Information Extraction

Named Entity Recognition, Relation Extraction

Gustavo Gonçalves, Research Assistant @ LTI/CMU and NOVA LINCS
ggoncalv@xcs.cmu.edux - remove x chars

About me

- I’m a Research Assistant at the Web and Media Search Lab @ NOVA LINCS, and at the Language and Technology Institute @ Carnegie Mellon University. I’m also pursuing my Dual PhD Degree. Check the CMU Portugal Program if you need funding

- I’ve been doing research since 2017, and my PhD topic is centered on improving conversational search sessions with knowledge-aware representations.

- My research interests include: Conversational Search; Knowledge-Representations; Machine Learning; and Socia-Media Analysis.

Lecture Outline

- Knowledge Graphs;
- What is Information Extraction?;
- Named Entity Recognition;
- Entity Linking;
- Knowledge Driven Information Retrieval;

The story so far

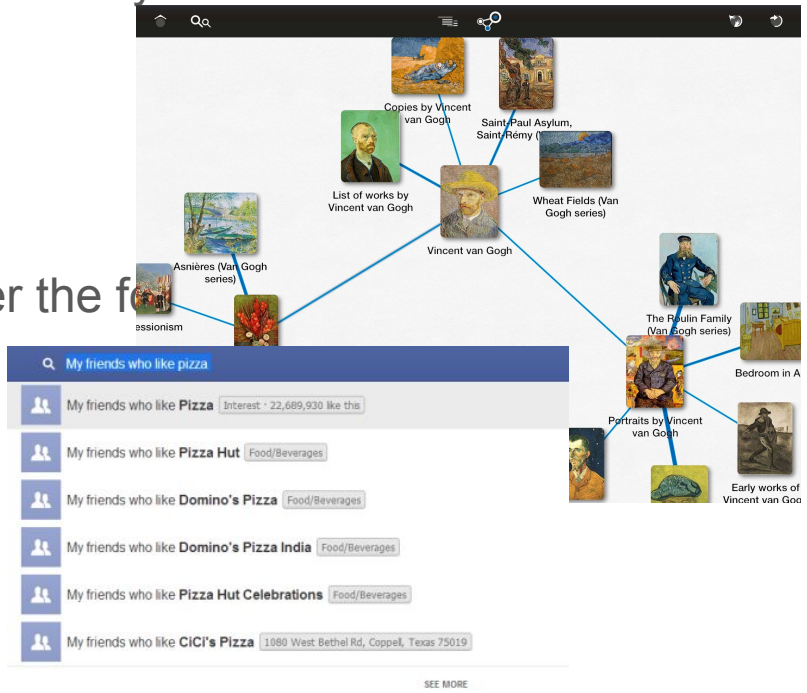
- Relevance based models
  - Based on the query-related documents (initial search results)

- Statistical correlations
  - *Term correlations across documents*
    - *(we will revisit this in the recommendation lecture)*
  - Term correlations across term's neighborhood (word embeddings)

- Knowledge-bases expansion
  - Linguistic thesaurus: e.g. MedLine: physician, syn: doc, doctor, MD, medico
  - Can be query rather than just synonyms

What are Knowledge Bases?

- Knowledge bases are graphs that have a manually curated source.
  - The most famous example probably is Wikipedia!
- Graph nodes are entities, and edges are relationships among entities;
- These relationships are usually stored under the f
  of triples
  - Example: Painted(Van Gogh, The Starry Night)



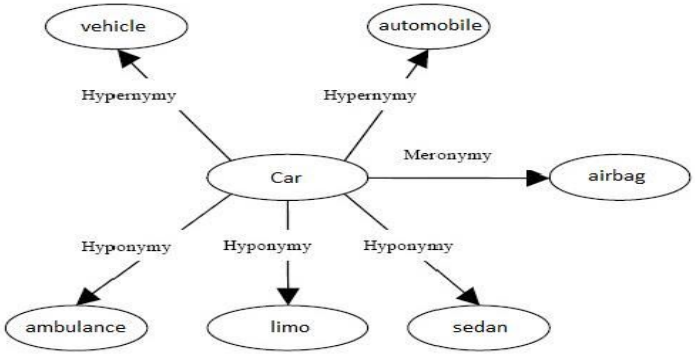
https://en.wikipedia.org/wiki/Facebook_Graph_Search

https://www.designnominees.com/apps/learn-discovery-mindmap-of-wikipedia

WordNet: A lexical database

"WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

Synsets are interlinked by means of conceptual-semantic and lexical relations. "



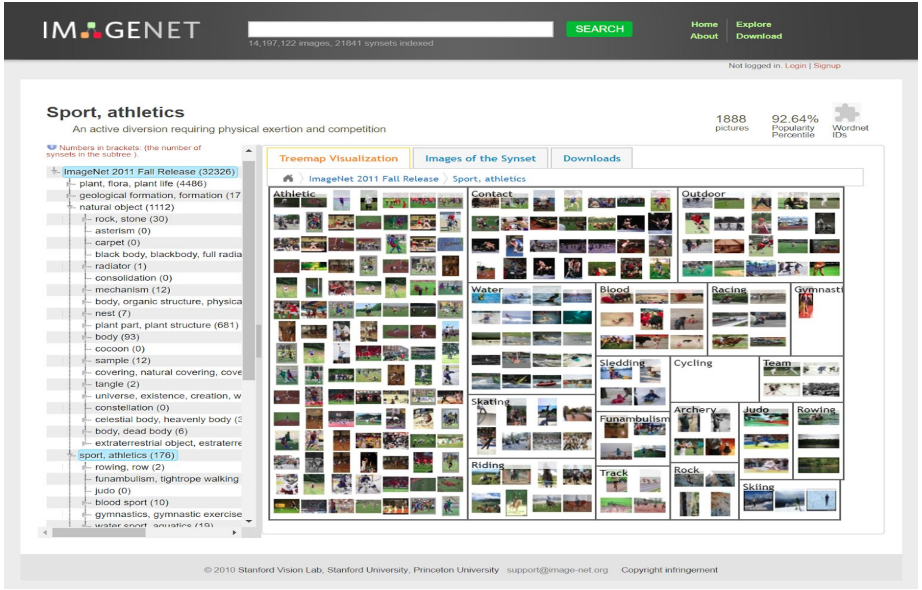
https://wordnet.princeton.edu/

"WordNet interlinks specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity."

ImageNet: A visual taxonomy

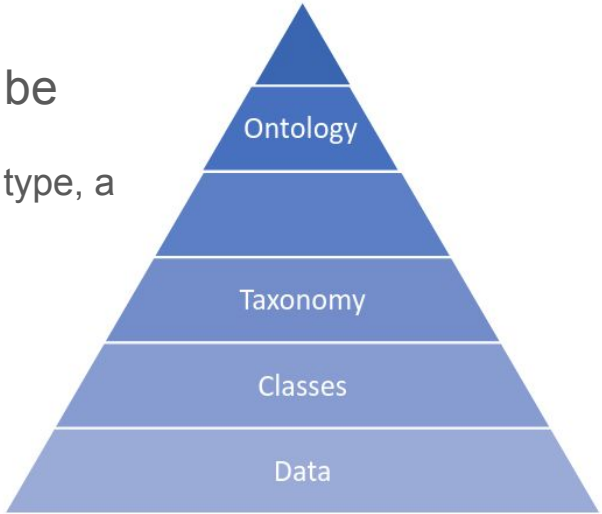
- Selected words of WordNet are illustrated in ImageNet.
- Currently, there are over 14.000 concepts illustrated.
- Roughly 1.000 concepts are used by VOC.
- Great impact in advancing the state of the art.

http://image-net.org/explore.php



From data to information

- A taxonomy is concerned with classifying and organizing hierarchically concepts of a specific domain.
- It is important to identify the list of items that need to be detected.
  - These items are domain specific, and can be a topic, a scene type, a visual object or a named entity.
  - They are normally associated to a class in a supervised learning task.



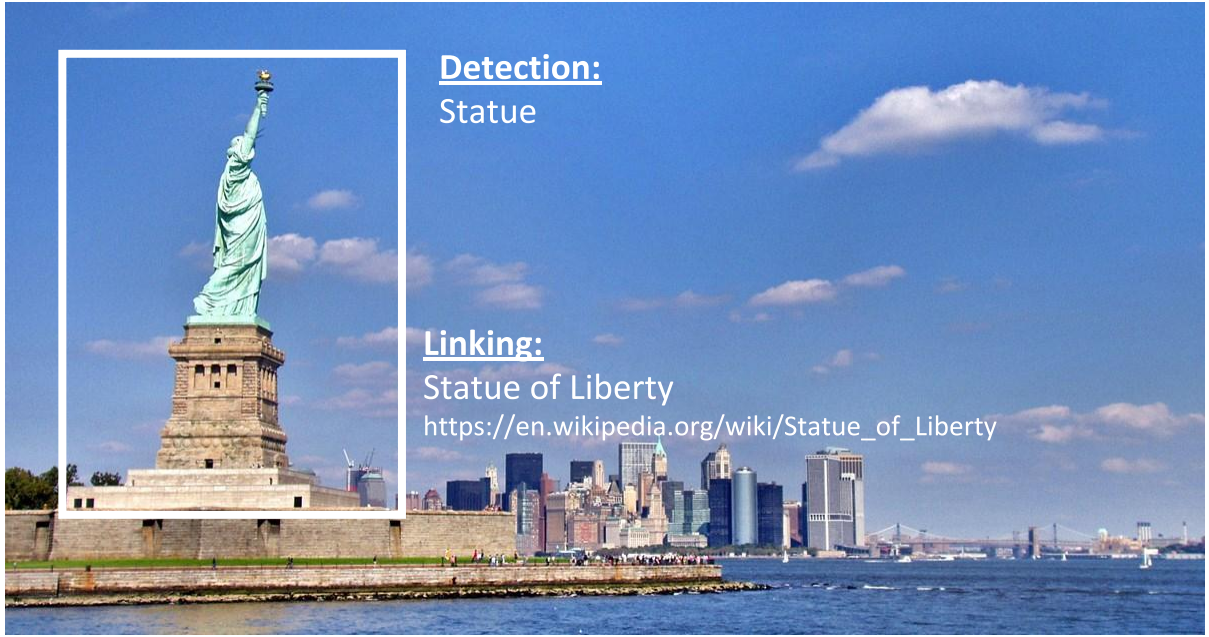
What is Information Extraction (IE)?

- We want to work with Things, not Strings!
- IE systems extract factual, and clear data;
  - What? Where? Who? When?
- We can roughly think of IE systems as entity recognition and relation extraction machines
  - For Example:

> **Tesla, Inc.** (formerly **Tesla Motors, Inc.**) is an American electric vehicle and clean energy company based in Palo Alto, California.[9]
>
> Location(“Tesla, Inc.”, “Palo Alto, California”)

Slide adapted from: Christopher Manning, Information Extraction, Stanford University, USA

Classification, detection, linking

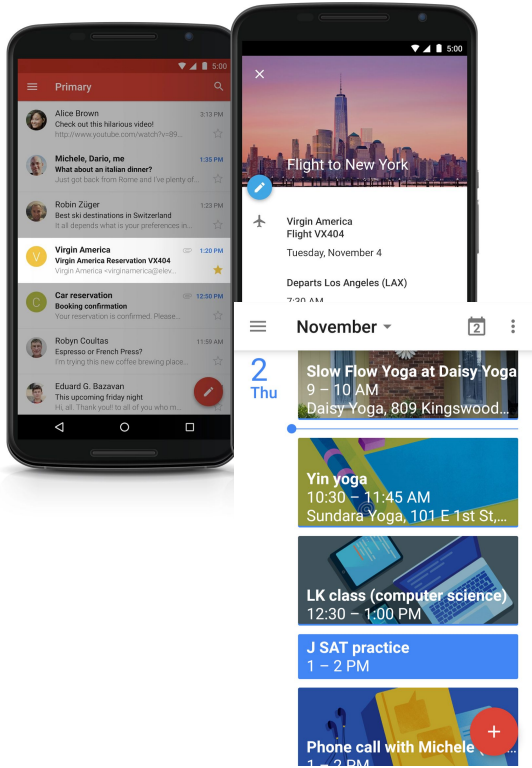


Classification:
Sea side
Statue
City
Sky

Linking:
New York City
https://en.wikipedia.org/wiki/New_York_City

You use IE every day!

- Rule-based and Machine Learning IE systems help us daily with convenient tasks;
- These tasks can go from low-level information extraction, such as with regular expressions;



Named Entity Recognition (NER)

- Is one of the sub-tasks in Information Extraction:
  Identify and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, aWer the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organization

- We've seen this kind of task in Machine Learning, it's a multi-label classification task. What are the features and evaluation metrics?

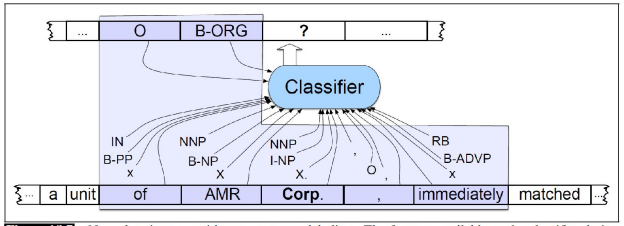


Figure 18.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

Slide from: Christopher Manning, Information Extraction, Stanford University, USA

Machine Learning approach to NER

Testing

1. Receive a set of testing documents
2. Run **sequence model** inference to label each token
3. Appropriately output the recognized entities

Training

1. Collect a set of representative training documents
2. Labe each token for its **entity class or other (O)**
3. Design feature extractors appropriate to the text and classes
4. Train a **sequence classifier** to predict the labels from the data

Slide from: Christopher Manning, Information Extraction, Stanford University, USA

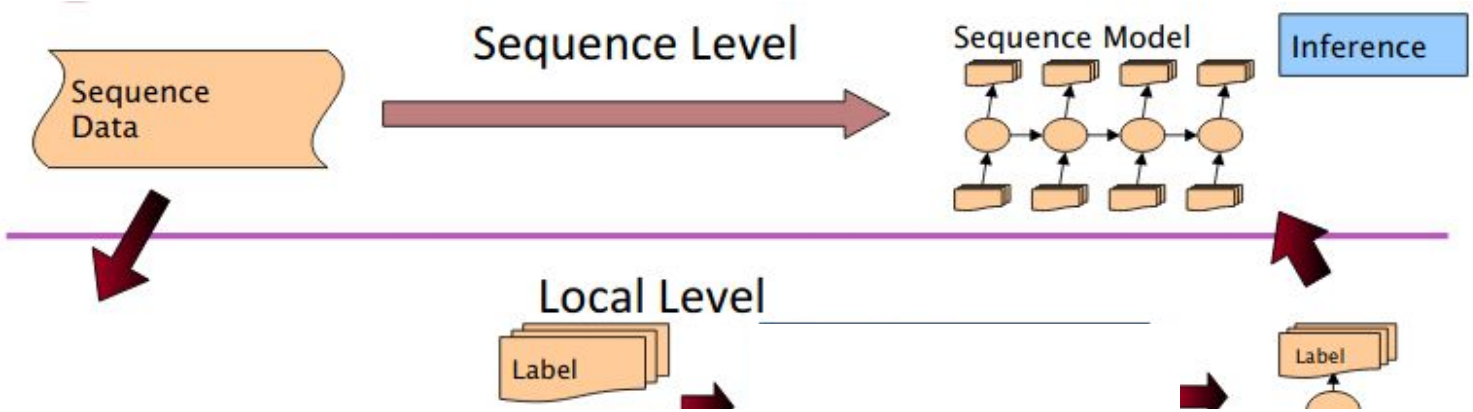
Language is Sequential, and so is IE

Test

Slide from: Christopher Manning, Information Extraction, Stanford University, USA

Language is Sequential, and so is IE

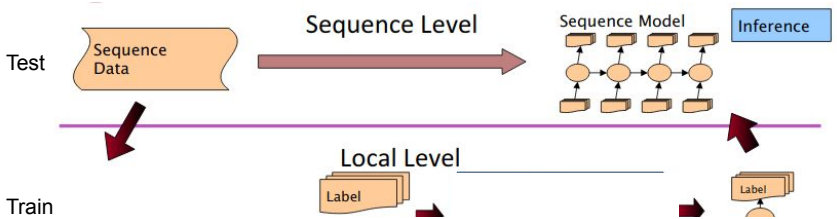
Test



Slide from: Christopher Manning, Information Extraction, Stanford University, USA

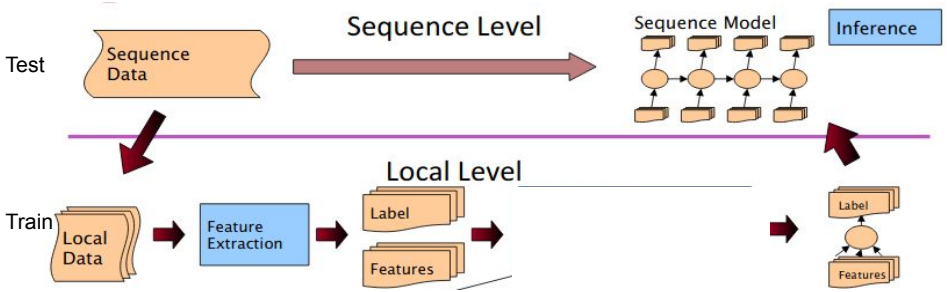
NER - Classes (Labels)

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O



Slide from: Christopher Manning, Information Extraction, Stanford University, USA

NER - Features



- Words
  - Current words
  - Previous/next word (context)
- Other kinds of inferred linguistic classification
  - Part-of-speech (POS) tags
- Label context
  - Previous (and perhaps next) label
- Word Substrings
  - Croatian surnames (many end with “ić”)
- Word shapes
  - Length, capitalization, numerals, greek letters, word punctuation
  - Examples: SARS-CoV-2 = XXXX-XxX-d / COVID-19 = XXXXX-dd

Slide adapted from: Christopher Manning, Information Extraction, Stanford University, USA

Important Sequence Modeling Problems in NLP

- Many problems in NLP have data which is a sequence of characters, words, phrases, lines, or sentences …
- We can think of our task as one of labeling each item

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

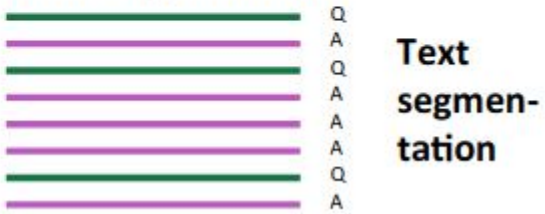
POS tagging

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

Named entity recognition

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

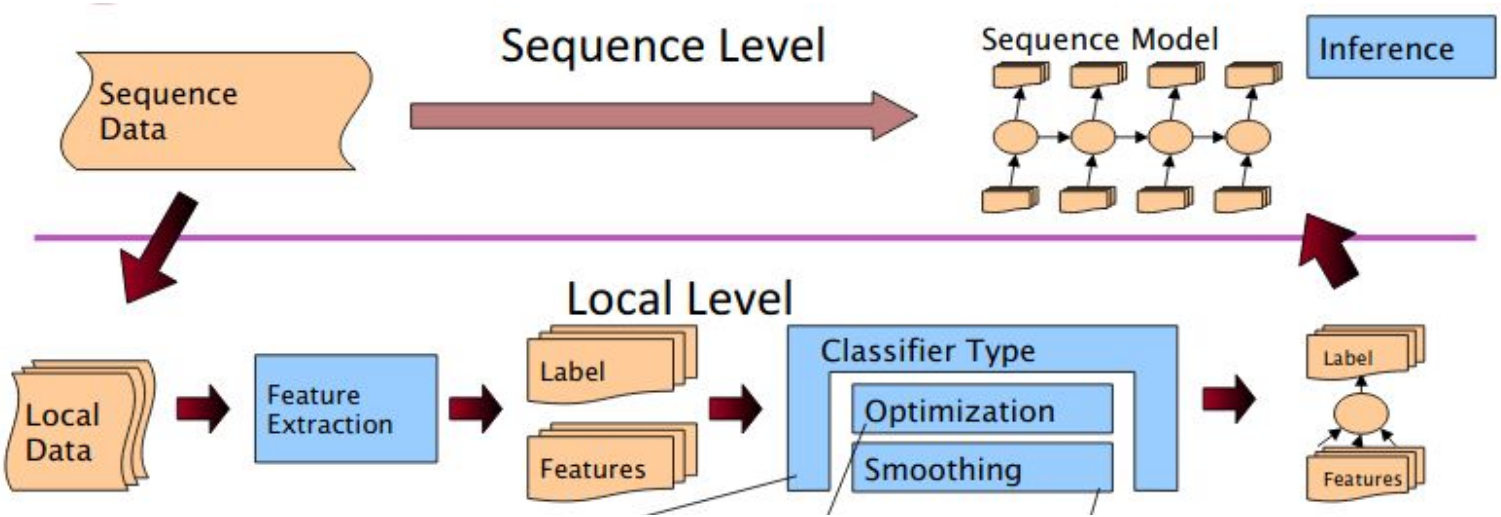
Word segmentation



Text segmen-tation

Slide from: Christopher Manning, Information Extraction, Stanford University, USA

Language is Sequential, and so is IE



Slide from: Christopher Manning, Information Extraction, Stanford University, USA

Entity Linking - What is an Entity?

- Depends on who you ask!
  - For simplicity let's follow Balog's[3] definition and consider Named Entities (Persons, Locations, Dates) Real-World objects and Concepts (Emotion, Gaussian Kernel, Peace) as abstract objects.
- Entity Linking is a special useful tool to inject knowledge into other representations. These representations can be raw count based, or vectorial (embeddings)
- What are the elements involved in a ML Entity Linker?
  - NER system to detect mentions;
  - Knowledge Base to relate candidate pairs of entitities (triples);
  - Feature Extractor (substrings, words, patterns);
  - Labeled Disambiguation Data (For supervised approaches);

Disambiguation - How many Ronaldo's do you know?

WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

Article Talk | Read View source View history | Search Wikipedia

Ronaldo

From Wikipedia, the free encyclopedia
(Redirected from Ronaldo (disambiguation))

For the documentary about the Portuguese footballer, see Ronaldo (film).

Ronaldo is a Portuguese[1] name equivalent to the Scottish Ronald. From the football super stars, it became a very common name in all Portuguese speaking countries, being also prevalent in Italy and Spanish speaking countries.

Contents [hide]

People

Notable people known as Ronaldo include:

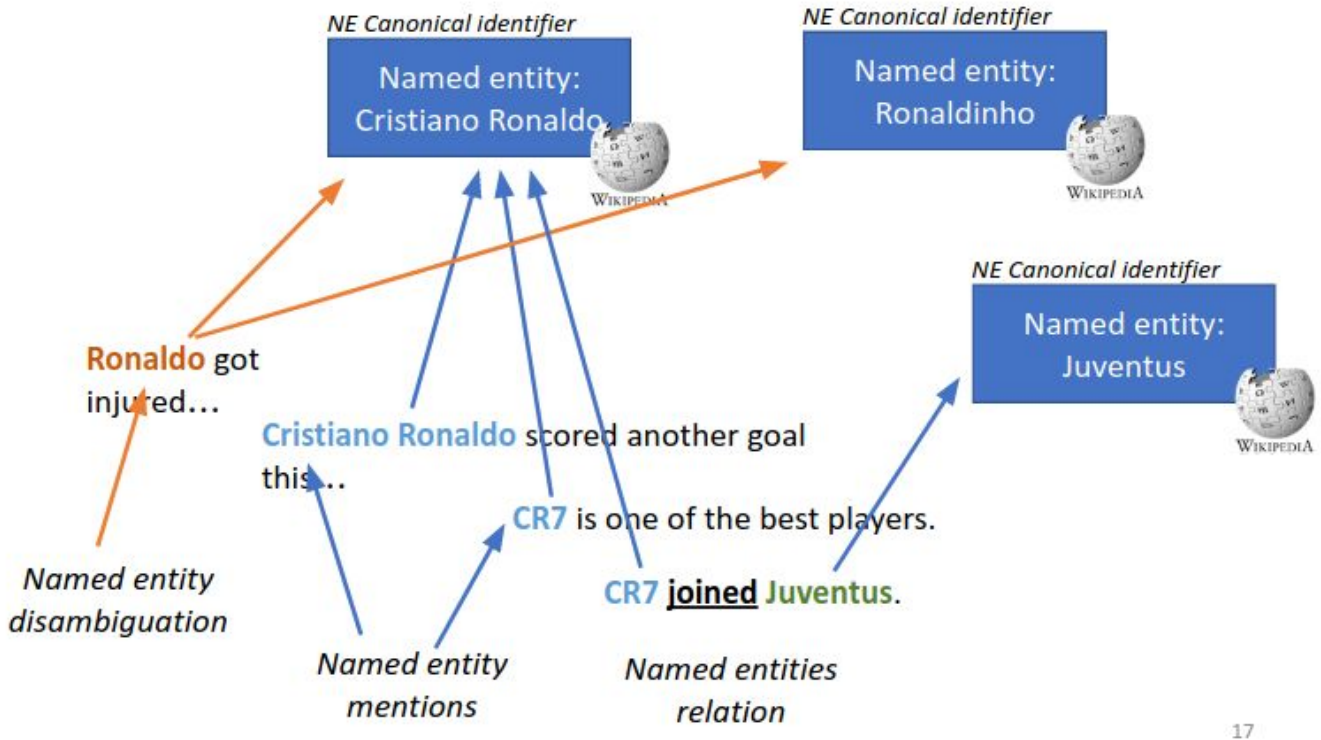
- Ronaldo (Brazilian footballer) (born 1976), Ronaldo Luís Nazário de Lima, was known as "Ronaldinho" in his early career to distinguish with Ronaldo Rodrigues de Jesus
- Cristiano Ronaldo (born 1985), Portuguese international footballer
- Ronaldinho, full name Ronaldo de Assis Moreira (born 1980), Brazilian international footballer, also known as "Ronaldinho Gaúcho"

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

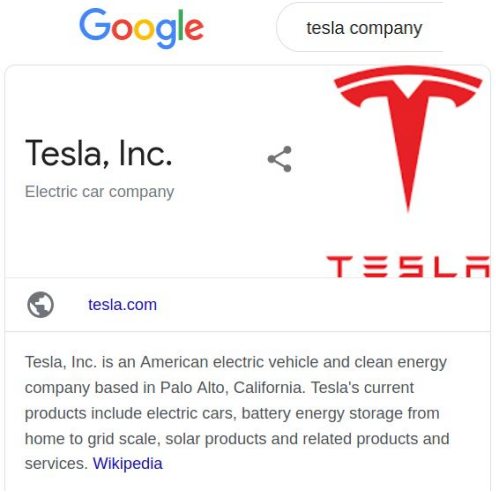
Tools
What links here
Related changes
Special pages
Permanent link
Page information

Entity Disambiguation Example



Knowledge Driven Information Retrieval

- Spoiler alert, all of you interact with entity retrieval daily.
- Entity cards are displayed by identifying and generating tabular and concise information about the entities contained in the query.
- The hypothesis behind this is that entities are important words that should be identified and treated a bit differently when possible. Users seem to enjoy it!



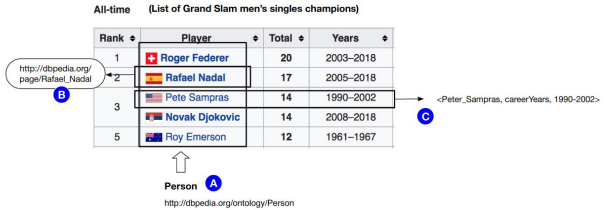


Figure 3: Illustration of table interpretation: (A) Column Type Identification. (B) Entity Linking. (C) Relation extraction.

Use case - Medical Search

- Domain specific terminologies are curated by domain experts and are designed with specific tasks and workflows in mind.

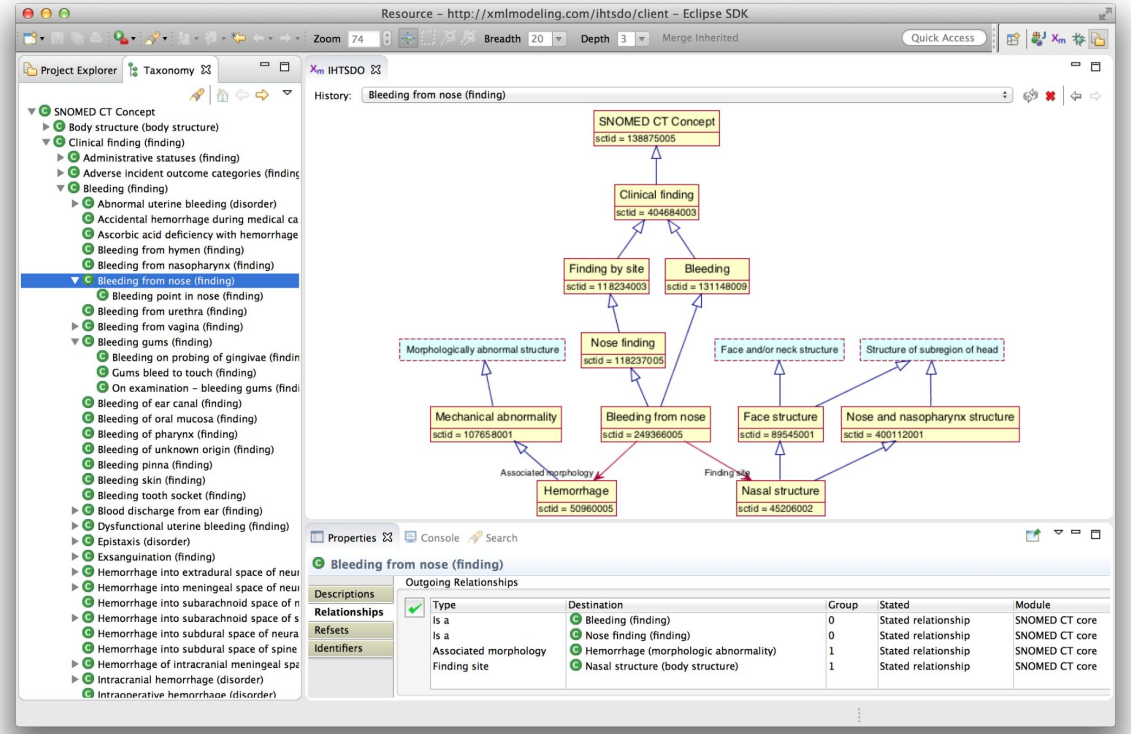
- In the medical domain, the SNOMED-CT is intended to describe medical conditions, procedures, admin, etc. http://browser.ihtsdotools.org/

Use case - Medical Search

- Domain specific terminologies are curated by domain experts and are designed with specific tasks and workflows in mind.

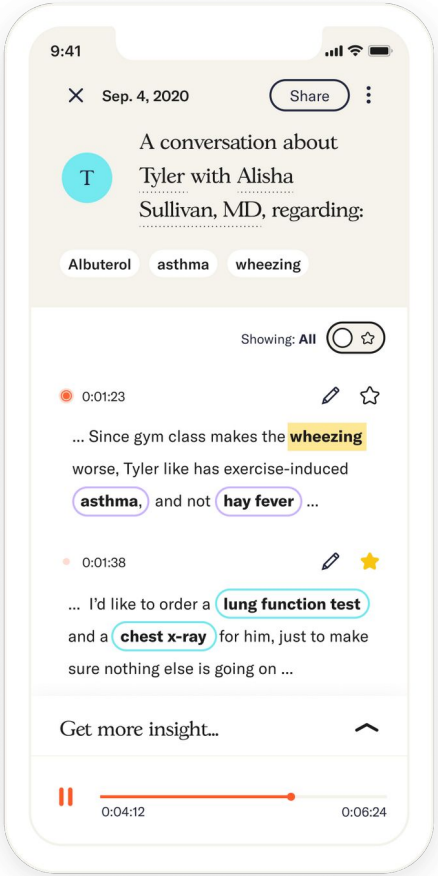
- In the medical domain, the SNOMED-CT is intended to describe medical conditions, procedures, admin, etc. http://browser.ihtsdotools.org/

Visualizing the Medical Taxonomy



Solving real world problems

- We are only getting to the good stuff!
- By detecting patterns and relating information we are able to make powerful inferences;
- This app is an example of product that uses IE and very possibly taxonomies to create a network of concepts that are discussed during a doctor's appointment.



https://www.abridge.com/

Summary

- In this class we discussed a complementary approach to NLP and IR using IE.
- You now have under your belt another tool that can be used in your work.
- This completes the third possible representation when working with NLP, the three representations that we have learned so far are:
  - Count Based Language Models (Using the Vector Space Model);
    - Coarse
    - Cheap
  - Co-ocurrence based language models (word embeddings);
    - Complete
    - Noisy
  - Manually curated knowledge bases (at the very least a controlled vocabulary);
    - Sparse
    - Reliable

Paper and Book References

[1] Daniel Jurafsky & James H. Martin. Speech and Language Processing. Information Extraction - Chapter 18. 2019
https://web.stanford.edu/~jurafsky/slp3/18.pdf

[2] Shuo Zhang and Krisztian Balog. Web Table Extraction, Retrieval, and Augmentation: A Survey. ACM Trans. Intell. Syst. Technol. 2020
https://arxiv.org/pdf/2002.00207.pdf

[3] Krisztian Balog. Entity Retrieval. Encyclopedia of Database Systems, Second Edition. 2018 https://link.springer.com/book/10.1007%2F978-3-319-93935-3