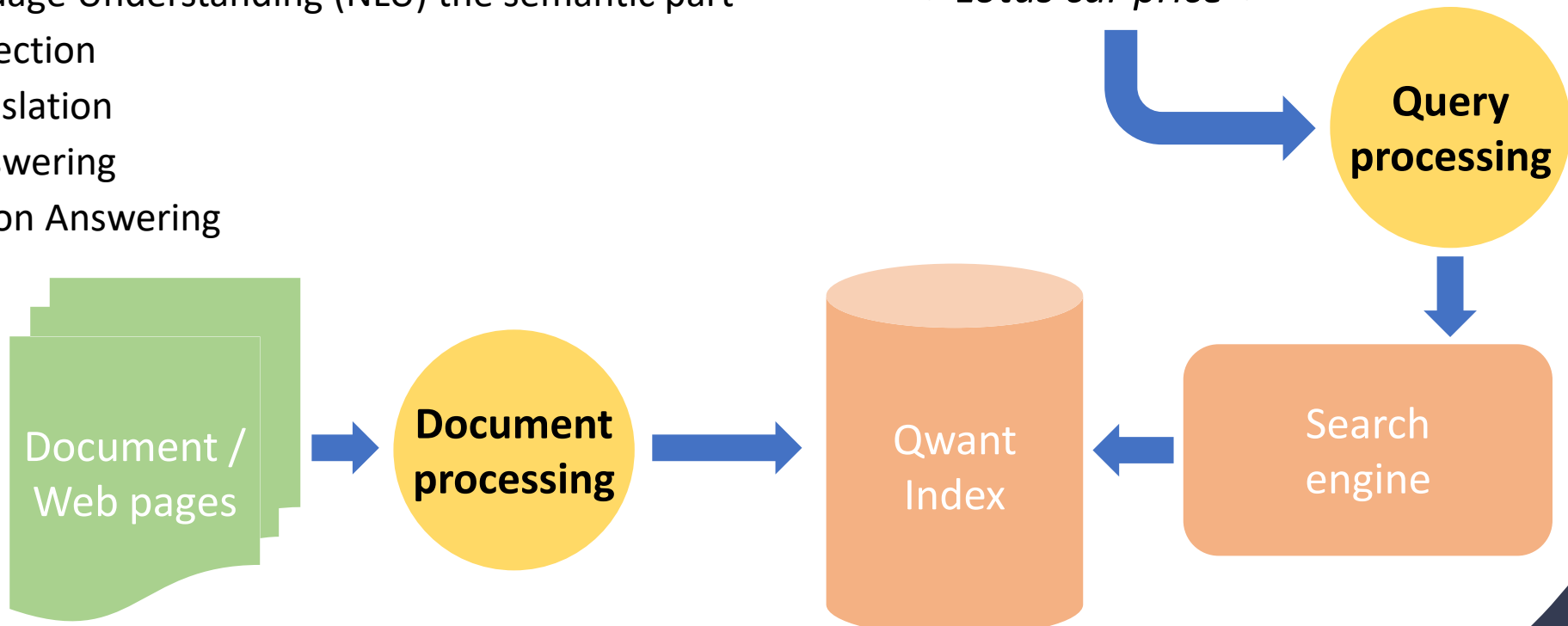
⇒ That's All!

NLP & IR @ QWANT

NATURAL LANGUAGE PROCESSING (NLP)

- **NLP** is crucial for the Qwant search engine

- Integration inside the search engine through **document and query** processing :
  - Name Entity Recognition (NER) like dates, locations, names, etc.
  - Natural Language Understanding (NLU) the semantic part
  - Intention detection
  - Machine Translation
  - Question/Answering
  - Visual Question Answering



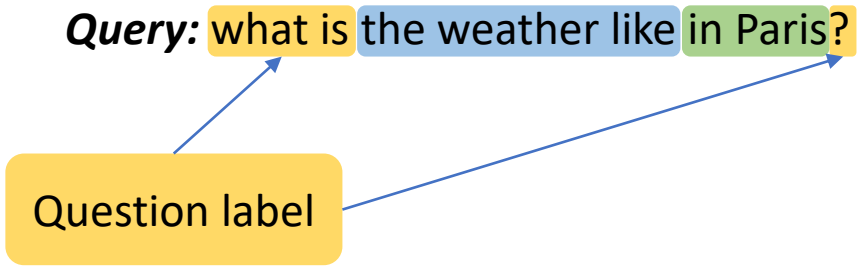
INFORMATION EXTRACTION / NER

Query: what is the weather like in Paris?

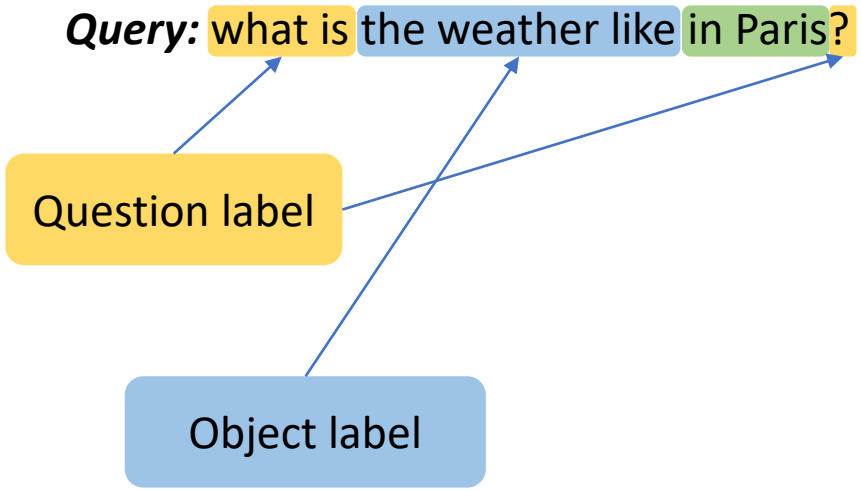
INFORMATION EXTRACTION / NER

Query: what is the weather like in Paris?

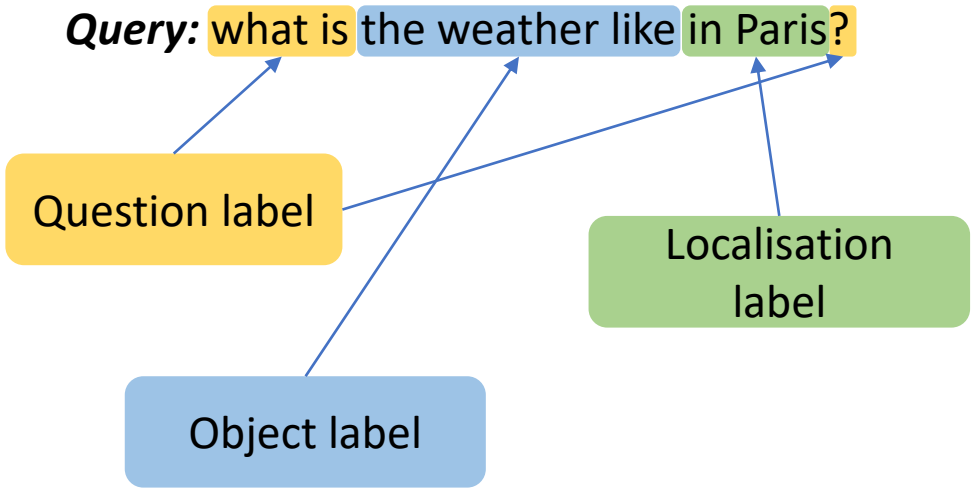
INFORMATION EXTRACTION / NER



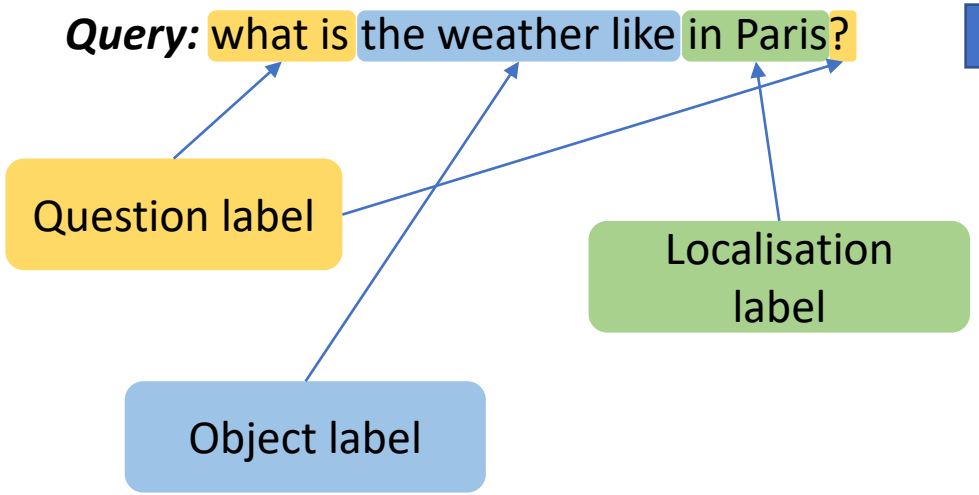
INFORMATION EXTRACTION / NER

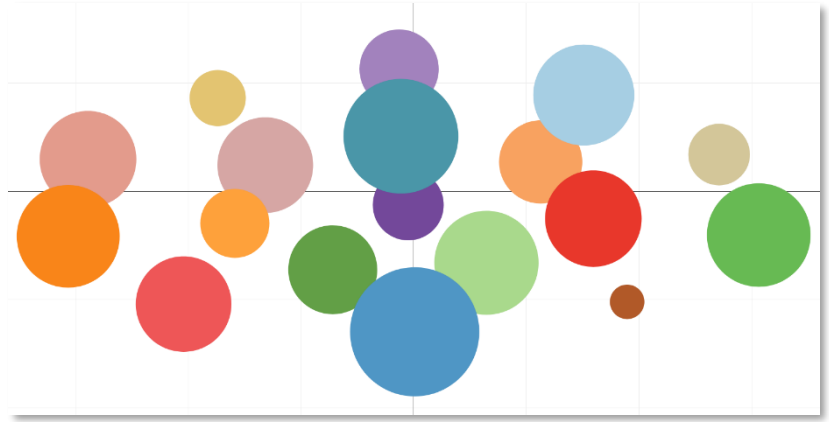


INFORMATION EXTRACTION / NER

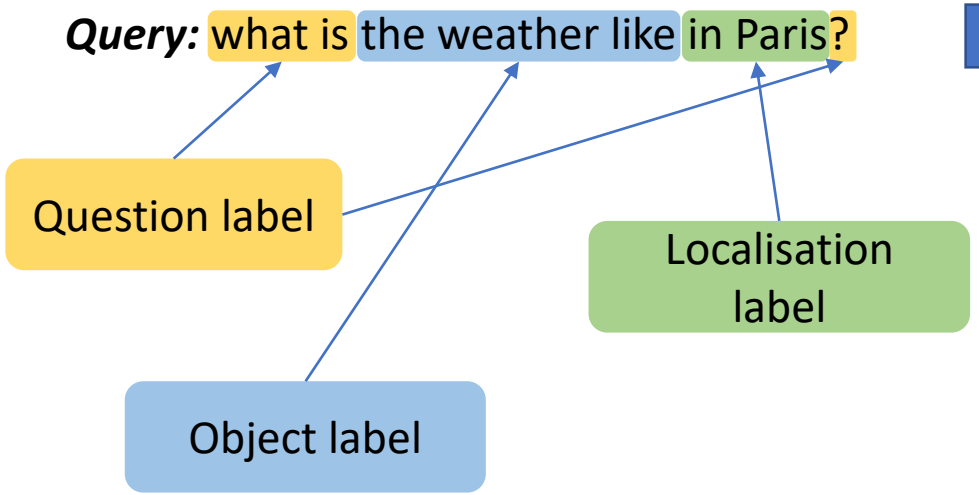


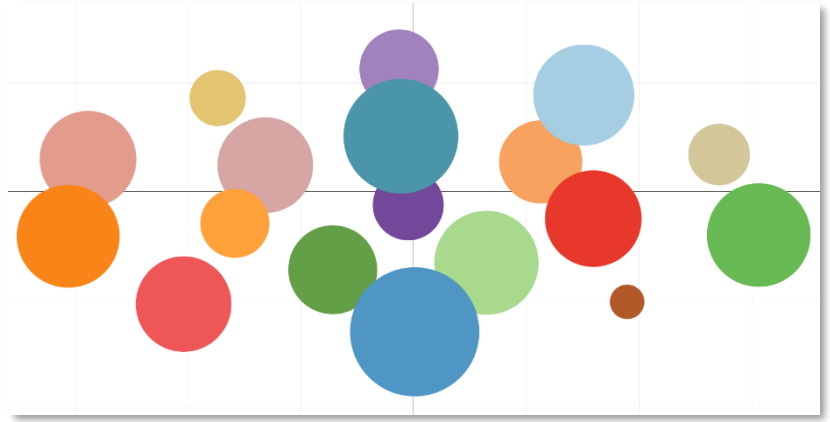
INFORMATION EXTRACTION / NER





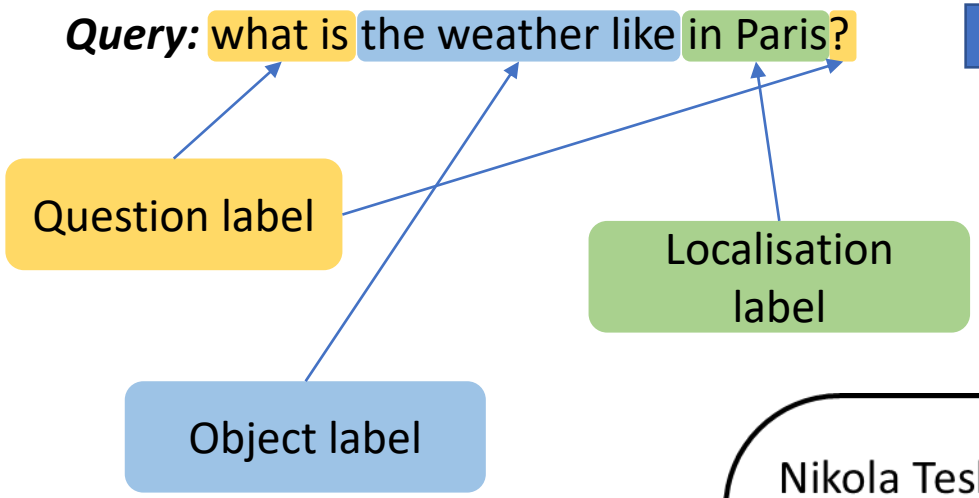
INFORMATION EXTRACTION / NER

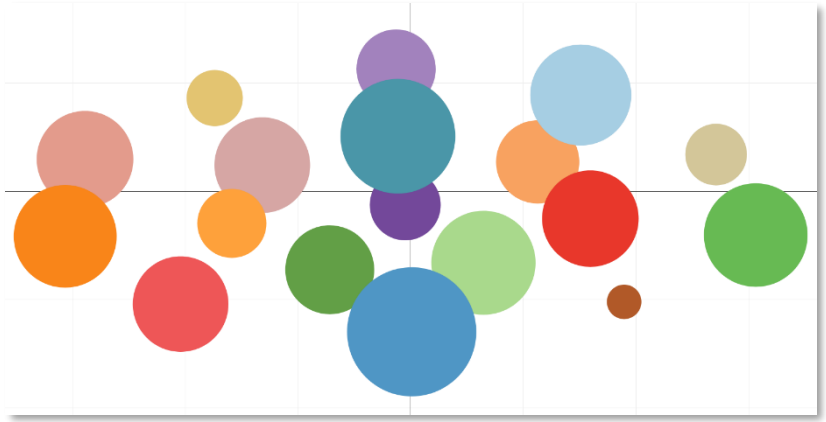




Document:

INFORMATION EXTRACTION / NER





Document:

Nikola Tesla was born an ethnic Serb in the village Smiljan, Lika county, in the Austrian Empire (present day Croatia), on 10 July 1856. His father, Milutin Tesla (1819–1879), was an Eastern Orthodox priest.
Tesla's mother, Đuka Tesla (née Mandić; 1822–1892), whose father was also an Orthodox priest, had a talent for making home craft tools and mechanical appliances and the ability to memorize Serbian epic poems.
Đuka had never received a formal education.
Tesla credited his eidetic memory and creative abilities to his mother's genetics and influence. Tesla's progenitors were from western Serbia, near Montenegro

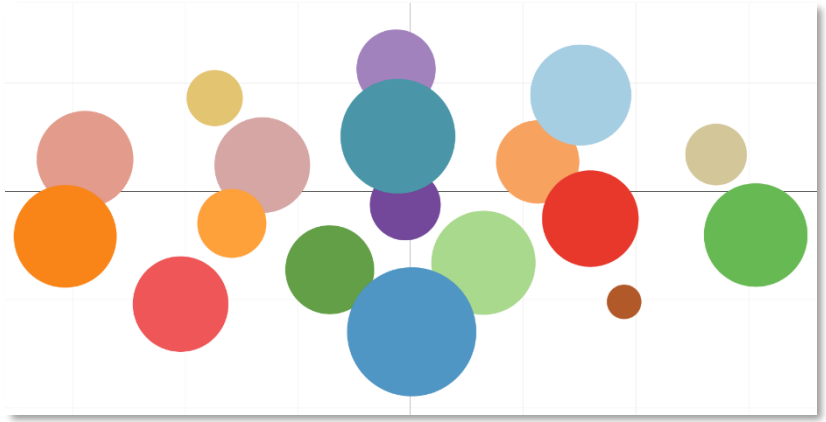
INFORMATION EXTRACTION / NER

Query: what is the weather like in Paris?

Question label

Localisation label

Object label



Document:

Nikola Tesla was born an ethnic Serb in the village Smiljan, Lika county, in the Austrian Empire (present day Croatia), on 10 July 1856. His father, Milutin Tesla (1819–1879), was an Eastern Orthodox priest.
Tesla's mother, Đuka Tesla (née Mandić; 1822–1892), whose father was also an Orthodox priest, had a talent for making home craft tools and mechanical appliances and the ability to memorize Serbian epic poems.
Đuka had never received a formal education.
Tesla credited his eidetic memory and creative abilities to his mother's genetics and influence. Tesla's progenitors were from western Serbia, near Montenegro

INFORMATION EXTRACTION / NER

- Use of OpenNMT toolkit
  $\Rightarrow$ MT approach

- Model inspired from [Ma and Hovy, 2016]

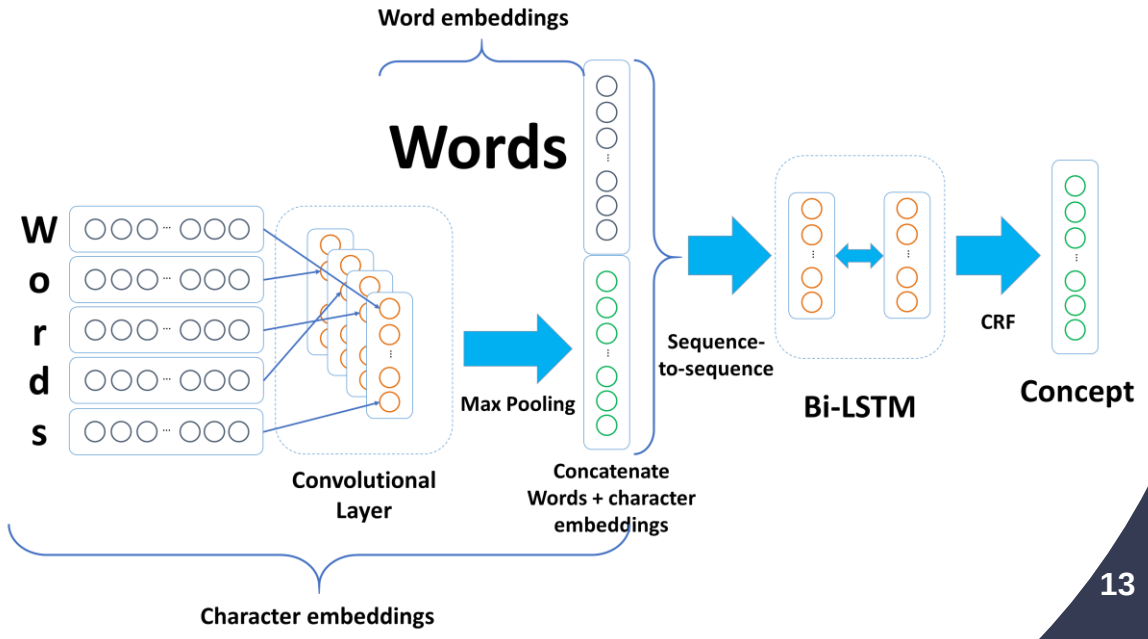
- Best performances on SLU SoA tasks (ATIS, MEDIA)

INFORMATION EXTRACTION / NER

- Use of OpenNMT toolkit
  ⇒ MT approach

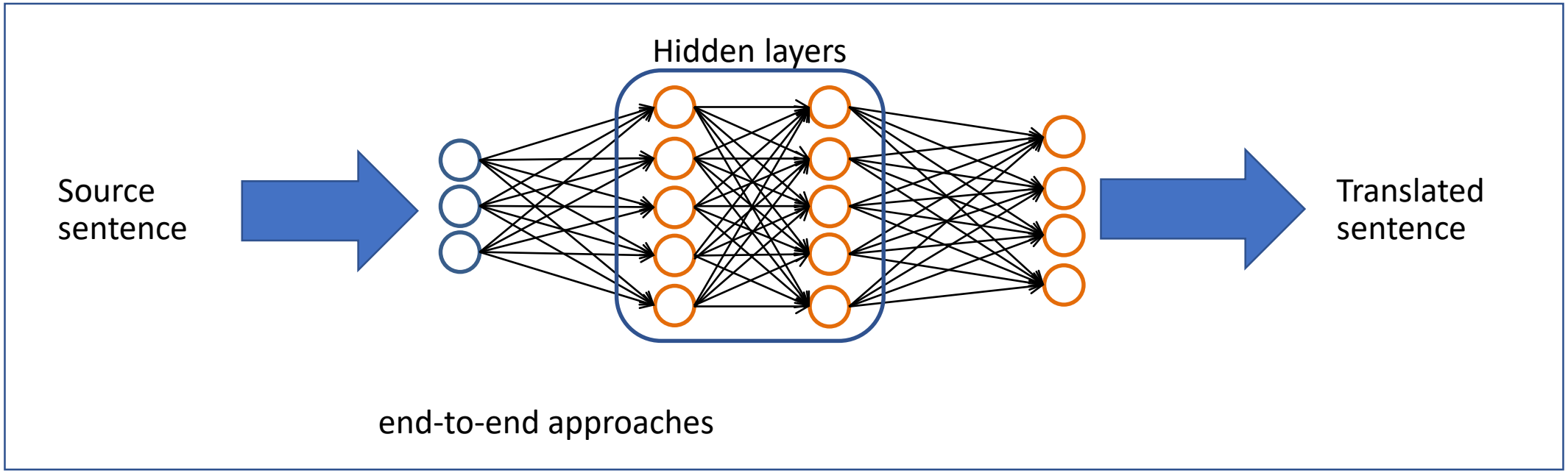
- Model inspired from [Ma and Hovy, 2016]

- Best performances on SLU SoA tasks (ATIS, MEDIA)



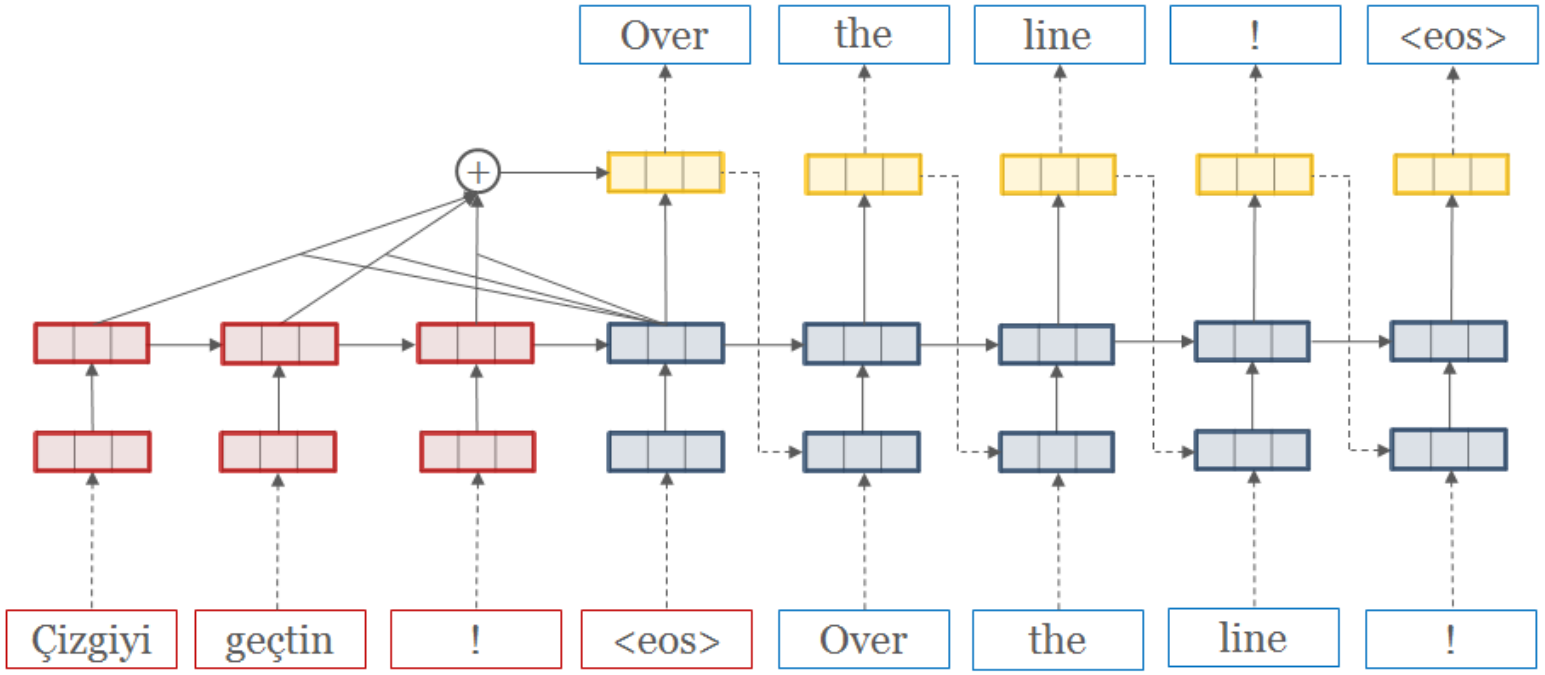
NEURAL MACHINE TRANSLATION

- Replace the TM and LM in the translation process [Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Jean et al., 2015]



NEURAL MACHINE TRANSLATION

- Sequence-to-Sequence Approach
  - With attention models
- Transformer models (a.k.a full attention model)

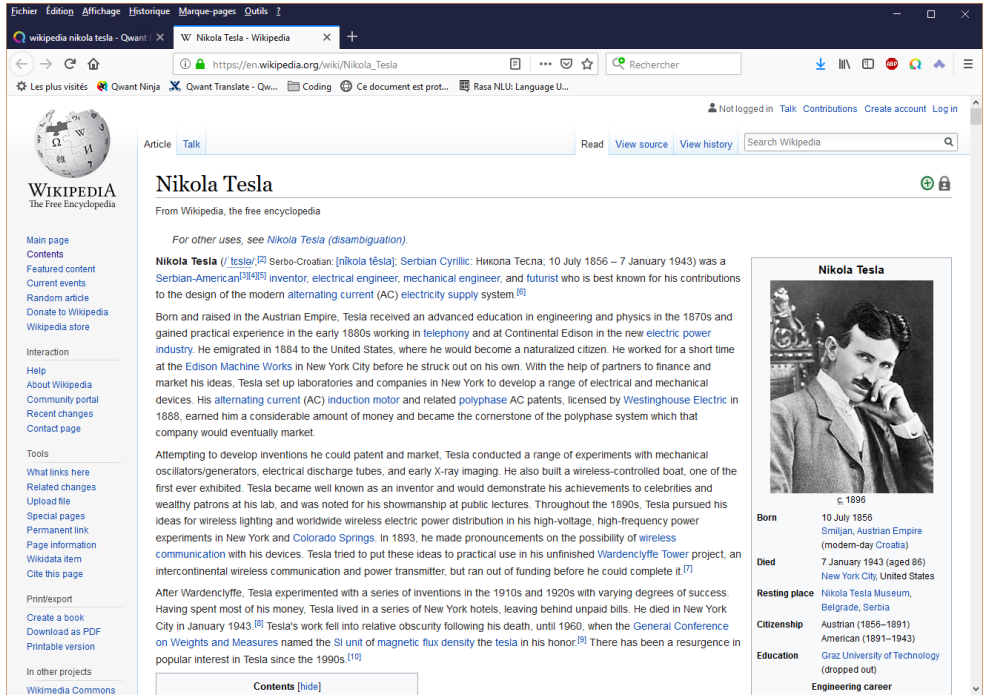


QUERYING IN SEVERAL LANGUAGES

Query: Nikola Tesla wikipedia

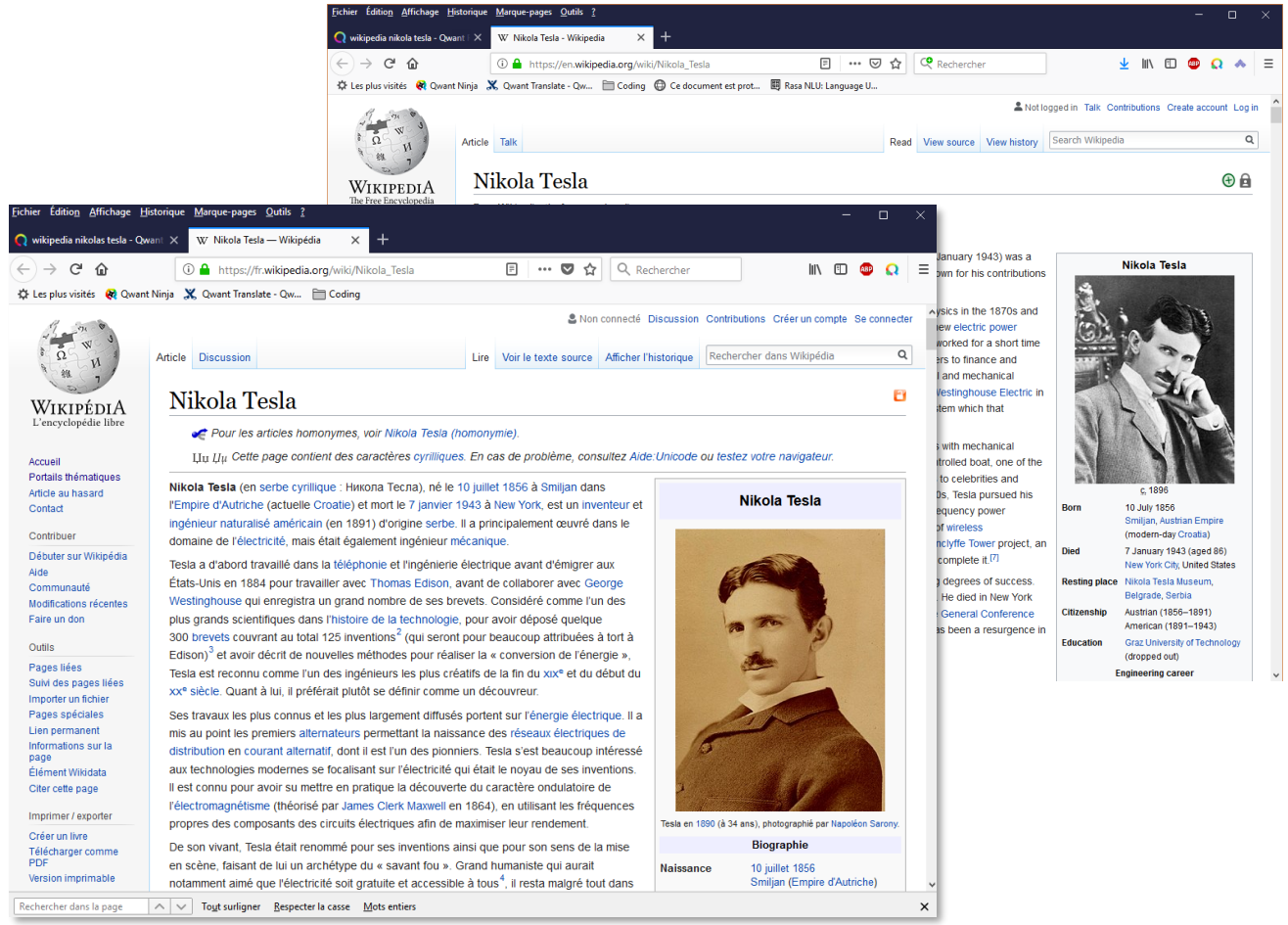
QUERYING IN SEVERAL LANGUAGES

Query: Nikola Tesla wikipedia



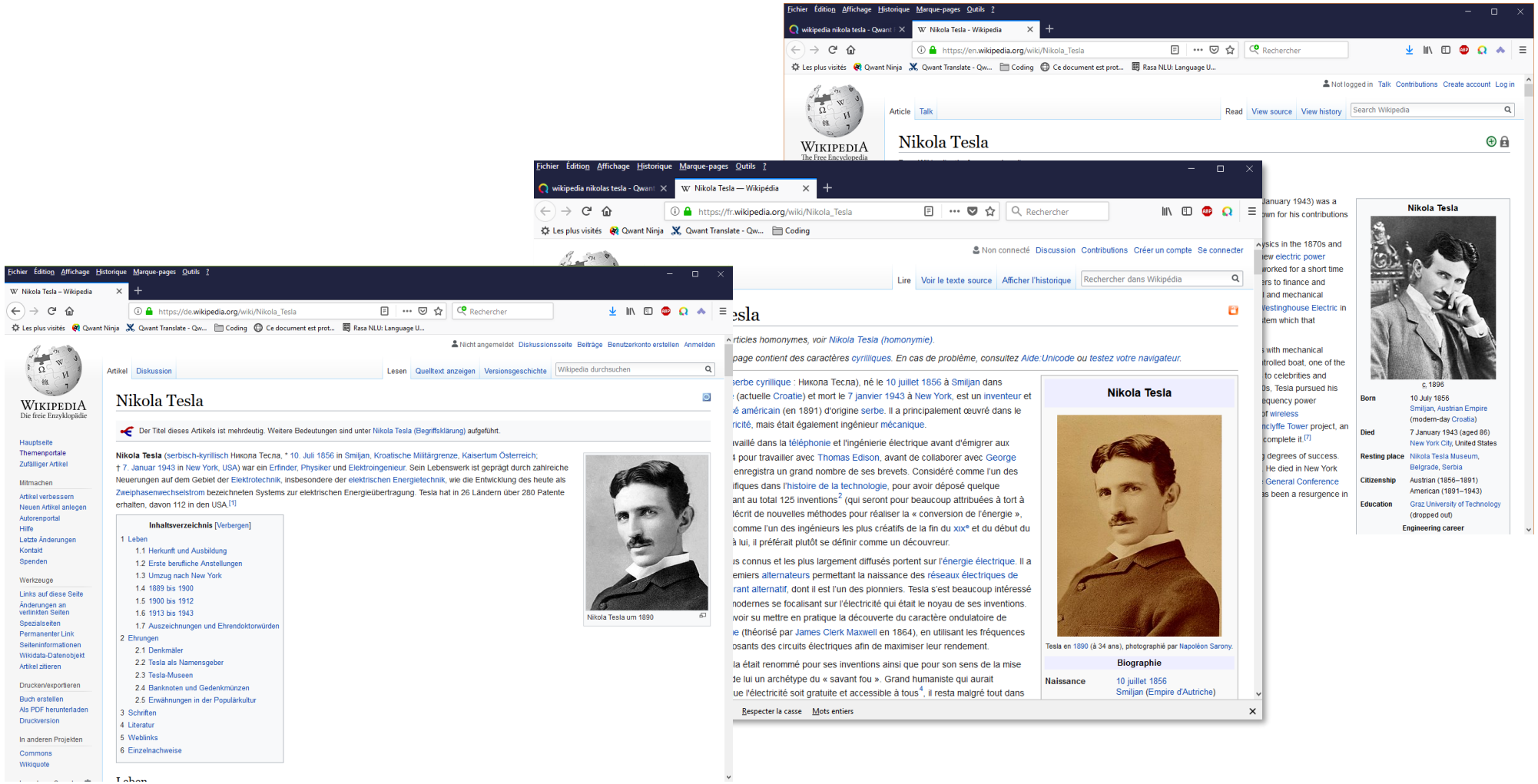
QUERYING IN SEVERAL LANGUAGES

Query: Nikola Tesla wikipedia



QUERYING IN SEVERAL LANGUAGES

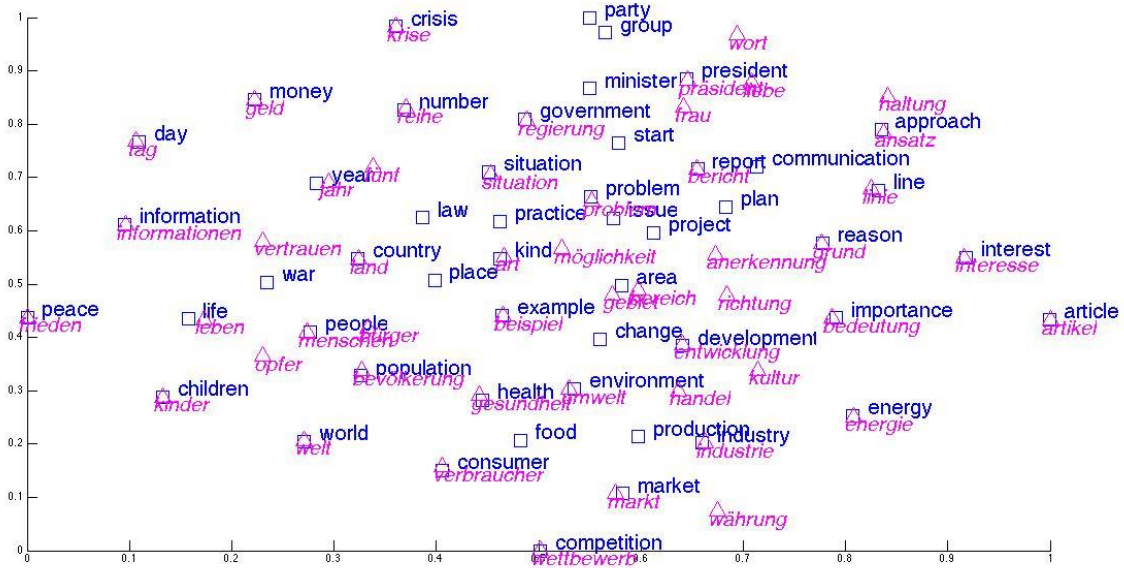
Query: Nikola Tesla wikipedia



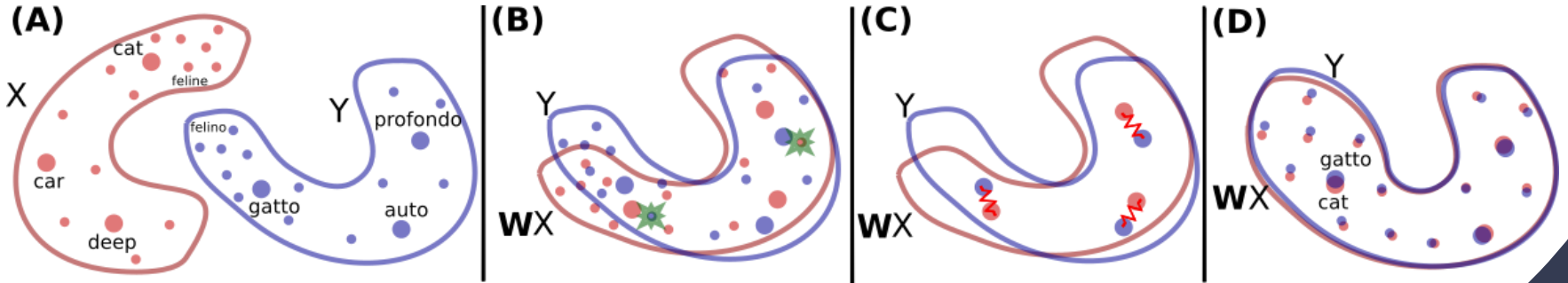
QUERYING IN SEVERAL LANGUAGES

Embeddings multilingues:

- BIVEC
  - Joint learning of word embeddings for two languages



- MUSE
  - Independant learning of several language
  - Cross-lingual alignment



QUERYING IMAGES IN SEVERAL LANGUAGES

Query: eine Frau Geige spielt

QUERYING IMAGES IN SEVERAL LANGUAGES

Query: eine Frau Geige spielt

(a) “eine Frau Geige spielt”

QUERYING IMAGES IN SEVERAL LANGUAGES

Query: eine Frau Geige spielt

(a) “eine Frau Geige spielt”

(b) “Une femme jouant du violon”

QUERYING IMAGES IN SEVERAL LANGUAGES

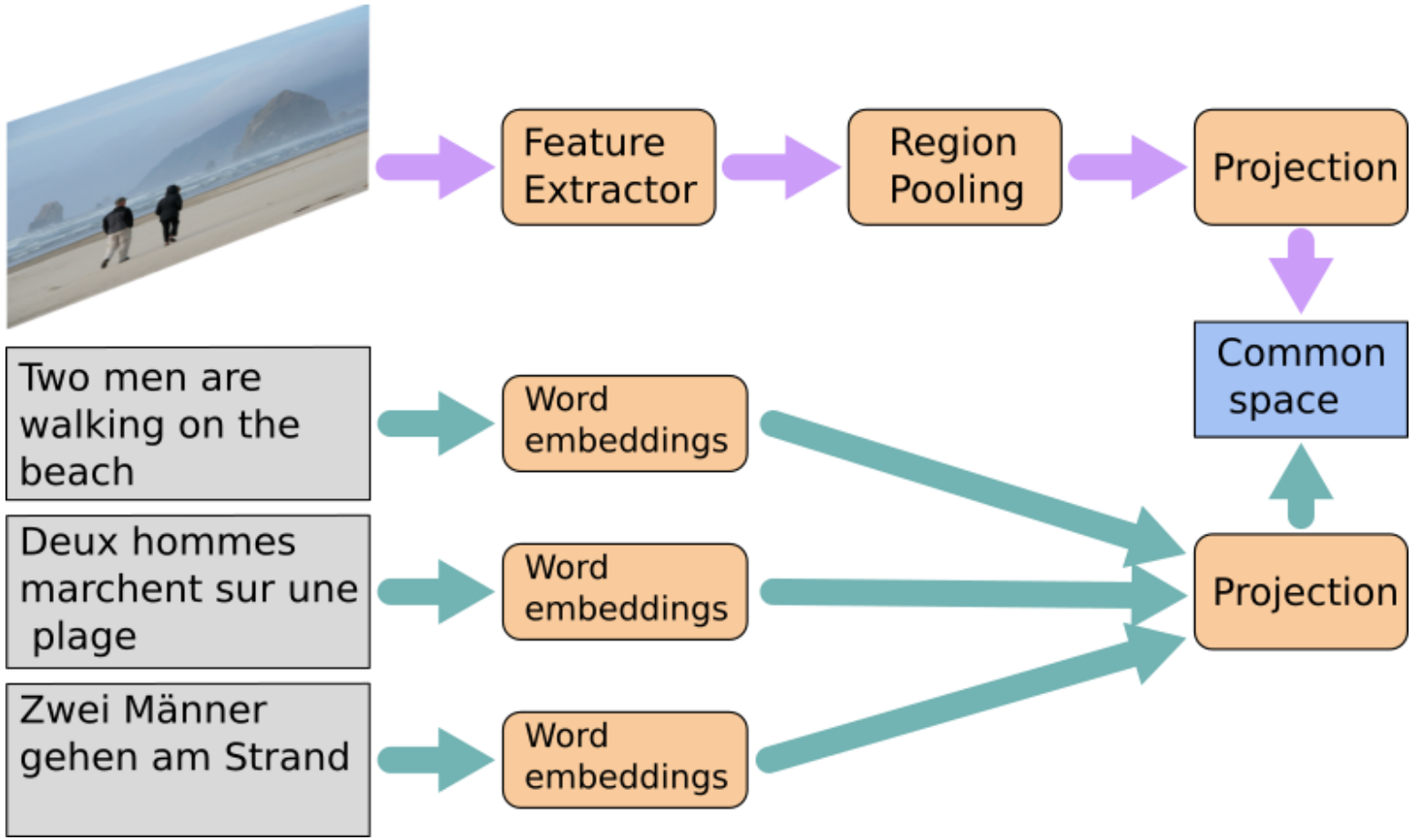
Query: eine Frau Geige spielt

(a) “eine Frau Geige spielt”

(b) “Une femme jouant du violon”

(c) “A woman playing violon”

QUERYING IN SEVERAL LANGUAGES

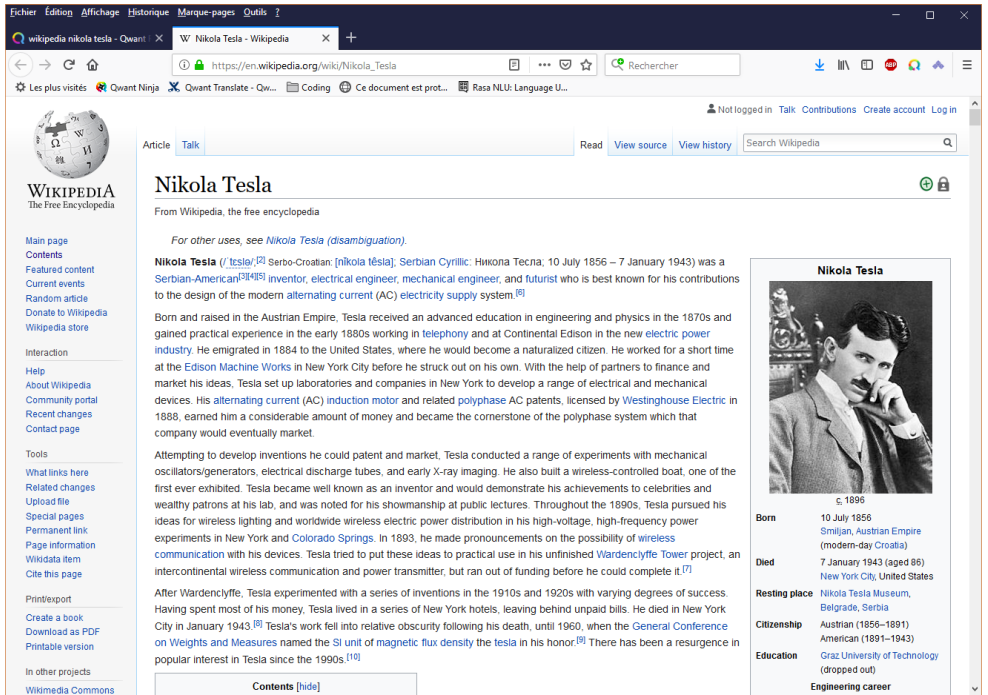


QUESTION ANSWERING IN NATURAL LANGUAGE

- **Query**: When was born Nikolas Tesla?

QUESTION ANSWERING IN NATURAL LANGUAGE

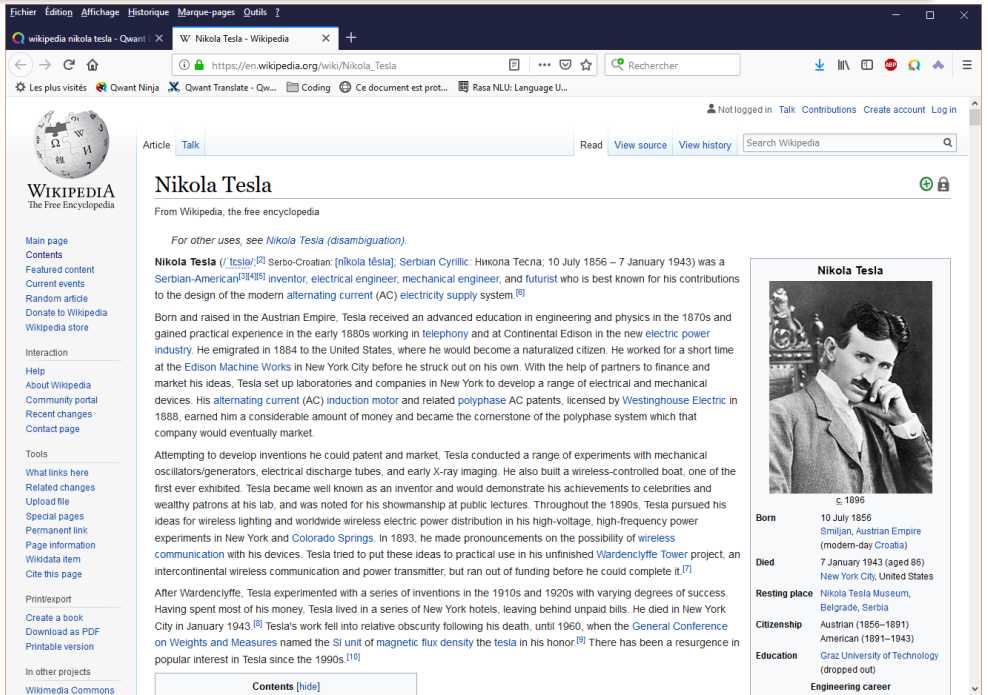
- **Query**: When was born Nikolas Tesla?



QUESTION ANSWERING IN NATURAL LANGUAGE

- **Query**: When was born Nikolas Tesla?

> « **Nikola Tesla** (/ˈtɛslə/;[2] Serbo-Croatian: [nǐkola têsla]; Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian-American[3][4][5] inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system. »

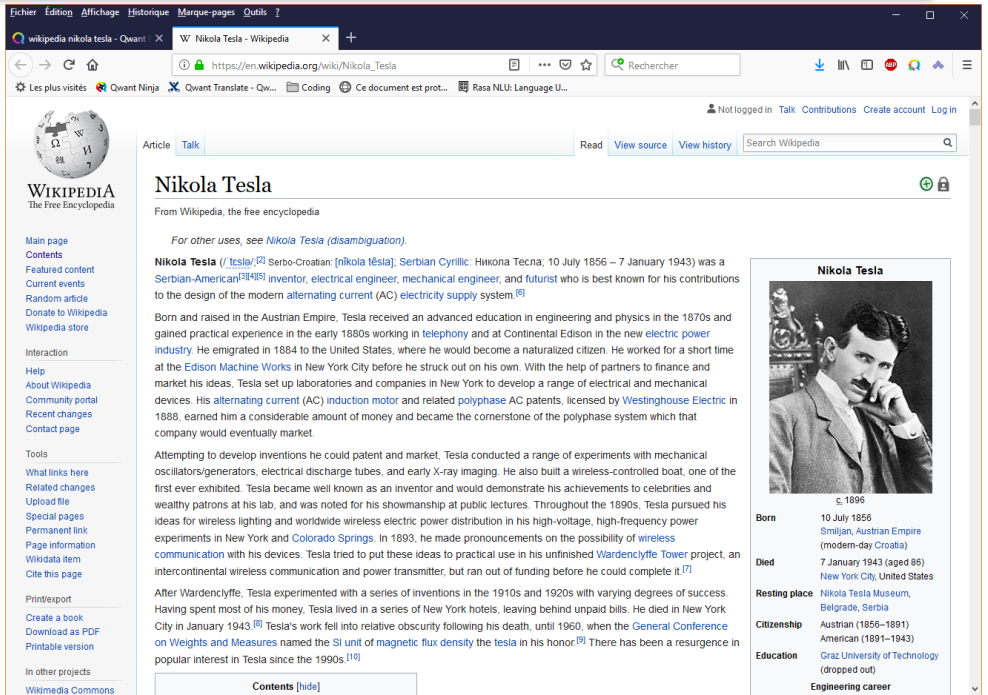


QUESTION ANSWERING IN NATURAL LANGUAGE

- **Query**: When was born Nikolas Tesla?

> « **Nikola Tesla** (/ˈtɛslə/;[2] Serbo-Croatian: [nǐkola têsla]; Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian-American[3][4][5] inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system. »

10 July 1856



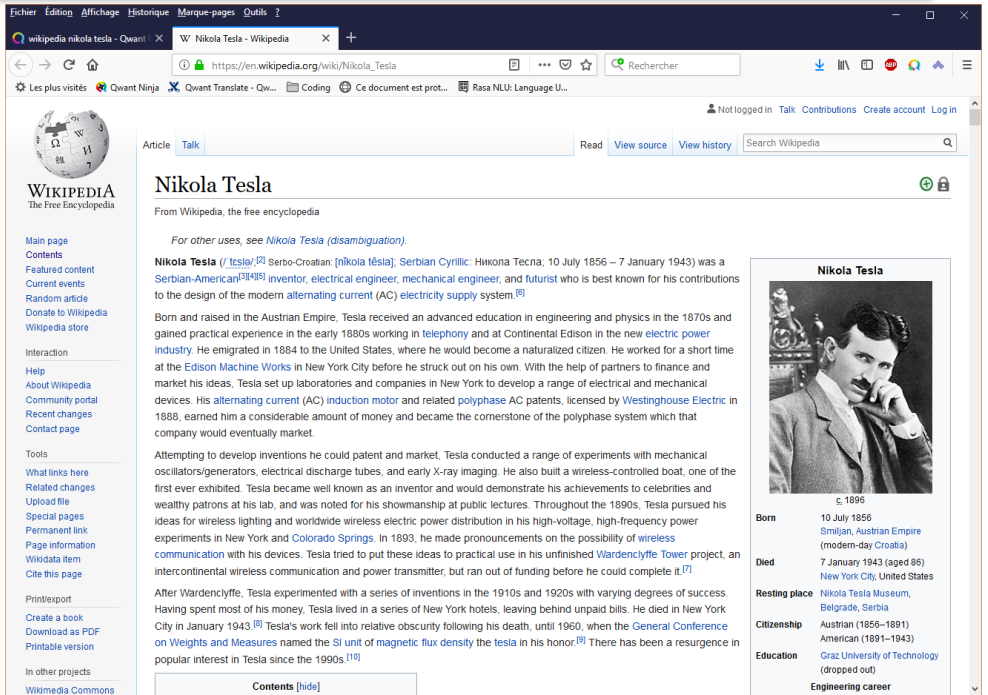
QUESTION ANSWERING IN NATURAL LANGUAGE

- **Query**: When was born Nikolas Tesla?

> « **Nikola Tesla** (/ˈtɛslə/;[2] Serbo-Croatian: [nǐkola têsla]; Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian-American[3][4][5] inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system. »

10 July 1856

« Nikolas Tesla was born on the 10 July 1856 »



LEARNING PROCESSES IN A PRACTICAL WAY

Qwant

LEARNING PROCESSES

- Lightly supervised learning
- Bootstrapping
- Continuous learning

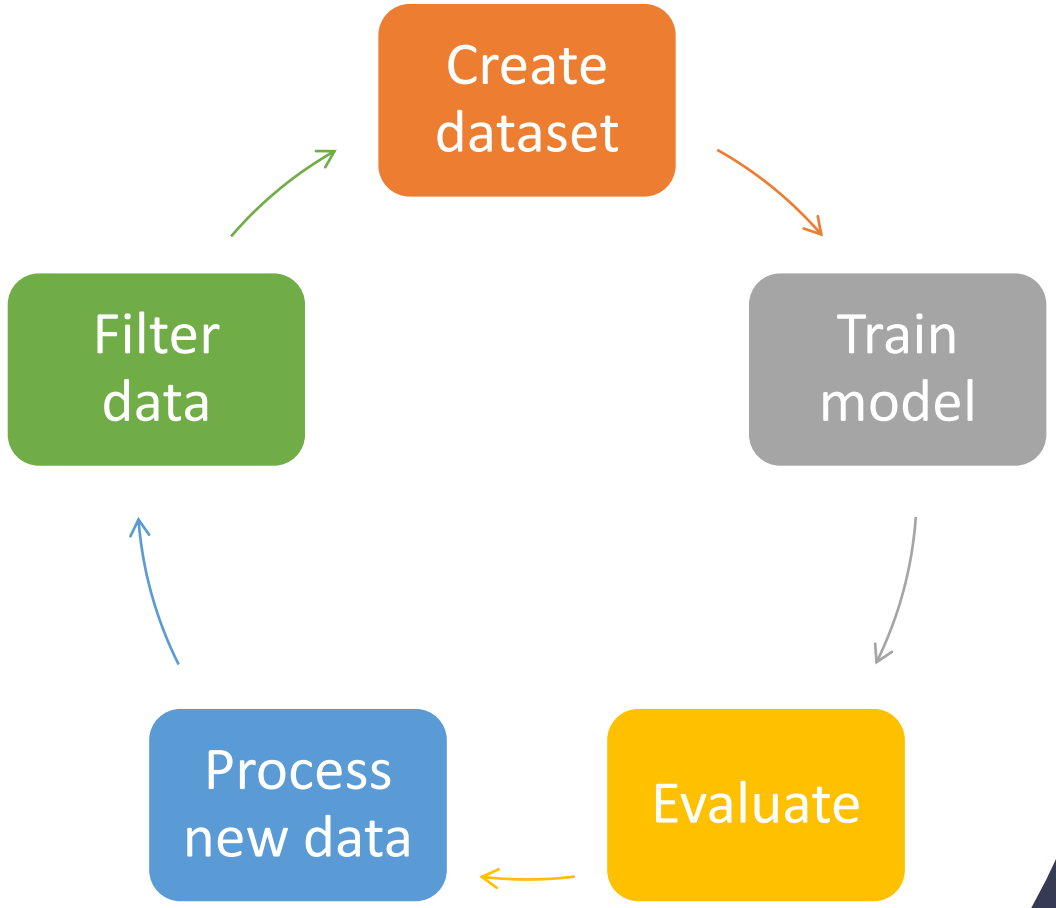
LIGHTLY SUPERVISED LEARNING

- Automatic extraction of information
  - Query & clic logs, databases, etc.
- Automatic generation of example using templates
  - I want a hotel in ${CITY} for 3 days
- Clustering
  - Automatic extraction of interesting clusters
  - Selection of interesting clusters (manually done)

$\Rightarrow$ Human in the loop

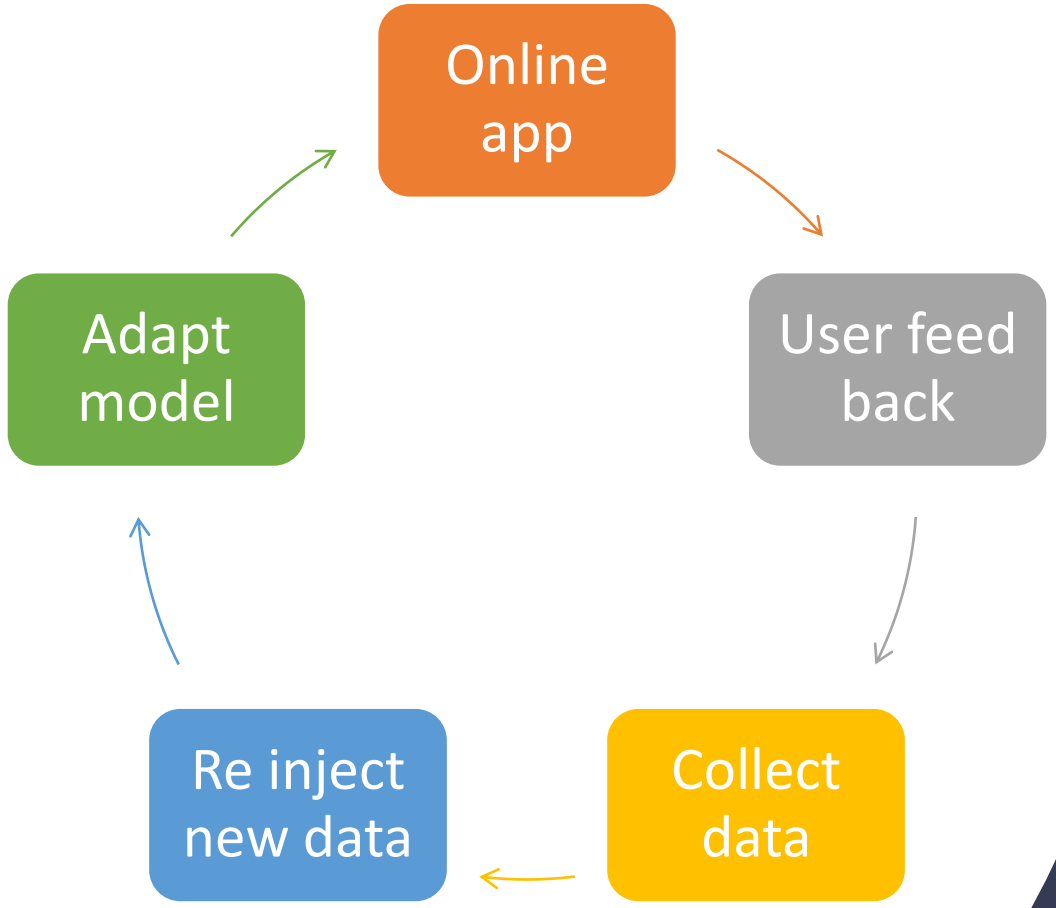
BOOTSTRAPPING

- Star from a small amount of data
- Train et evaluate it
- Apply the new model on new data
- Filter them / correction
- Add them into the new data



CONTINUOUS LEARNING

- From runnng data & models
- Request user feedback
- Collect data
- Re-inject them into the model
- Adapt model



CONCLUSION

Qwant

NLP 4 IR

- Language Models
- Machine Translation
- Intention Detection
- NER
- Spoken and Natural Language Understanding
- Word Embeddings
- …and many others!

PROJECTS

- **Projets internes:**
- Détection d'Intention
- Shopping
- Agent conversationnel (Machine Reading, Thèse)
- Qwant Junior (QR, calcul, définitions, conjugaisons…)
- Traduction Automatique

- **Projets Externes:**
- Cosy Cloud: indexation
- Caliopen: classification de courrier électronique

- **Projets Subventionnés:**
- INRIA CodeScope: Analyse de code & indexation
- INRIA Fake NewsPIA: Fact checking, Comprehension, Classification
- H2020 Answers: Analyse de sentiment
- H2020 AI4EU: Indexation de contenu web
- PIA Social Truth: Analyse de sentiment, reseaux sociaux

NLP 4 IR CHALLENGES

- Improving IR Precision

- Scaling (searching in a billion document collection)

- Speed (answering in less than ms)

⇒ improve user satisfaction
⇒ improve user experience
⇒ github.com/QwantResearch/

THE END!

QUESTIONS?

c.servan[AT]qwant[DOT]com

Qwant