A GUIDED TOUR ON NLP APPLICATIONS FROM DOCUMENT CLASSIFICATION TO SENTENCE UNDERSTANDING

Agro–IODAA, semestre 1

Vincent Guigue

INRAE AgroParisTech université PARIS-SACLAY

Around textual data: a tour on NLP tasks

Think as a data-scientist:

- No magic !
- For each application :
  1. Identify the problem class (regression, classification, ...)
  2. Find global Input / output
  3. Think about the data format (I/O)
  4. Find a model to deal with this data
  5. **Optimize all parameters & hyper-parameters**
     - **Conduct an experiment campaign**

$$\{(\mathbf{x}_i, y_i)\}_{i=1,\dots,N} \quad \Rightarrow \text{ learn:} \quad f_{\theta,\phi}(\mathbf{x}) \approx y$$

Let's have a tour on NLP applications !

Your role = identifying (x, y) for all applications

Different levels of analysis

1. At the document level
   - Topic/sentiment classification
   - etc...
2. At the paragraph / sentence level
   - Co-reference resolution
   - Nammed Entity Recognition / Relation Extraction
   - Sentiment classification (also)
3. At the word level
   - Synonymy
4. At the stream level
   - Topic detection & tracking

$\Rightarrow$ Different levels often refer to different format of data (tabular, sequence, stream)

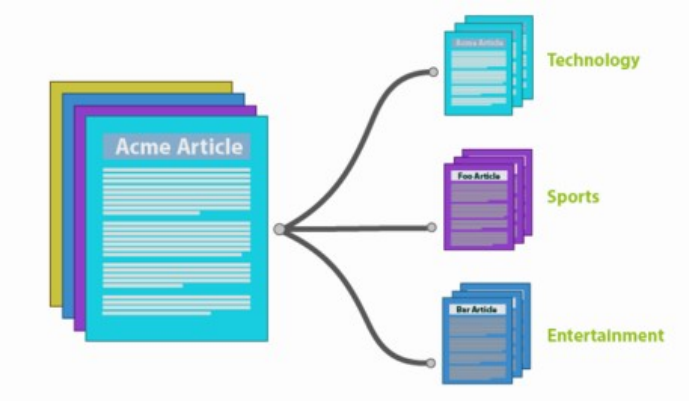
At the document level

Tasks

- **Indexing**
  - cf Information Retrieval
- **Classification / filtering**
  - topic
- **Counting / survey**
  - Sentiment classification
- **Clustering** (topic analysis)
  - Unsupervised formulation
  - Large corpus primary exploration
  - Lexical field extraction

Tasks

- **Indexing**
  - cf Information Retrieval
- **Classification / filtering**
  - topic
- **Counting / survey**
  - Sentiment classification
- **Clustering** (topic analysis)
  - Unsupervised formulation
  - Large corpus primary exploration
  - Lexical field extraction

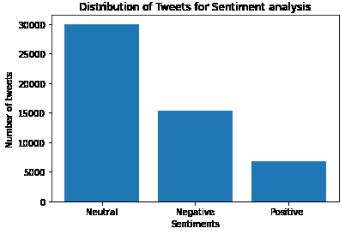


Tasks

- **Indexing**
  - cf Information Retrieval
- **Classification / filtering**
  - topic
- **Counting / survey**
  - Sentiment classification
- **Clustering** (topic analysis)
  - Unsupervised formulation
  - Large corpus primary exploration
  - Lexical field extraction

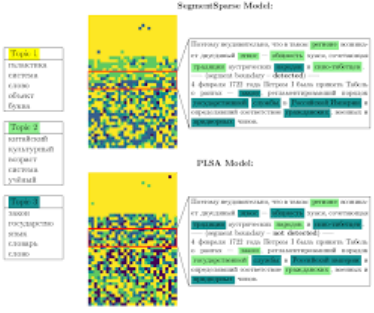
Tasks

- **Indexing**
  - cf Information Retrieval
- **Classification / filtering**
  - topic
- **Counting / survey**
  - Sentiment classification
- **Clustering** (topic analysis)
  - Unsupervised formulation
  - Large corpus primary exploration
  - Lexical field extraction



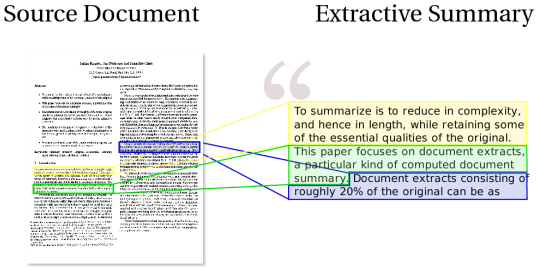
Still at the document level?

- **Document Segmentation**
  - Sentence classification
  - Link with stream
  - Break detection
- **Automated summary**
  - Sentence extraction
  - Generative architecture
- **Translation**
  - Mostly at the sentence level



Still at the document level?

- **Document Segmentation**
  - Sentence classification
  - Link with stream
  - Break detection
- **Automated summary**
  - Sentence extraction
  - Generative architecture
- **Translation**
  - Mostly at the sentence level



At the paragraph level

At the paragraph level

- **Question Answering (QA)**
  - Classification formulation
  - Extraction approach (Siri/Google assistant)
  - An emerging task... For several other tasks!
- **Coreference resolution**
- **Information extraction**
  - Mainly a sentence level task...
  - Sometimes at the paragraph level (with coreference / QA)
- **Dialog State Tracking**
  - Chatbot...

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

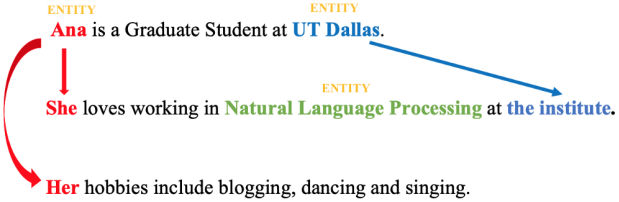
- Between question and answer

cause---gravity
precipitation---gravity
fall---gravity
what---gravity

WARNING: a mix of old & new tasks

At the paragraph level

- **Question Answering (QA)**
  - Classification formulation
  - Extraction approach (Siri/Google assistant)
  - An emerging task... For several other tasks!
- **Coreference resolution**
- **Information extraction**
  - Mainly a sentence level task...
  - Sometimes at the paragraph level (with coreference / QA)
- **Dialog State Tracking**
  - Chatbot...



WARNING: a mix of old & new tasks

At the paragraph level

- **Question Answering (QA)**
  - Classification formulation
  - Extraction approach (Siri/Google assistant)
  - An emerging task... For several other tasks!
- **Coreference resolution**
- **Information extraction**
  - Mainly a sentence level task...
  - Sometimes at the paragraph level (with coreference / QA)
- **Dialog State Tracking**
  - Chatbot...

```
Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.
Q: Where is the apple?
A. Bedroom
```

Note that for each question, a

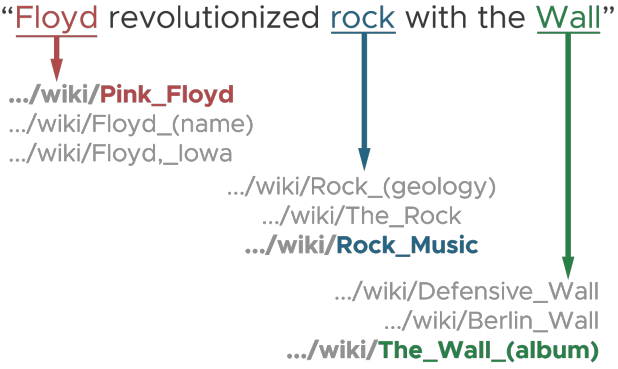
WARNING: a mix of old & new tasks

At the sentence level

At the sentence level

- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**



At the sentence level

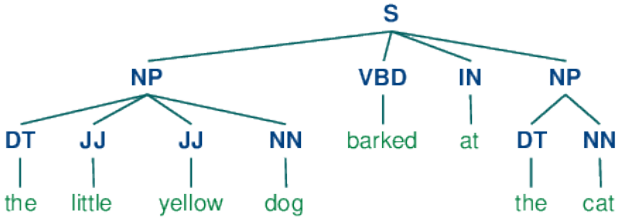
- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**

At the sentence level

- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

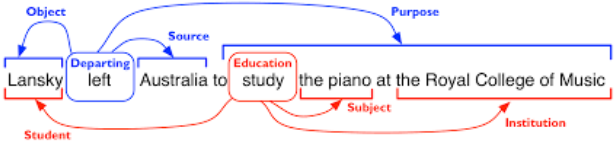
  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**



At the sentence level

- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

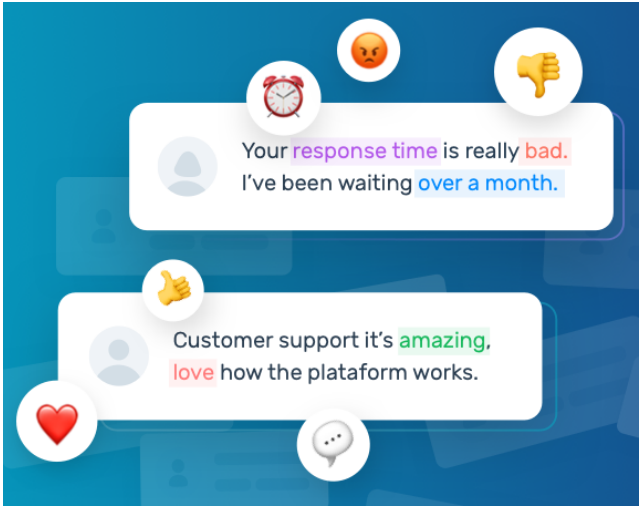
  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**



At the sentence level

- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**



At the sentence level

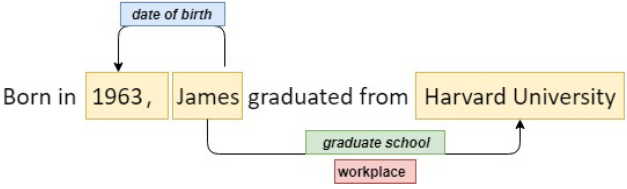
- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**

At the sentence level

- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

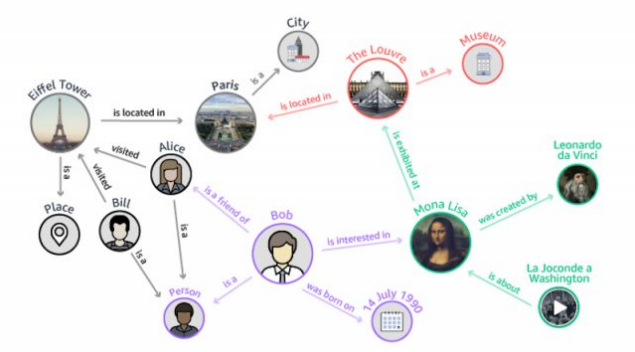
  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**



At the sentence level

- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

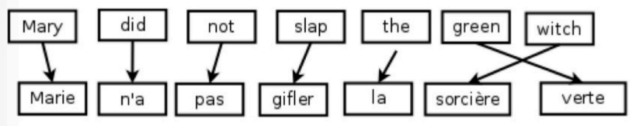
  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**



At the sentence level

- **Word/entity resolution**
- **Part-Of-Speech** tagging (POS)
- **Chunking** (sentence segmentation) = Syntactic Parsing
- **Semantic Role Labeling** (SRL)
- **Sentiment Analysis** (fine grained)
- **Named Entity Recognition** (NER)
- **Information Extraction**
  - Entity resolution
  - Relation extraction

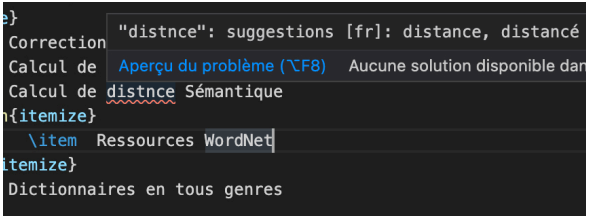
  $\Rightarrow$ Knowledge Graph ([opt] + ontologies...)
- **Machine Translation**



At the word level

At the word level

- Syntactic similarities
  - spelling correction
  - Levenshtein = DTW
- Semantic distance (synonymy)
  - WordNet (& other resources)
  - Representation learning



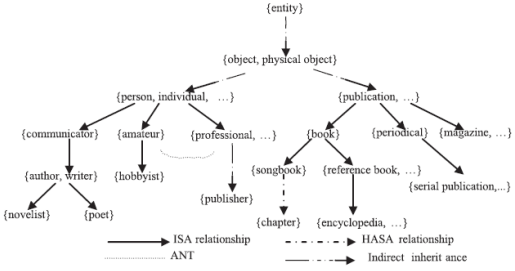
At the word level

- Syntactic similarities
  - spelling correction
  - Levenshtein = DTW
- Semantic distance (synonymy)
  - WordNet (& other resources)
  - Representation learning

		H	Y	U	N	D	A	I
	0	1	2	3	4	5	6	7
H	1	0	1	2	3	4	5	6
O	2	1	1	2	3	4	5	6
N	3	2	2	2	2	3	4	5
D	4	3	3	3	3	2	3	4
A	5	4	4	4	4	3	2	3

At the word level

- Syntactic similarities
  - spelling correction
  - Levenshtein = DTW
- Semantic distance (synonymy)
  - WordNet (& other resources)
  - Representation learning



Toward composite or multimodal analysis

Temporal streams

- Historical tasks: **Topic Detection and Tracking (TDT)**
- = adding a temporal axis to the topic detection
  - topic quatification
  - topic appearance
  - topic vanishing

Temporal streams

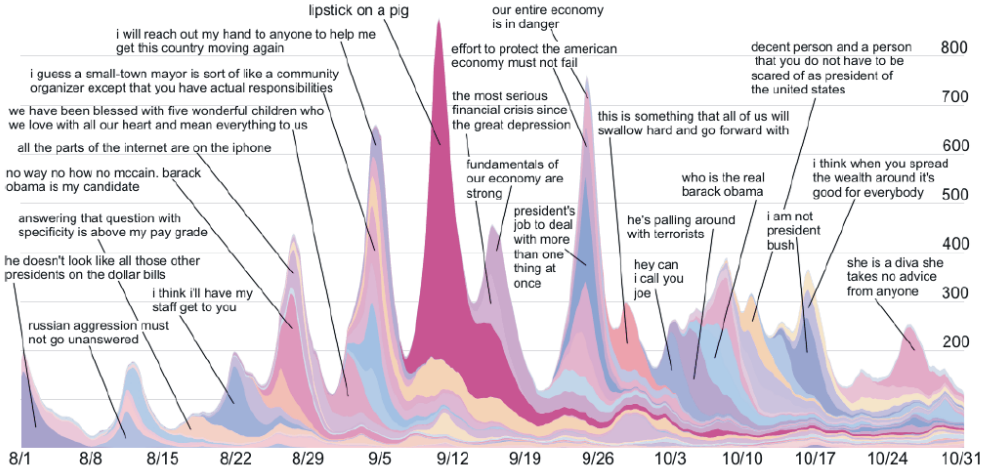


Figure 4: Top 50 threads in the news cycle with highest volume for the period Aug. 1 – Oct. 31, 2008. Each thread consists of all news articles and blog posts containing a textual variant of a particular quoted phrases. (Phrase variants for the two largest threads in each week are shown as labels pointing to the corresponding thread.) The data is drawn as a stacked plot in which the thickness of the strand corresponding to each thread indicates its volume over time. Interactive visualization is available at http://memetracker.org.

Profiling

Profiling = recommender system

- modern User Interface (UI)
- Information Access
  - Personalization
  - Active suggestion (user = query)

Recent proposal: mixing user interaction & textual review

- user interaction = best item similarity
- review = in depth textual description

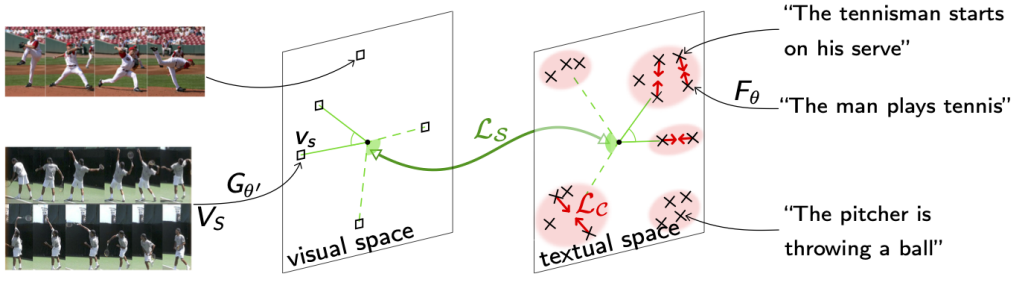
McAuley & Leskovec 2013

Text & Image

- Visual grounding
- Visual Question Answering
- Captioning
- Image Generation

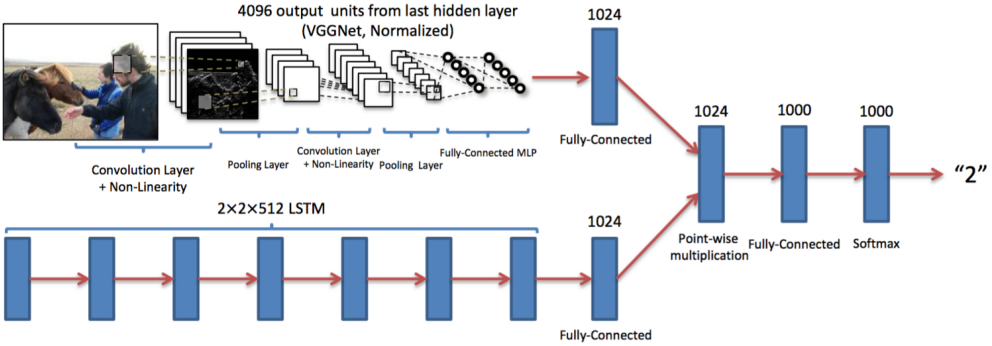
Word	*Teraword*	*Knext*
Spoke	11,577,917	372,042
Laughed	3,904,519	179,395
Murdered	2,843,529	16,890
Inhaled	984,613	5,617
Breathed	725,034	41,215

Word	*Teraword*	*Knext*
Hugged	610,040	11,453
Blinked	390,692	21,973
Was late	368,922	31,168
Exhaled	168,985	4,052
Was on time	23,997	14



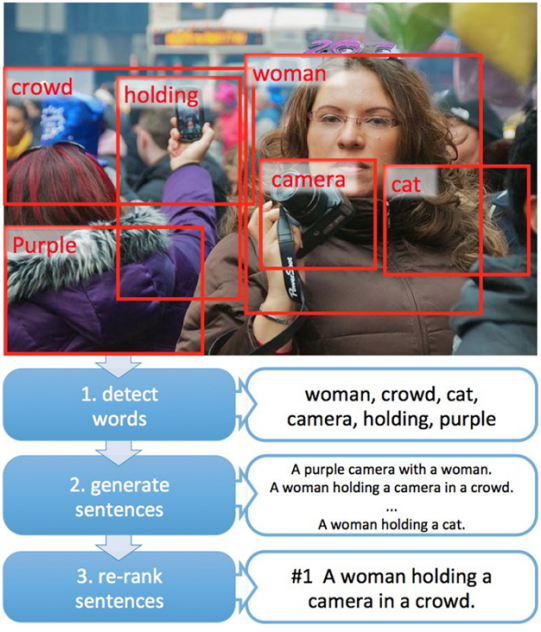
Text & Image

- Visual grounding
- Visual Question Answering
- Captioning
- Image Generation



Text & Image

- Visual grounding
- Visual Question Answering
- Captioning
- Image Generation



Text & Image

- Visual grounding
- Visual Question Answering
- Captioning
- Image Generation

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

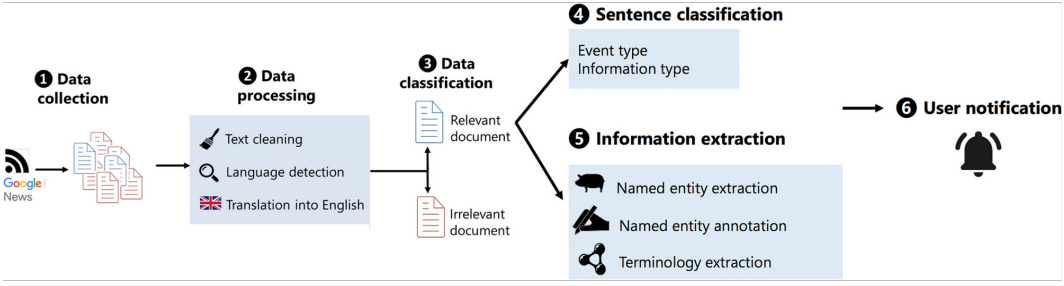
riding a horse lounging in a tropical resort in space playing basketball with cats in space

in a photorealistic style in the style of Andy Warhol as a pencil drawing

→

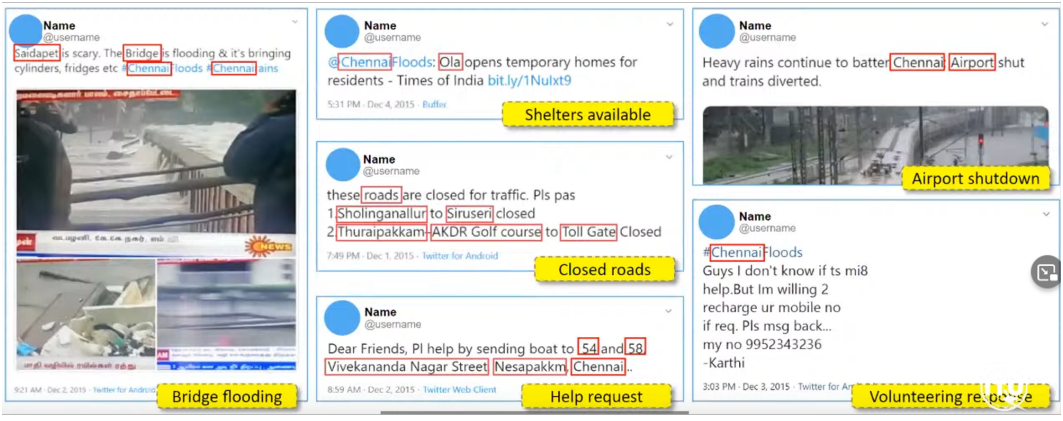
DALL·E 2

Epidemiology



Valentin S. et al., PADI-web 3.0: framework for extracting fine-grained information from news for animal disease surveillance.

Event detection



Suwaileh R., Elsayed T., Imran M. (2023). IDRISI-RE: A generalizable dataset with benchmarks for location mention recognition on disaster tweets. Information Processing & Management, https://doi.org/10.1016/j.ipm.2023.103340

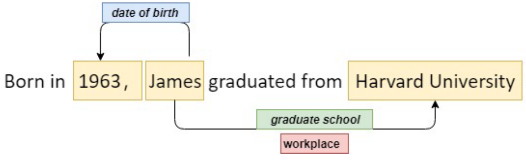
From text (NLP) to Knowledge Bases (KB)

Text = lot of resources ⇒ Universal knowledge?

- Noisy (both at the syntactic & semantic levels)
- Hard to query & exploit (search, knowledge inference,...)

(RDF) knowledge

- Def: Entity 1 (subject/head) Relation (predicate) Entity 2 (target/tail)
- Simpler version: Key / value
- Easy to handle



From text (NLP) to Knowledge Bases (KB)

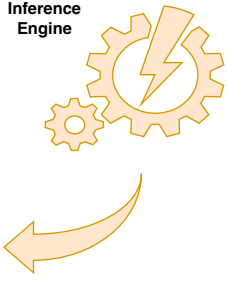
Text = lot of resources ⇒ Universal knowledge?

- Noisy (both at the syntactic & semantic levels)
- Hard to query & exploit (search, knowledge inference,...)

From RDF to ontology: adding hierarchies + reasoning

Fact base
Barack Obama was born in Honolulu
Honolulu is the capital of Hawaii
Hawaii is a state of the USA
...

Rule base
capital => part of
state of => part of
born in => citizen of [USA, France]
...



From text (NLP) to Knowledge Bases (KB)

Text = lot of resources ⇒ Universal knowledge?

- Noisy (both at the syntactic & semantic levels)
- Hard to query & exploit (search, knowledge inference,...)

And advanced queries based on this structure:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
PREFIX up: <http://purl.uniprot.org/core/>
SELECT ?protein ?organism ?isoform ?sequence
WHERE
{
    ?protein a up:Protein .
    ?protein up:organism ?organism .
    # Taxon subclasses are materialized, do not use rdfs:subClassOf+
    ?organism rdfs:subClassOf taxon:83333 .
    ?protein up:sequence ?isoform .
    ?isoform rdf:value ?sequence .
}
```

Overview

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC
Einstein met with UN officials in Princeton		

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Dan Jurafsky

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity name

Where is *A Bug's Life* playing
Let It Be was recorded …
… a mutation on the *for* gen

Data formats & processing chain

What is textual data?

- A **series** of letters

```
t h e _ c a t _ i s ...
```

- A **series** of words

```
the cat is ...
```

- A **set** of words

in alphabetical order

```
cat
is
the
...
```

- N-gram dictionary:

```
BEG_the the_cat cat_is is_...
```

Standard processing chain

0. Reading : watch the encoding (utf8, latin1, ...) / file format (txt, pdf, ...)

1. Preprocessing

- encoding (latin, utf8, ...)
- ponctuation
- stemming
- lemmatization
- tokenization
- capitals/lower case
- regex
- ...

2. Formatting

- Dictionary
- + reversed Index
- Vectorial format
- Sequence format

3. Learning

- Doc / sentence / word classification
- Semantic

...
Perceptron, HMM, RNN, Transformer?

4. Hyper-parameter optimization

Conclusion

- Large number of high value applications
- (One of the) biggest community (both academic & industrial)
- A community of its own
  - (very) Specific data
  - Specific pre-processing
  - Specific models

$\Rightarrow$ Specific companies, conferences, ... until the end of 2010 decade !

NLP = Very high potential, with a significant cost