Information Extraction based on Named Entities for Tourism

Ahmed Hamdi

University of La Rochelle
Laboratoire Informatique, Image, Interaction (L3i)

April 29, 2022

Who I am?

- I am **Ahmed Hamdi** research engineer in computer science at the university of La Rochelle. I received my PhD in computational linguistics from Aix-Marseille university.

- I work on information extraction and natural language processing.

ahmed.hamdi@univ-lr.fr

Interventions (CET)

- April 29, 2022: Information Extraction based on Named Entities for Tourism
  - 3.00 p.m – 4.30 p.m : **course**
  - 4.45 p.m – 6.15 p.m : **practice work**

- May 06, 2022: Stance Detection for Tourism
  - 3.00 p.m – 4.30 p.m : **course**
  - 4.45 p.m – 6.15 p.m : **practice work**

Overview

- Context: tourism, computer science, information extraction
- Information extraction for tourism
- Word embedding for information extraction
- Named entities
- Named entity recognition and linking

Context

Information extraction is the process of extracting information from unstructured textual sources to enable finding entities as well as classifying and storing them in a database.

Information extraction using machines

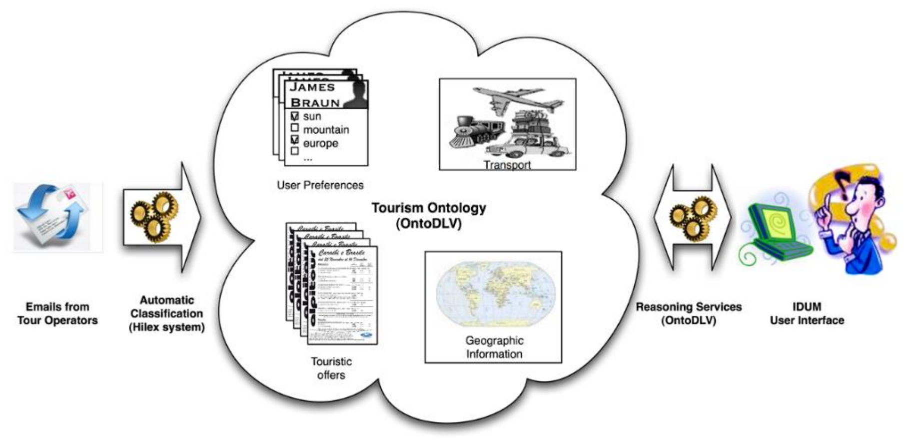
Information extraction using machines allows extracting relevant information from large amount of textual data in a short period of time

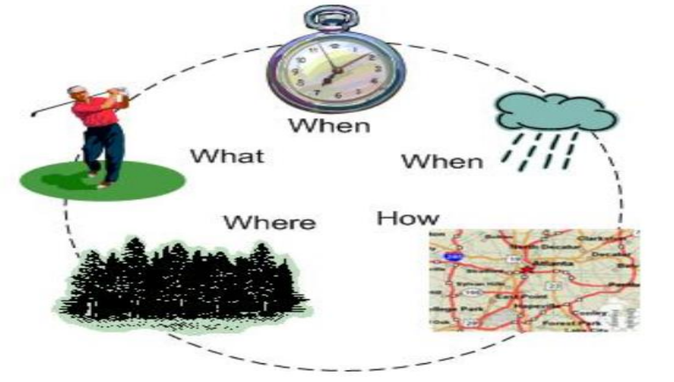
Examples:

- Trend analysis
- Topic modeling
- Classification of opinions
- Summarisation

What about tourism?

- Scattered Tourism information

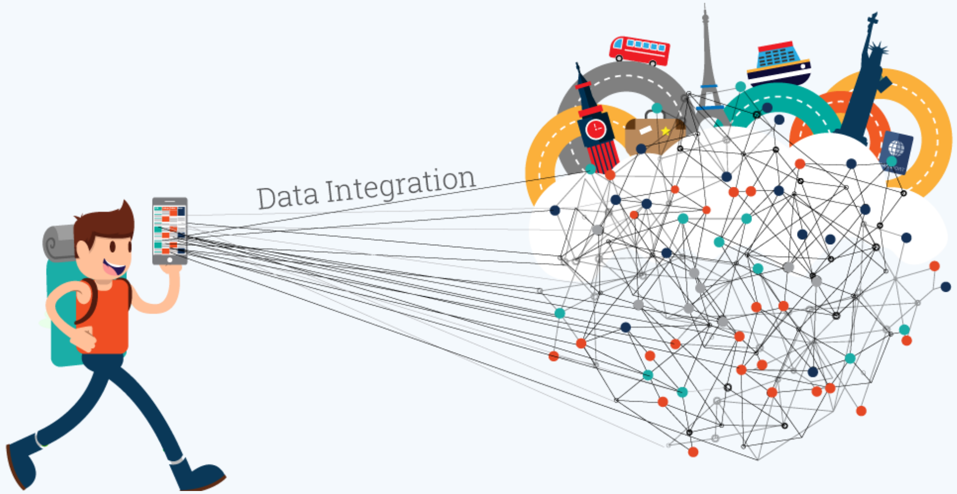




Inofrmation Extraction for Tourism

- Extract relevant information about a topic
- Build ontologies
- Classify the opinions of customers
- Determine fake news

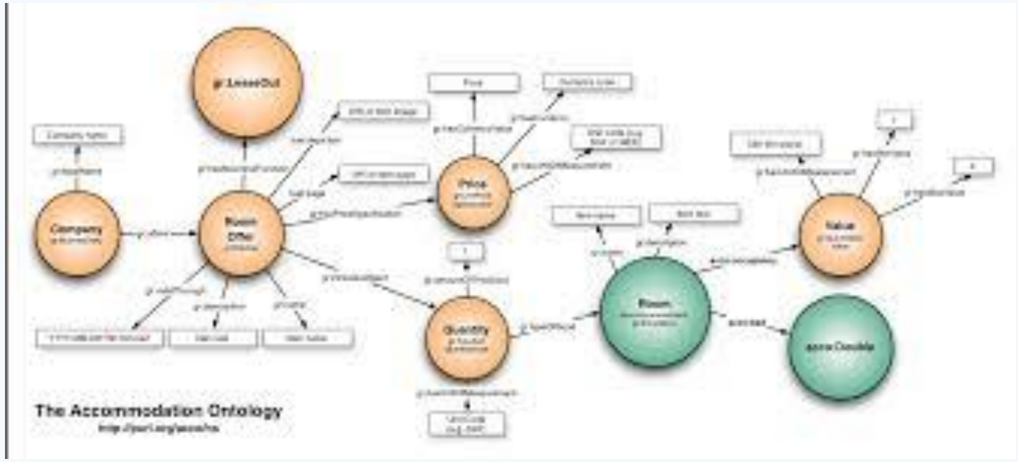
DATAtourisme



Ontologies for Tourism

➔ Describes hotel rooms, hotels, camping sites, and other types of accommodations, their features, and modelling compound prices as frequently found in the tourism sector

- Existing tourism ontologies
  - Morocco tourism ontology.
  - Mondeca tourism ontology in OnTour
  - the accommodation ontology STI Innsbruck



Summary

- Find and understand limited relevant parts of texts
- Gather information from many pieces of text
- Produce structured representations of relevant information
  - Relations, knowledge base
- Goals
  - Organize information so that it is useful to people
  - Put information in semantically precise form

Word embedding

Word embedding

- Represent each word from a vocabulary by a vector of real numbers

home →	0.01	-0.05	0.58	0.19	0.98	0.67
dog →	-0.01	-0.33	0.09	0.27	0.77	0.15
cat →	0.08	0.87	-0.55	0.99	-0.91	0.04
kitten →	0.24	-0.22	-0.58	-0.64	0.48	-0.36

Bag-of-Words (BoW)

Sentence S1: I like room . room is comfortable !

Sentence S2: I do not like room. room not clean !

Vocabulary {*I, like, room, it, comfortable, do, not, clean*}

room	0	0	4	0	0	0	0	0
clean	0	0	0	0	0	0	0	1

Drawbacks

- *Frequent words may not be relevant (i.e. room)*
- *Do not take into account the meaning of words*

word2vec

- Two papers published by (Mikolov et al. 2013)

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

Greg Corrado
Google Inc.
Mountain View
gcorrado@google.com

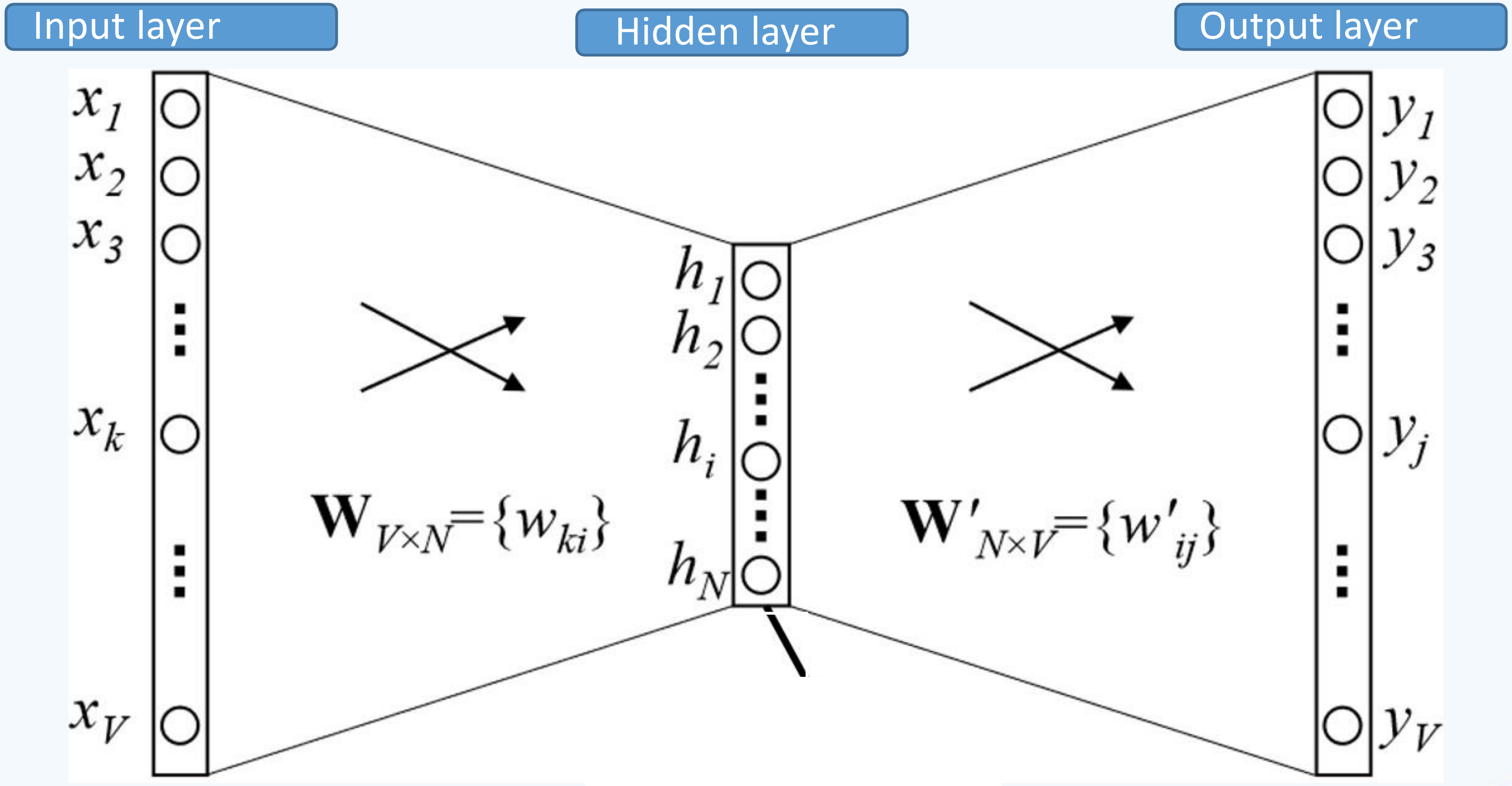
Jeffrey Dean
Google Inc.
Mountain View
jeff@google.com

Abstract

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada" and "Air" cannot be easily combined to obtain "Air Canada". Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible.

word2vec

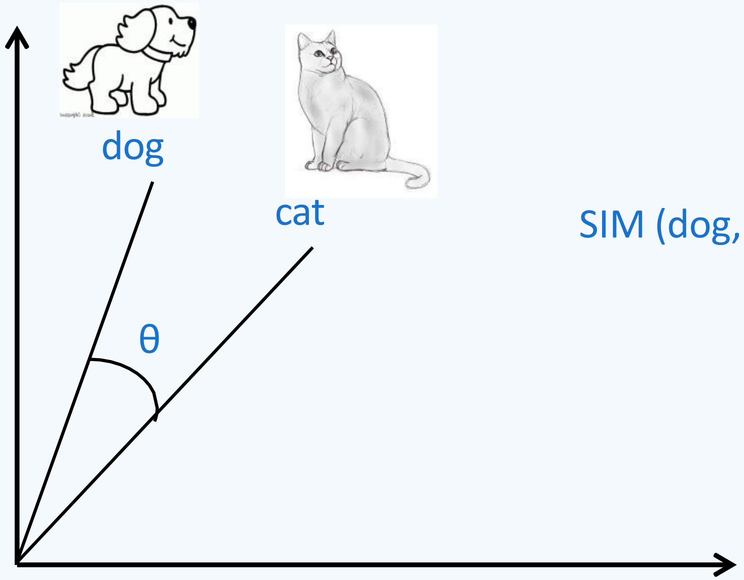


Input layer
Hidden layer
Output layer
x_1
x_2
x_3
x_k
x_V
h_1
h_2
h_i
h_N
y_1
y_2
y_3
y_j
y_V
$\mathbf{W}_{V\times N}=\{w_{ki}\}$
$\mathbf{W}'_{N\times V}=\{w'_{ij}\}$

word2vec

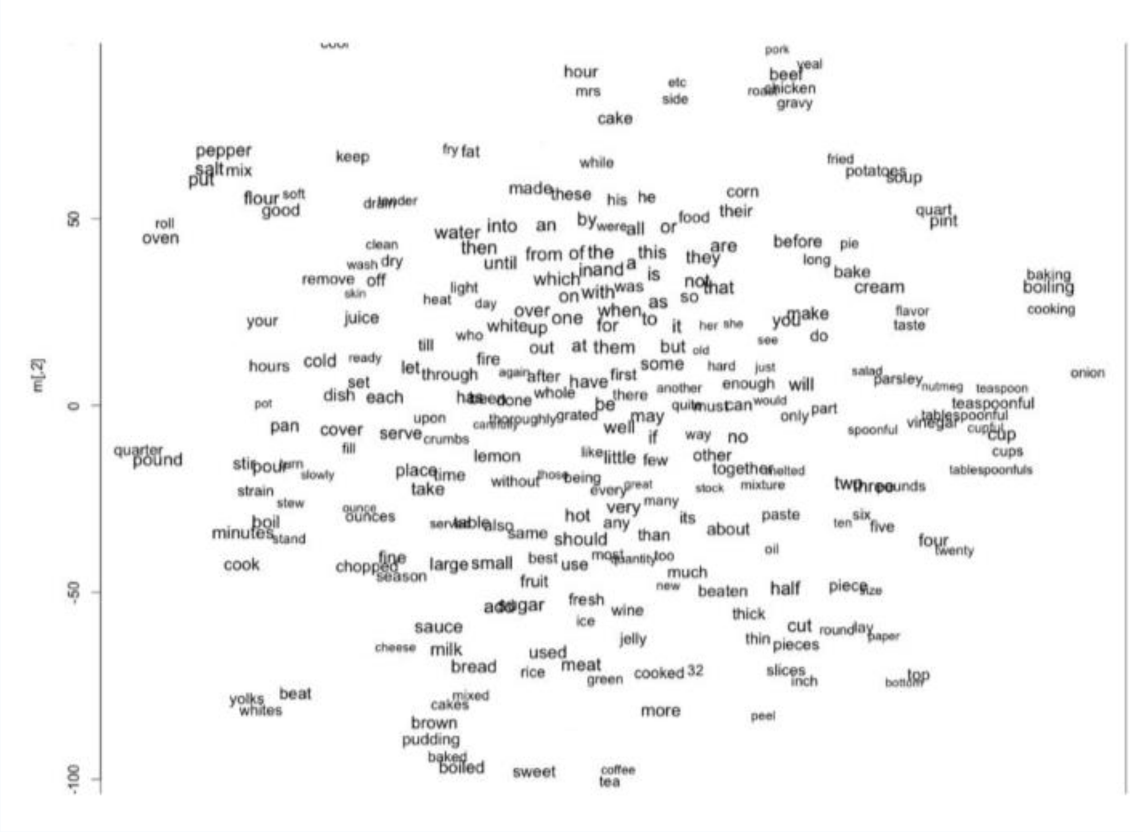
- Word similarity ➔ vector similarity

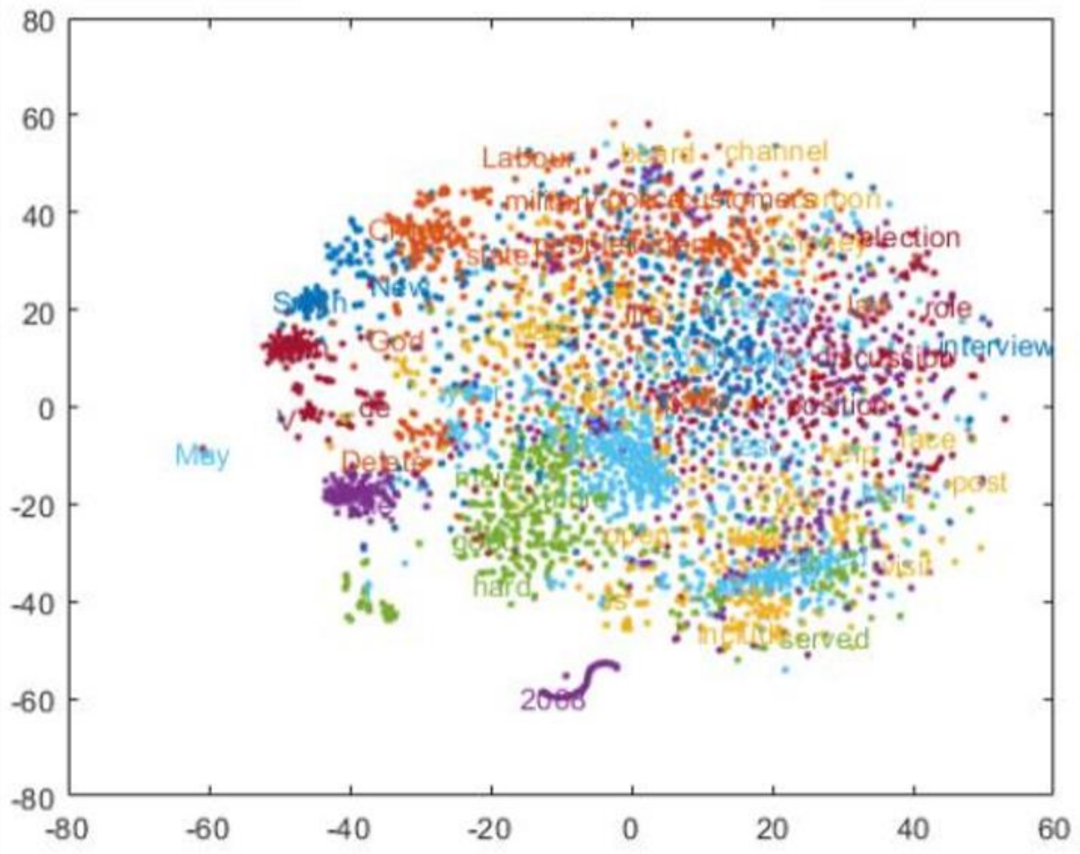
  > Cosine similarity
  > Euclidian similarity
  > Jaccard similarity



SIM (dog, cat) = $\cos(\theta)$

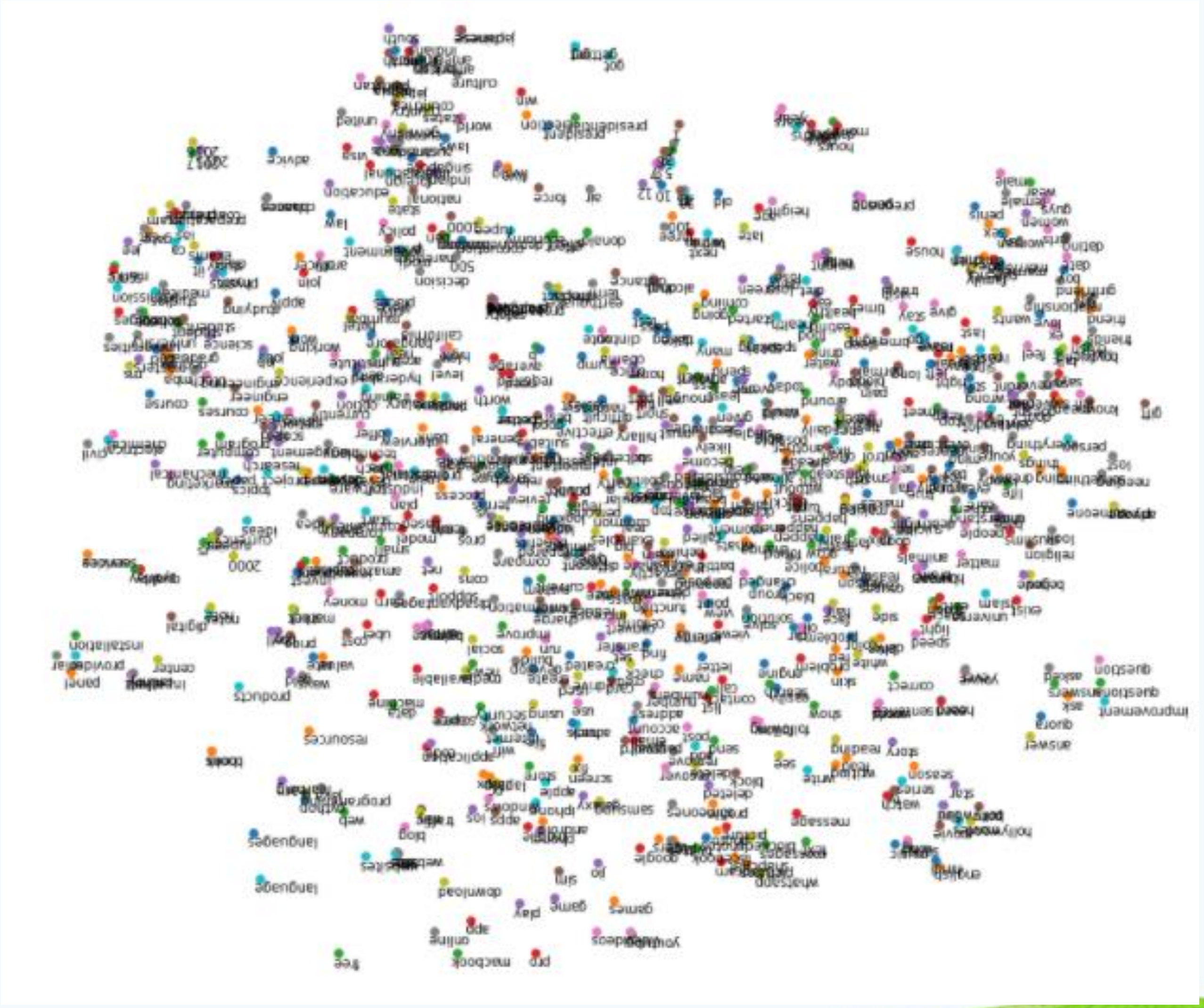
Semantic space



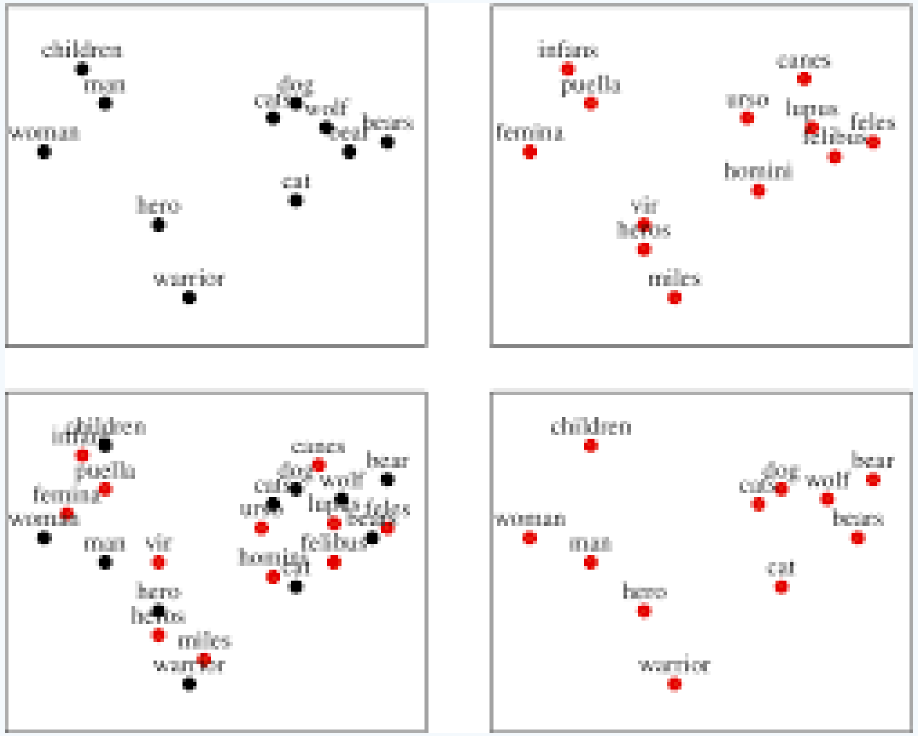


Semantic space

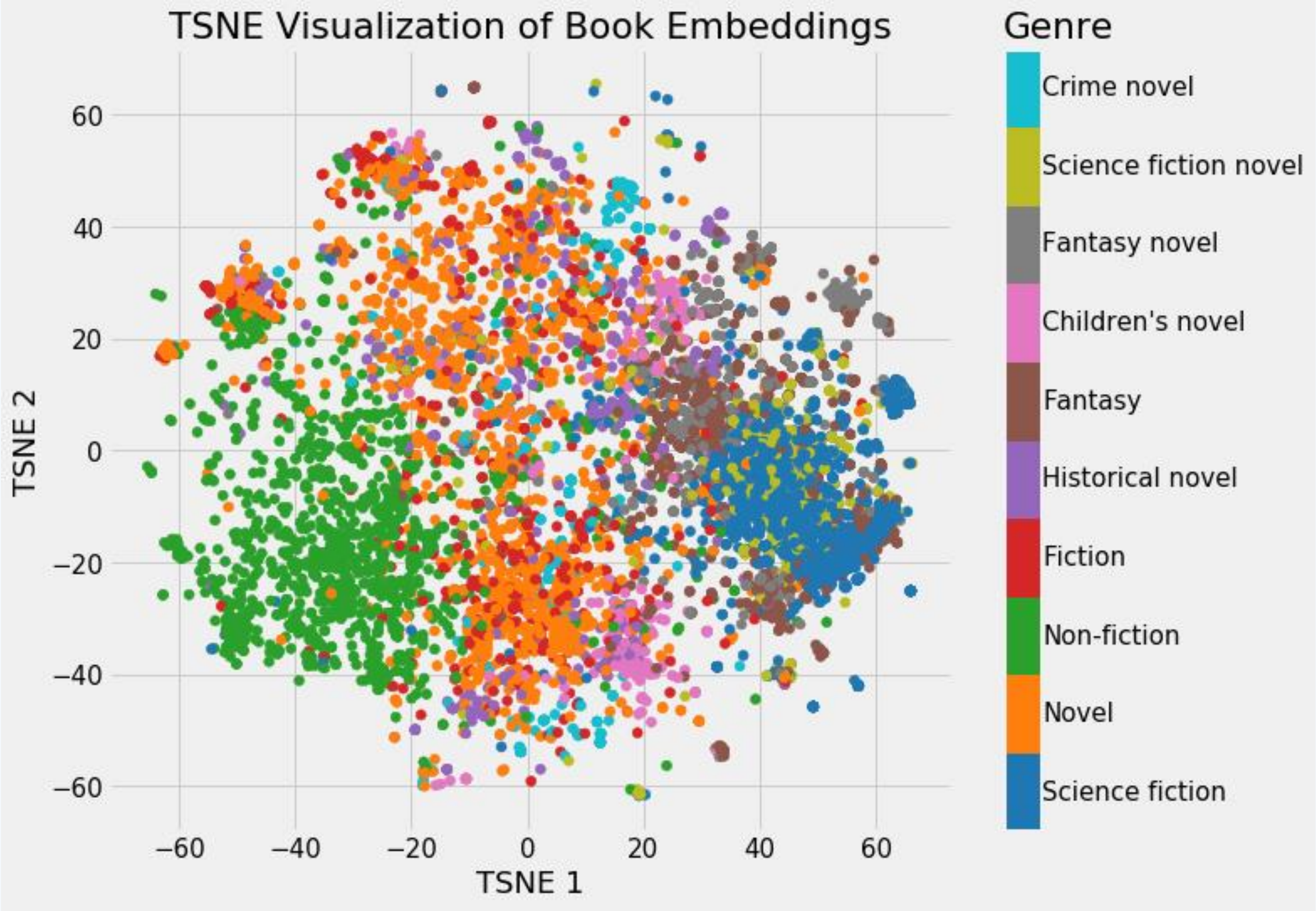
Projection of English words



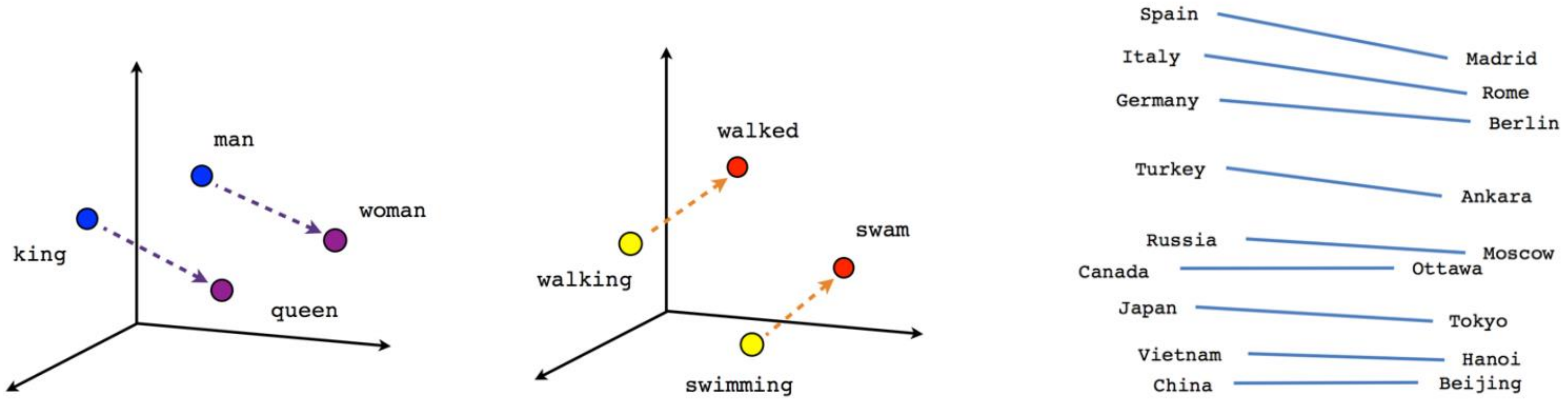
Semantic space



Semantic space



Semantic encoding



Male-Female

Verb tense

Country-Capital

Semantic similarity

- Paris – France + Germany = ?
- Women – Queen + Men = ?
- France – Euro + Russia = ?
- Good – Best + Bad = ?

Well known word embeddings

- **Word2vec** (French, English, German)
  - Google News → 3B words
  - French Wikipedia → 500M words
  - German Wikipedia → 651M words
- **FastText** (157 languages)
  - Common Crawl
  - Wikipedia
- **GloVe** (English)
  - Twitter → 27B words (2B tweets)
  - Wikipedia → 6B words

Named entities

Named entities (NE)

- A **named entity** is a real-world object denoting a unique individual with a proper name

Named entity categories

- **Person** (PER): individual or group, fictional character
- **Location** (LOC): place, city, country, zip code…
- **Organisation** (ORG): company, hotel, university…
- **Products** (PROD): human products
- **Other** (MISC): event, time, nationality…

Named entities: ambiguity

- Paris Hilton stays at the Paris Hilton
- New York Times is based in New York
- Moscow's as yet undisclosed proposals on Chechnya's political future

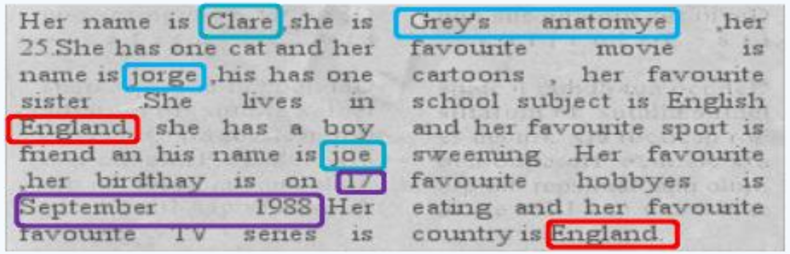
PER LOC ORG PROD

Named entity recognition

Named entity recognition (NER)

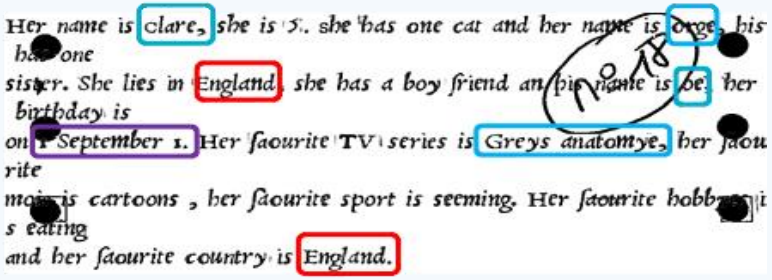
- The task consisting in locating named entities and categorizing them into classes (PER, LOC, ORG, PROD …)

Reasonable quality



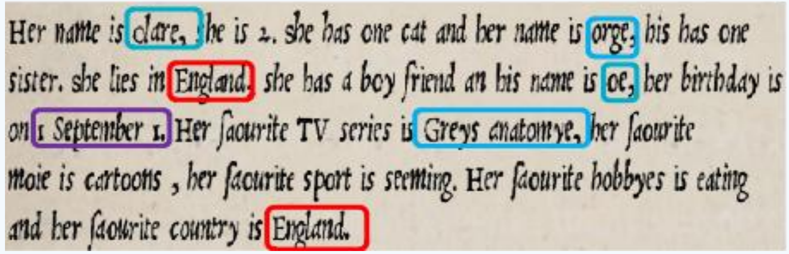
Her name is Clare she is 25.She has one cat and her name is jorge ,his has one sister .She lives in England, she has a boy friend an his name is joe ,her birdthay is on 17 September 1988 Her favourite TV series is Grey's anatomye ,her favourite movie is cartoons , her favourite school subject is English and her favourite sport is sweeming .Her favourite favourite hobbyes is eating and her favourite country is England.

Poor quality



Her name is clare, she is 5. she has one cat and her name is orge, his ha one sister. She lies in England, she has a boy friend an his name is oe, her birthday is on 1 September 1. Her faourite TV series is Greys anatomye, her faou rite moie is cartoons , her faourite sport is seeming. Her faourite hobbyes is eating and her faourite country is England.

Medium quality



Her name is clare, he is 2. she has one cat and her name is orge, his has one sister. she lies in England, she has a boy friend an his name is oe, her birthday is on 1 September 1, Her faourite TV series is Greys anatomye, her faourite moie is cartoons , her faourite sport is seeming. Her faourite hobbyes is eating and her faourite country is England.

Named entity recognition

- Rule based methods

  - **Lexicons**: list of proper names, places, organisations
  - **Trigger words**: i.e. Mr., Mrs., Ms., Dr…
  - **Regular expressions**: i.e. uppercase, acronyms…

Named entity recognition

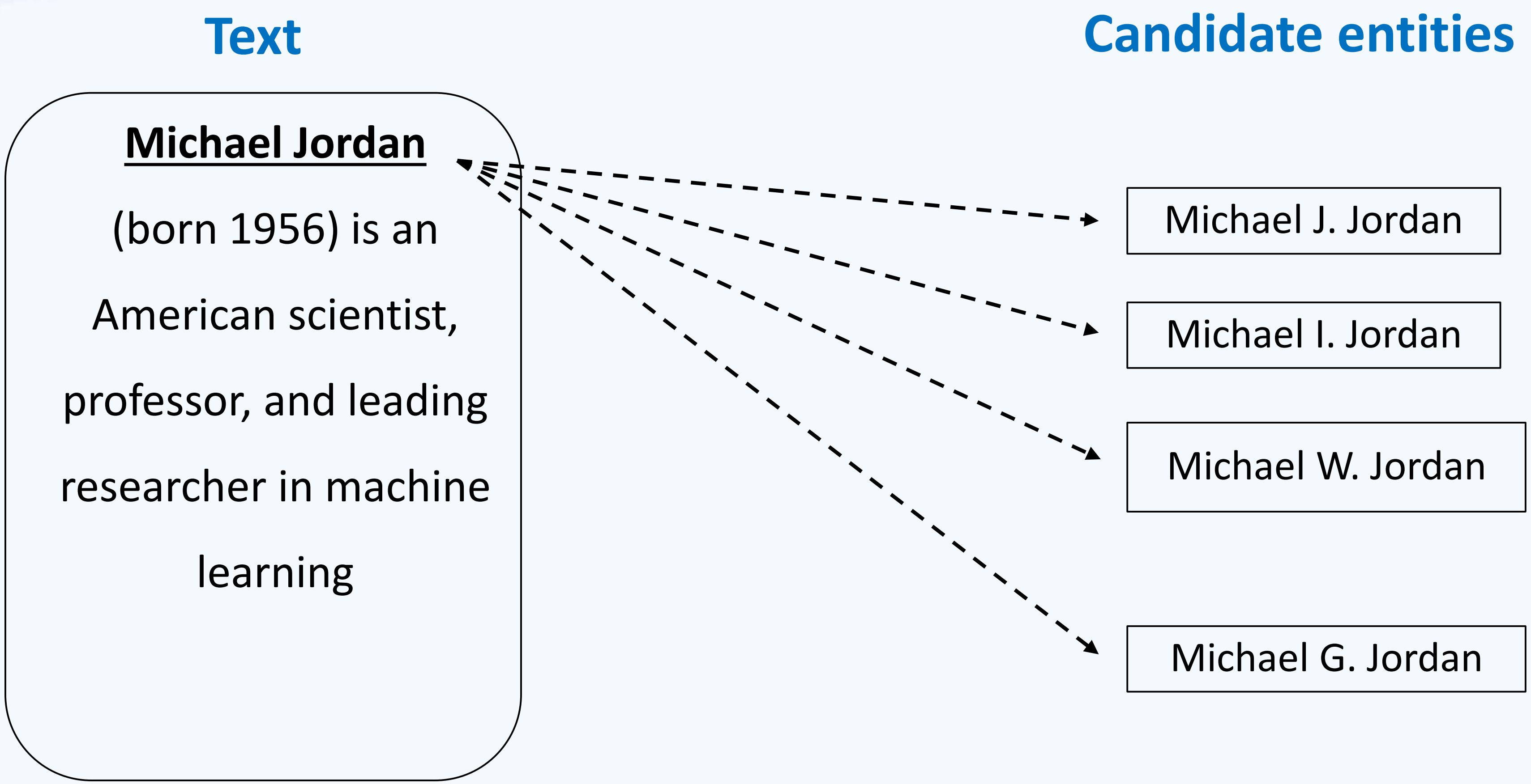
- Machine learning methods
    - Annotated corpora

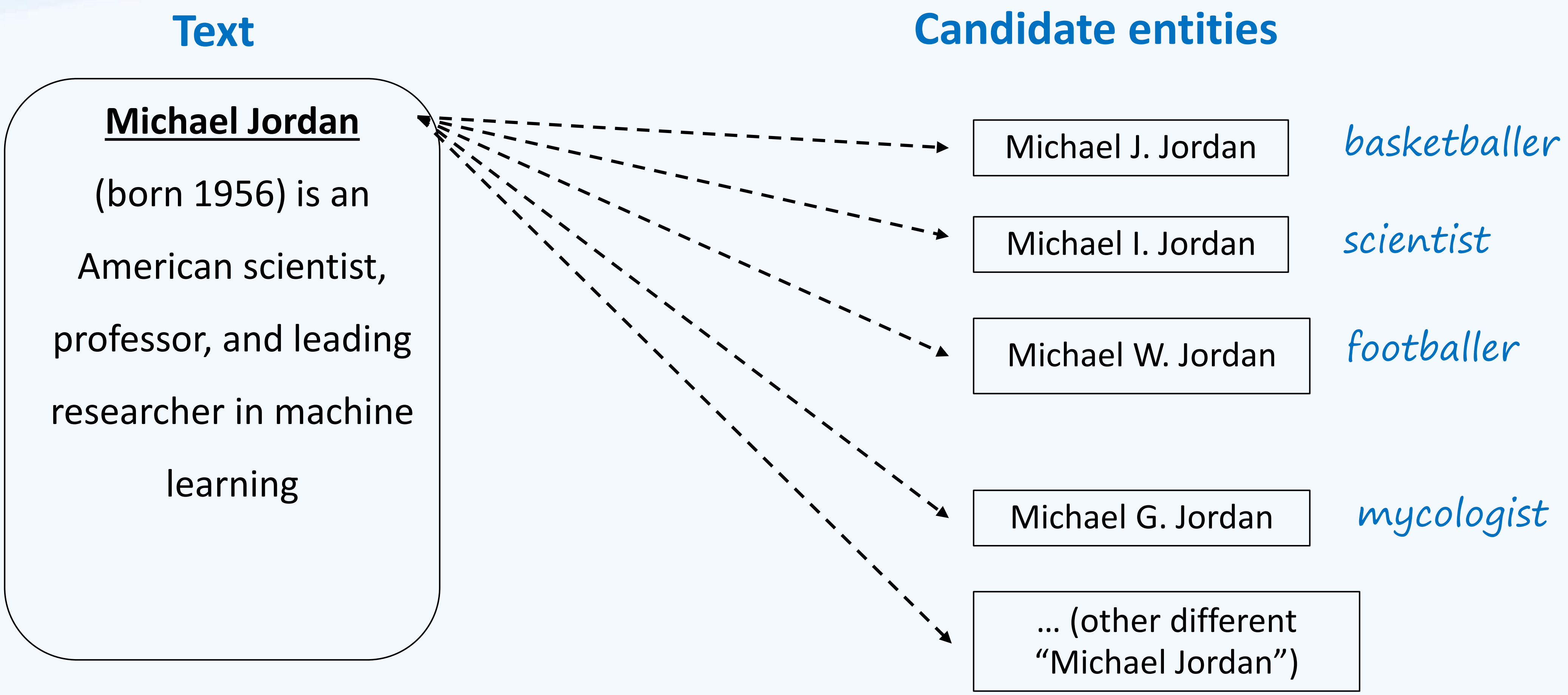
```
LONDON NNP I-NP I-LOC
1996-08-30 CD I-NP O

West NNP I-NP I-MISC
Indian NNP I-NP I-MISC
all-rounder NN I-NP O
Phil NNP I-NP I-PER
Simmons NNP I-NP I-PER
took VBD I-VP O
four CD I-NP O
for IN I-PP O
38 CD I-NP O
on IN I-PP O
Friday NNP I-NP O
as IN I-PP O
Leicestershire NNP I-NP I-ORG
```

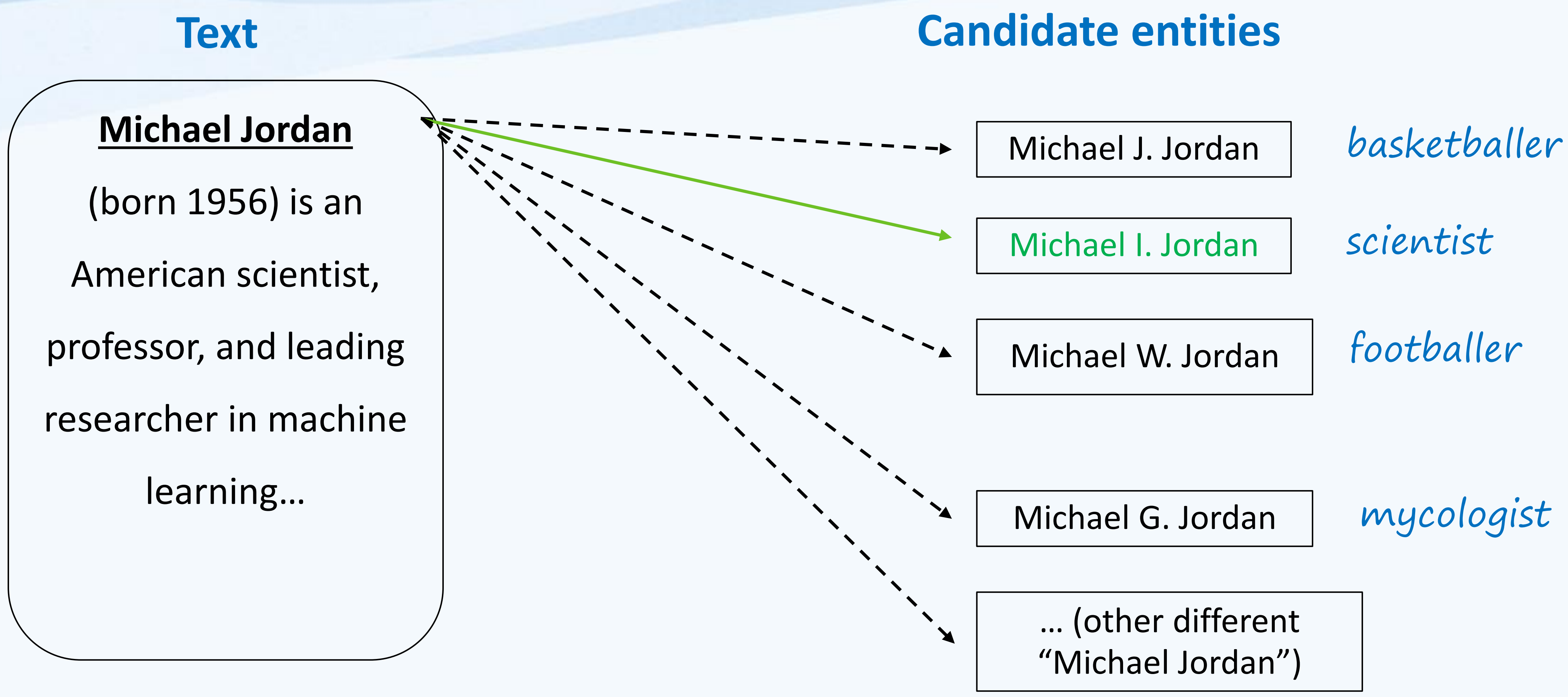
NER systems Evaluation

Model	F1	Paper / Source	Code
LUKE (Yamada et al., 2020)	94.3	LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention	Official
CNN Large + fine-tune (Baevski et al., 2019)	93.5	Cloze-driven Pretraining of Self-attention Networks	
RNN-CRF+Flair	93.47	Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition	
CrossWeigh + Flair (Wang et al., 2019)♦	93.43	CrossWeigh: Training Named Entity Tagger from Imperfect Annotations	Official
LSTM-CRF+ELMo+BERT+Flair	93.38	Neural Architectures for Nested NER through Linearization	Official
Flair embeddings (Akbik et al., 2018)♦	93.09	Contextual String Embeddings for Sequence Labeling	Flair framework
BERT Large (Devlin et al., 2018)	92.8	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	

Named entity Linking


Text
Candidate entities
Michael Jordan (born 1956) is an American scientist, professor, and leading researcher in machine learning
Michael J. Jordan
Michael I. Jordan
Michael W. Jordan
Michael G. Jordan


Text
Candidate entities
Michael Jordan
(born 1956) is an American scientist, professor, and leading researcher in machine learning
Michael J. Jordan
basketballer
Michael I. Jordan
scientist
Michael W. Jordan
footballer
Michael G. Jordan
mycologist
… (other different "Michael Jordan")



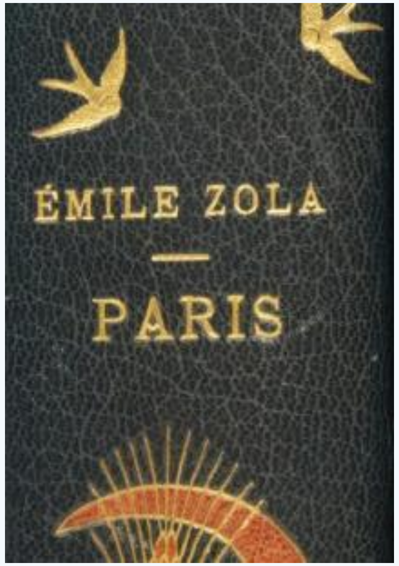
https://en.wikipedia.org/wiki/Michael_I._Jordan

Named entity linking (NEL)

- The task cinsisting in identifing entities and linking them to a knowledge base (such as Wikipedia)

List of cities called **Paris**

1. France
2. Danemark
3. United States
4. Canada
5. Panama
6. Gabon
7. Russia



Disambiguation

"Paris is the capital of France."

fr.wikipedia.org/wiki/Paris

fr.wikipedia.org/wiki/France

Knowledge Bases

"Paris est the capital of France."

- How to choose the right answer?
- There is about ten cities with the name Paris (persons or ships' names as well)
- Knowledge bases:
  - Rich in information about entities
  - Examples : Wikidata, Wikipedia, DBpedia, ...

Paris (Q90)

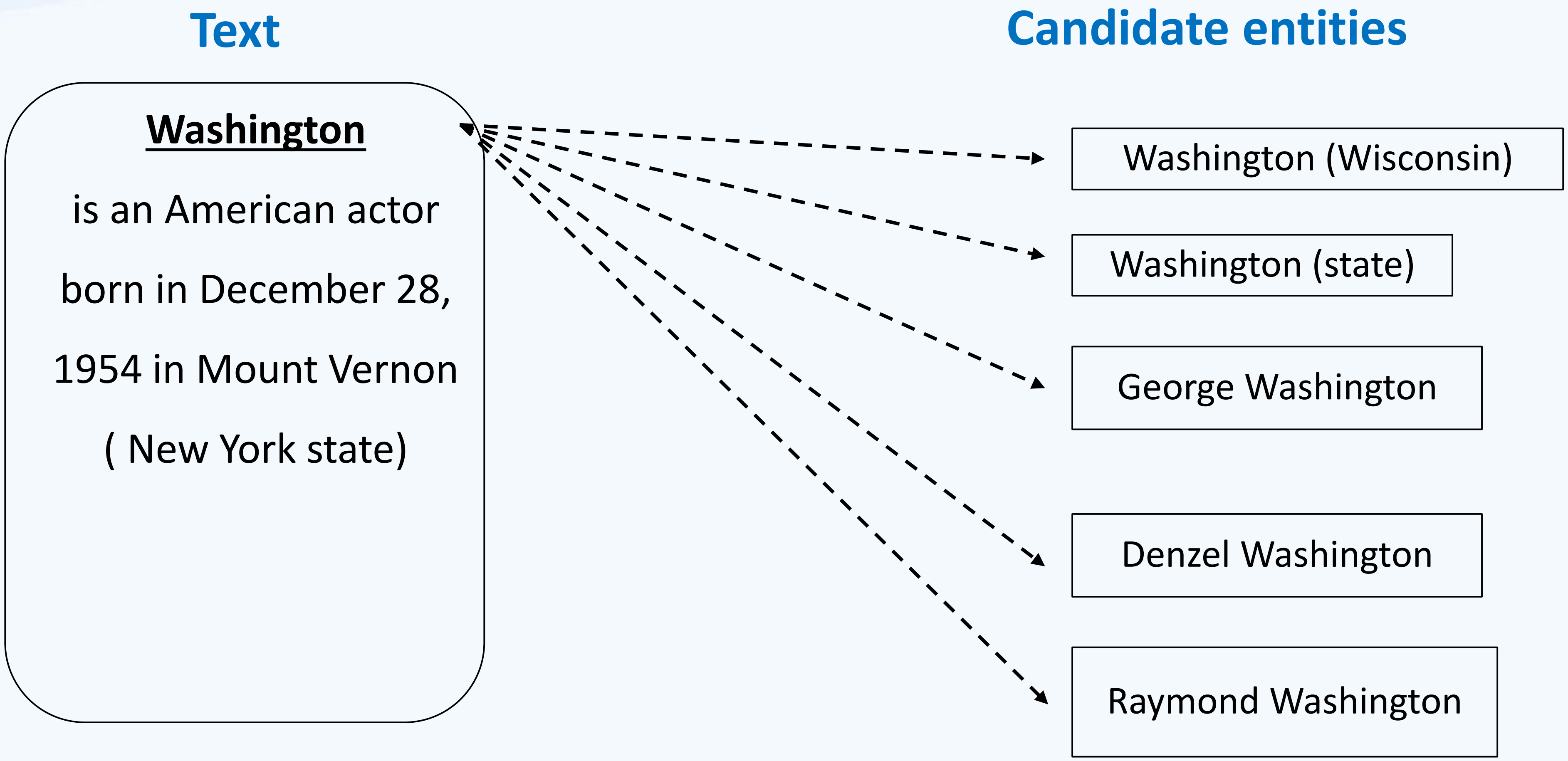
capital and largest city of France

City of Light | Paris, France

In more languages

Configure

Language	Label	Description	Also known as
English	Paris	capital and largest city of France	City of Light Paris, France
French	Paris	capitale de la France	Ville-Lumière Paname 75 Lutèce Ville de l'Amour 7.5
Spanish	París	capital de Francia	La Ciudad de la Luz Paris La Ciudad Luz
German	Paris	Hauptstadt von Frankreich	


Text
Candidate entities
Washington
is an American actor
born in December 28,
1954 in Mount Vernon
(New York state)
Washington (Wisconsin)
Washington (state)
George Washington
Denzel Washington
Raymond Washington

Disambiguation

Floyd revolutionized rock with the Wall

.../wiki/Pink_Floyd
.../wiki/Floyd_(name)
.../wiki/Pink_Iowa

.../wiki/Rock_(geology)
.../wiki/The_Rock
.../wiki/Musique_Rock

.../wiki/Berlin_Wall
.../wiki/The_Wall_(album)
.../wiki/Defensive_Wall

Challenge

../wiki/England
../wiki/England_football_team
../wiki/England_football

../wiki/football_world_cup
../wiki/rugby_world_cup
../wiki/world_cup

../wiki/England
../wiki/England_football_team
../wiki/England_football

The 1966 world Cup was held in England ... England won ...

In the final, England beat West Germany.

../wiki/England
../wiki/England_football_team
../wiki/England_football

../wiki/West_Germany
../wiki/German_Cup_of_football
../wiki/German_football_team

Challenge for artificial intelligence *(sometimes for human also)*

- Non-existant entities in the knowledge base
  - New companies
  - Little known people
- Lack of context

"Paris is beautiful"

- Paris city (France)
- Actor Paris Hilton
- City of Tennessee (USA)
- Prince of Troy (Greek mythology)
- Genus of plant Liliaceae
- City in Bourbon County, Kentucky (USA)
- City (Illinois, USA)
- ...

NEL systems Evaluation

Disambiguation-Only Models

Paper / Source	Micro-Precision	Macro-Precision	Paper / Source	Code
Mulang' et al. (2020)	94.94	-	Evaluating the Impact of Knowledge Graph Context on Entity Disambiguation Models	-
Raiman et al. (2018)	94.88	-	DeepType: Multilingual Entity Linking by Neural Type System Evolution	Official
Sil et al. (2018)	94.0	-	Neural Cross-Lingual Entity Linking	
Radhakrishnan et al. (2018)	93.0	93.7	ELDEN: Improved Entity Linking using Densified Knowledge Graphs	

End-to-End Models

Paper / Source	Micro-F1-strong	Macro-F1-strong	Paper / Source	Code
Kolitsas et al. (2018)	82.6	**82.4**	End-to-End Neural Entity Linking	Official
van Hulst et al. (2020)	**83.3**	81.3	REL: An Entity Linker Standing on the Shoulders of Giants	Official
Piccinno et al. (2014)	70.8	73.0	From TagME to WAT: a new entity annotator	
Hoffart et al. (2011)	71.9	72.8	Robust Disambiguation of Named Entities in Text	

References and important links

- OnTour ontology: https://hal.archives-ouvertes.fr/hal-02131145/document
- Eiffel project by Mondeca: https://mondeca.com/
- SEED (Semantic E-Tourism Dynamic packaging):
  https://jorge-cardoso.github.io/publications/Papers/OT-006-2006-R&D-Project-Report-SEED.pdf
- CRUZAR W3C usecase of Zaracoza:
  https://www.w3.org/2001/sw/sweo/public/UseCases/Zaragoza-2

Thank you!

Questions?

Stance Detection for Tourism

Ahmed Hamdi

University of La Rochelle
Laboratoire Informatique, Image, Interaction (L3i)

March 12, 2021

Interventions (CET)

- Mars 5, 2021: Information Extraction based on Named Entities for Tourism
  - 1.15 p.m – 2.45 p.m : **course**
  - 3.00 p.m – 4.30 p.m : **practice work**

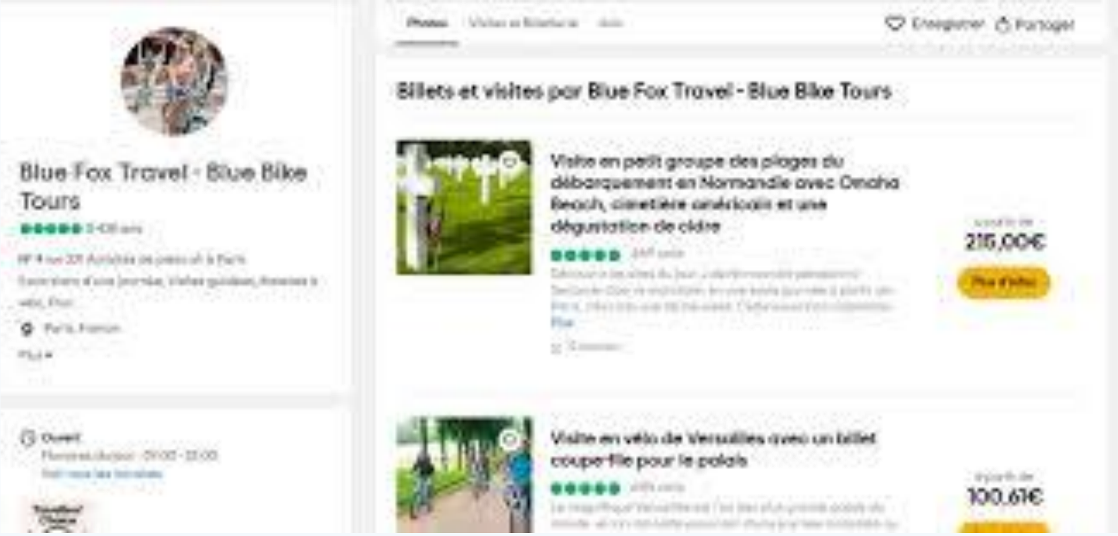
- **Mars 12, 2021: Stance Detection for Tourism**
  - 1.15 p.m – 2.45 p.m : **course**
  - 3.00 p.m – 4.30 p.m : **practice work**

Overview

- Context
- Stance detection
- Text classification
- Text embedding
- Application to stance detection for tourism

Context

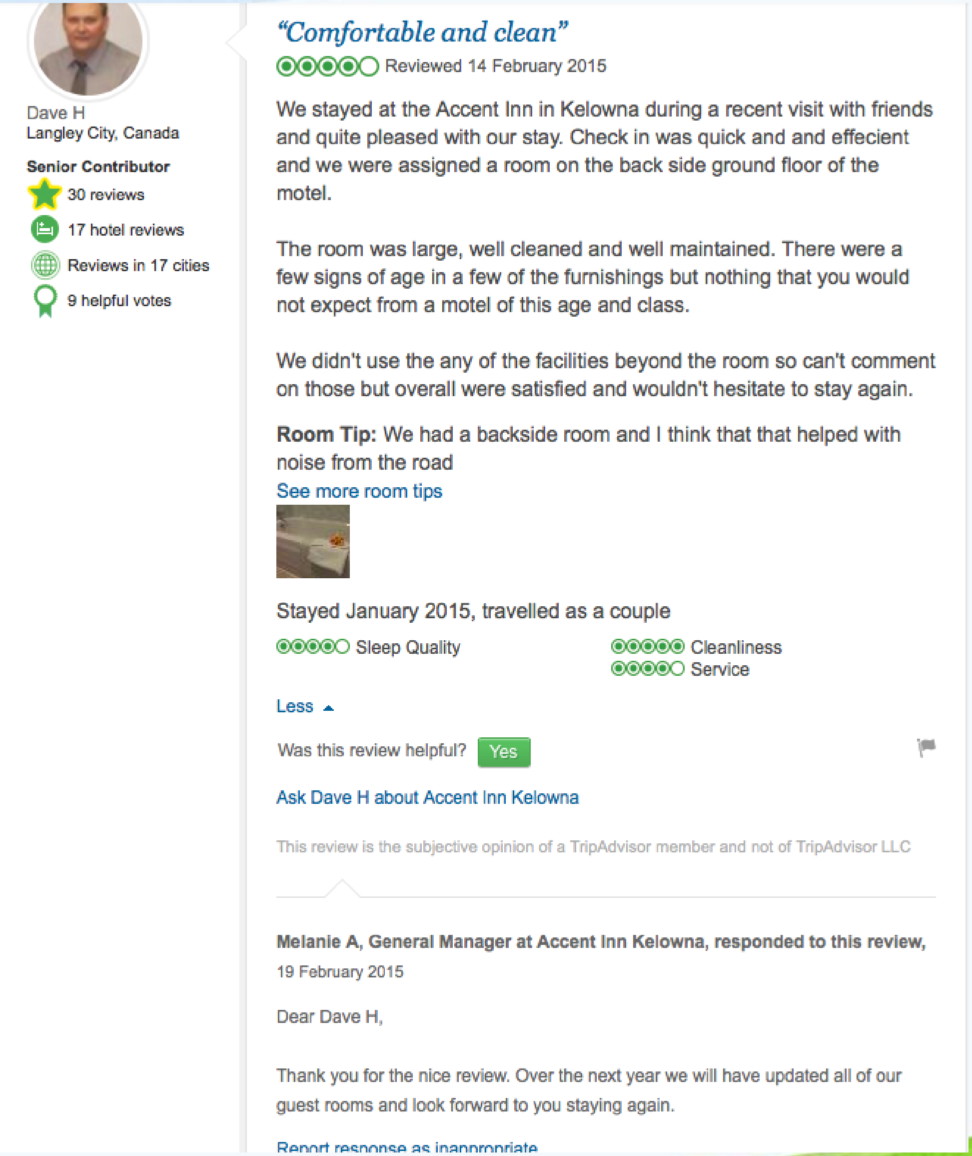
- Lot of tourists' reviews are available
- Online reviews remain a trusted source of information



Context

Questionnaire survey among tourists

- **Service quality**
- **Hospitality**
- **Safety**
- **Price**
- **Location and Closeness**
- **Comfort**
- **(beautiful nature, historic sites)**



Goal

- Automatically extract stances (opinions), emotions from reviews
- Tracking attitudes and feelings from reviews and comments about all sorts of tourism
- Determining whether they are viewed positively or negatively

Stance detection

Stance detection

- The task consisting in determining form pieces of texts the authors' opinions towards a topic

  - Negative ☹ : I do not recommend this hotel
  - Positive ☺ : the staff was friendly
  - Neutral 😐 : it could be better

Subjective stances

Exceptionnel

· Breath taking views and beautiful facilities in the room, very clean and well update in terms of how luxury and modern it looks

· I loved everything not one bad thing to say

Terrible Do not stay!!

· There is Nothing I like about this hotel! This Hotel is awful they have charged me an extra £108 for a cancellation I haven't even made The staff wrongfully advise me and speak very rude to you I would not stay in this property again absolutely terrible service

· Reception Staff are very rude And very unhelpful and advise you wrongly

Subjective detection

· The staff was absolutely charming and helpful, we got all possible attention and support. Breakfast was also very nice, especially taking into account current situation. And of course the location - the very heart of the city, walking distance from Eiffel Tower, with lots of nice little restaurants around.

· The room was very small, old, dark and it smelled paint terribly. One could see the Eiffel Tower, but only a very little bit, over a narrow street. I had somewhat greater impression from the advertised pictures...

Why stance detection?

- It allows business to track:
  - Flame detection
  - New service perception
  - Reputation management

- It allows individuals to get:
  - A global opinion on something

Opinions vs Facts

- The task allows analyzing
  - **Opinion**: personal belief or judgment that is not founded on proof or certainty.

    I like this hotel

  - **Fact**: statement that can be verified or proved to be true. It almost relies on observations and describes an objective reality.

    The booking is more expensive than usual

Stance detection for machines

The problem has several dimensions:

1. How does a machine define objectivity and subjectivity?
2. How does a machine analyze polarity?
3. How does a machine deal with word senses?
4. How does a machine assign an opinion rating?
5. How does a machine know about feeling intensity?

What is a stance to a machine

It is a quintuple (o, f, s, h, t)

1. o — the thing in question (i.e. hotel) → named entity recognition
2. f — features extracted from the text → information extraction
3. s — stance value → classification
4. h — stance holder → information extraction
5. t — time when opinion is expressed → data analysis

Text classification

Text classification

- Stance detection can be seen as a text classification problem
- Assigning a class (neutral, positive, negative) to a piece of text

Examples:

☺ *I liked the room. It was comfortable!*

☹ *I do not like the room. It is not clean!*

Preprocessing

- **Stop words:** irrelevant words (the, a, from, of…) should be removed from the text being analyzed
  - ☺ *I liked room. It comfortable!*
  - ☹ *I don't like room. It not clean!*
- **Tokenization**: splits the text into very simple tokens such as words, numbers, punctuation marks.
  - ☺ *I liked room . It comfortable !*
  - ☹ *I do not like room . It not clean !*
- **Stemming**: produces a stem for each word in the text
  - ☺ *I like room . It comfortable !*
  - ☹ *I do not like room. It not clean !*

Approaches

1. Feature-based
2. Sentiment-based
3. Machine learning-based

Feature-based

- Word polarity:
  - Positive: good, like, nice…
  - Negative: bad, dislike…
- Emoticons:
  - Positive: 🙂 🤩 😍 🤗
  - Negative: 😞 😔 😠 😈
- Uppercases
- Rating

Lexicon-based

- WordNet is a lexical database for the English language that groups English word into set of synonyms called SynSet
- WordNet distinguishes between:
  - Nouns
  - Verbs
  - Adjectives
  - adverbs

Word Sense Disambiguation (WSD)

The techniques of WSD aim to determine the meaning of each word in its context

In this case, the disambiguation happens selecting for each words in a comment the SynSet in wordnet that best represents the word in its context

Sentiment-based

- SentiWordNet is an extension of WordNet that adds to each SynSet 3 scores between 0 and 1:
  - PosScore: positivity measure
  - NegScore: negativity measure
  - ObjScore: objective measure

PosScore + NegScore + ObjScore = 1

```
# Positive-Score <tab> Negative-Score <tab> Synset
1       0       true#a#2 real#a#4
1       0       illustrious#a#1 famous#a#1 far-famed#a#1 noted#a#1 celebrated#a#1 notable#a#2 renowned#a#1 famed#a#1
0.5     0       real#a#6 tangible#a#2
0.25    0       existent#a#2 real#a#1
```

Sentiment-based

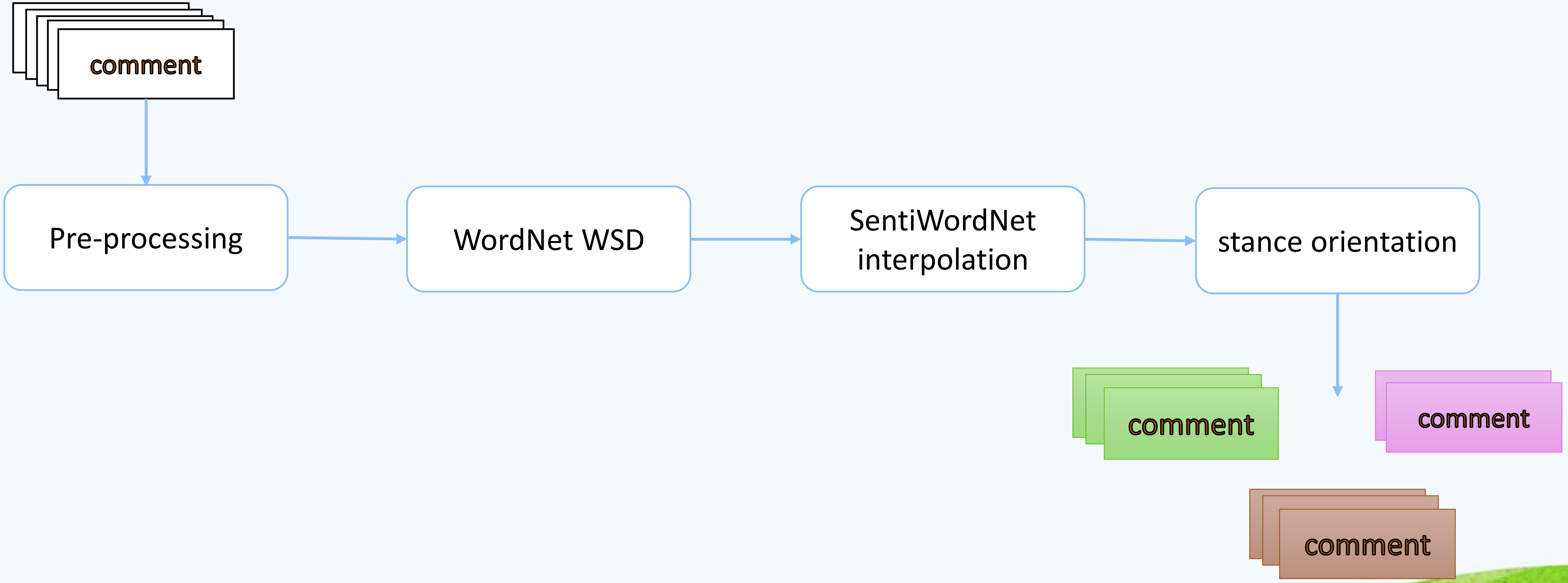
- Given a SynSet, we can search in SentiWordNet, the scores associated to this SynSet

This is very **accurate**. Well done ☺

accurate: conforming exactly or almost exactly to fact or to a standard or performing with total accuracy

posScore	NegScore	ObjScore
0.5	0	0.5

Sentiment-based



Sentiment-based

- **Sum :**

  The positive and negative scores for each term found in a comment are summed separately to get the positive and negative scores

$$s_+ = \sum_{i \in t} pos\_score_i$$

$$s_- = \sum_{i \in t} neg\_score_i$$

Sentiment-based

- **Average :**

  The positive and negative scores for each comment are determined by calculating the average of positive and negative scores

$$s_+ = \frac{\sum_{i \in t} pos\_score_i}{n}$$

$$s_- = \frac{\sum_{i \in t} neg\_score_i}{n}$$

Sentiment-based

- **Average with threshold on objective score:**
  - The word with objective score < of a given threshold is discarded
  - Positive and negative scores for each comment are determined by calculating the average of positive and negative scores of all the words that are not been discarded

$$s_+ = \frac{\sum_{\substack{i \in t \\ obj\_score_i < \theta}} pos\_score_i}{n}$$

$$s_- = \frac{\sum_{\substack{i \in t \\ obj\_score_i < \theta}} neg\_score_i}{n}$$

Classification

- The stance is determined based on the higher value between S+ and S-

$$s_t = \begin{cases} positive & \text{if } s_+ > s_- \\ negative & \text{if } s_+ \leq s_- \end{cases}$$

Classification

- The average stance orientation of all the comments we gathered is computed
- This allows the machine to say something like:
  - Generally people like the hotel

    → they recommend it

  - Generally people dislike the hotel

    → they do not recommend it

Machine learning-based

- Large annotated corpora
- Train the machine to predict classes to unseen comments

→ How to represent the text to the machine?

Text embedding

Text embedding

Word embedding

- Represent each word from a vocabulary by a vector of real numbers

home →	0.01	-0.05	0.58	0.19	0.98	0.67
dog →	-0.01	-0.33	0.09	0.27	0.77	0.15
cat →	0.08	0.87	-0.55	0.99	-0.91	0.04
kitten →	0.24	-0.22	-0.58	-0.64	0.48	-0.36

From word embedding to sentence embedding

- Represent each sentence by a vector of real numbers
- Use the word vectors to calculate the sentence vector
  - Sum of words' vectors
  - Average of words' vectors

→ Classification basing on all the words

→ Taking into account the meaning of words

Thank you!

Questions?