DBpedia Spotlight

Teaching material

Manuel Fiorelli

fiorelli@info.uniroma2.it

Named Entity Recognition

Quoting https://en.wikipedia.org/wiki/Named-entity_recognition:

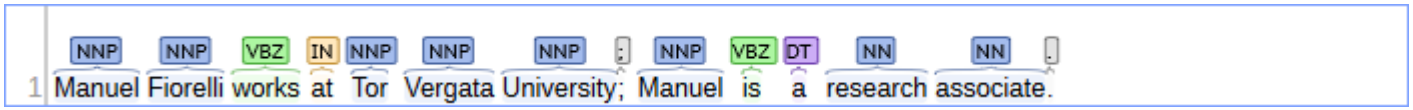
> In information extraction, a **named entity** is a real-world object, such as persons, locations, organizations, products, etc., that can be denoted with a proper name.

Quoting https://en.wikipedia.org/wiki/Named-entity_recognition

> **Named-entity recognition** (**NER**) (also known as **entity identification**, **entity chunking** and **entity extraction**) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc

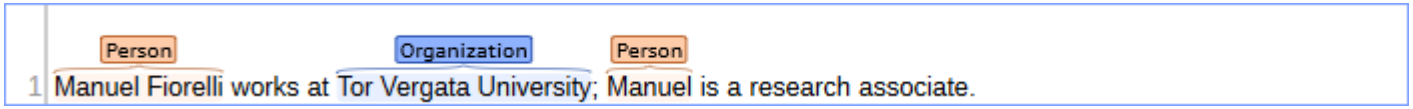
Named Entity Recognition (cont'd)

Part of speech (POS) tagging



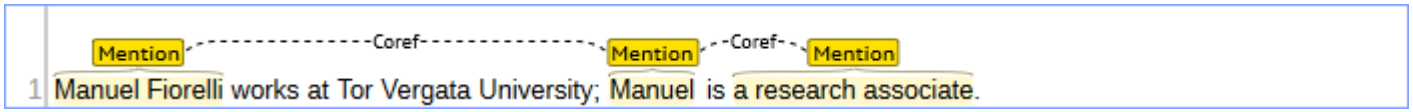
NNP = Proper Noun Singular
http://web.mit.edu/6.863/www/PennTreebankTags.html

Named Entity Recognition



Named Entities are first identified and then classified in a few categories (e.g. Person, Organization, Location)

Coreference resolution



Recognize mentions having the same referent

Images generated with http://nlp.stanford.edu:8080/corenlp/

Named Entity Disambiguation

Quoting https://en.wikipedia.org/wiki/Entity_linking:

> In natural language processing, **entity linking**, **named entity linking** (NEL),[1] **named entity disambiguation** (NED), **named entity recognition and disambiguation** (NERD) or **named entity normalization** (**NEN**)[2] is the task of determining the identity of entities mentioned in text. […] Entity linking requires a knowledge base containing the entities to which entity mentions can be linked.

DBpedia Spotlight

DBpedia Spotlight (http://www.dbpedia-spotlight.org/) annotates mentions of DBpedia resources in natural language texts.

Documentation:

https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

Online demo:

http://demo.dbpedia-spotlight.org/

Purpose of DBpedia Spotlight

The primary purpose of DBpedia Spotlight is to interlink unstructured content and DBpedia.

DBpedia (and the datasets it is linked to) provides background knowledge (e.g. types, relations between entities) supporting applications such as:

- Faceted document browsing
- Semantic search

Variants of DBpedia Spotlight

There exist two variants of DBpedia Spotlight:

- DBpedia-Spotlight-Lucene (based on *vector space model*)
  - https://github.com/dbpedia-spotlight/dbpedia-spotlight-lucene
  - Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems* (pp. 1-8). ACM. http://oa.upm.es/11477/2/INVE_MEM_2011_105377.pdf
- DBpedia-Spotlight-Model (based on *generative models*)
  - https://github.com/dbpedia-spotlight/dbpedia-spotlight-model
  - Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013, September). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems* (pp. 121-124). ACM. http://jodaiber.de/doc/entity.pdf

Variants of DBpedia Spotlight (cont'd)

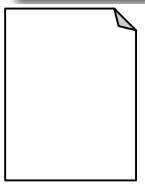
The two variants (*lucene* and *model*) share the:

- Application Programming Interface (API)
- Overall processing steps

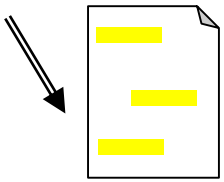
The key difference lies in how the same steps are implemented by each variant.

The *model* variant achieved improvements in accuracy, time and space requirements.

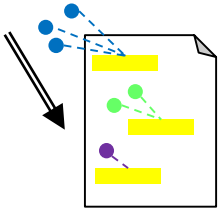
DBpedia Spotlight Flow



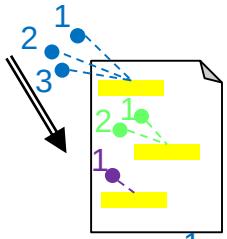
Input content.



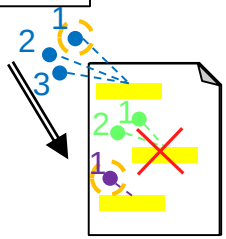
Spotting
Identifies portions of the input document that may be entity mentions



Candidate Selection
Associate the spotted mentions with DBpedia entities that are candidate meanings for those mentions



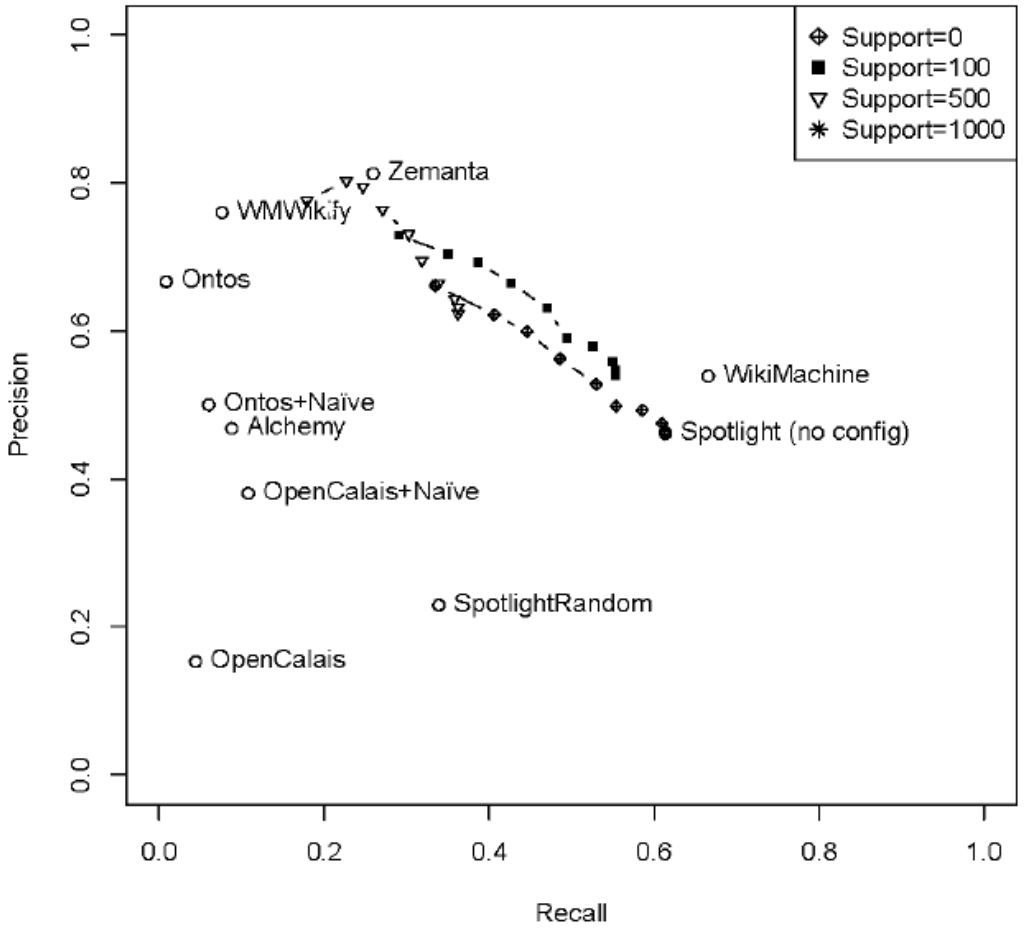
Disambiguation
Decide which is the meaning of each mention



Filtering
Adjust the annotation process by configuring some parameters

Manuel Fiorelli fiorelli@info.uniroma2.it
http://art.uniroma2.it/fiorelli

Evaluation and comparison with other systems



Source:
(Mendes et al., 2011)

Figure 3: DBpedia Spotlight with different configurations (lines) in comparison with other systems (points).

DBpedia Spotlight on premises

Università di Roma Tor Vergata

See https://github.com/dbpedia-spotlight/dbpedia-spotlight-lucene or https://github.com/dbpedia-spotlight/dbpedia-spotlight-model

```
wget http://downloads.dbpedia-spotlight.org/spotlight/dbpedia-spotlight-1.0.0.jar
wget http://downloads.dbpedia-spotlight.org/2016-04/en/model/en.tar.gz
tar xzf en.tar.gz
java -jar dbpedia-spotlight-1.0.jar en_2+2 http://localhost:2222/rest
```

DBpedia Spotlight API

```
curl http://model.dbpedia-spotlight.org/en/annotate  \
  --data-urlencode "text=President Obama called Wednesday on Congress to extend a tax break
  for students included in last year's economic stimulus package, arguing that the policy provides more generous assistance." \
  --data "confidence=0.35" \
  -H "Accept: application/json"
```


Manuel Fiorelli fiorelli@info.uniroma2.it
http://art.uniroma2.it/fiorelli

DBpedia Spotlight UIMA Integration

There is a UIMA annotator in the old* source tree of the DBpedia Spotlight project:

https://github.com/dbpedia-spotlight/dbpedia-spotlight/tree/master/uima

*before they decided to create two separate repositories for the *lucene*, respectively, the *model* variant.

Some remarks – Encyclopedic content

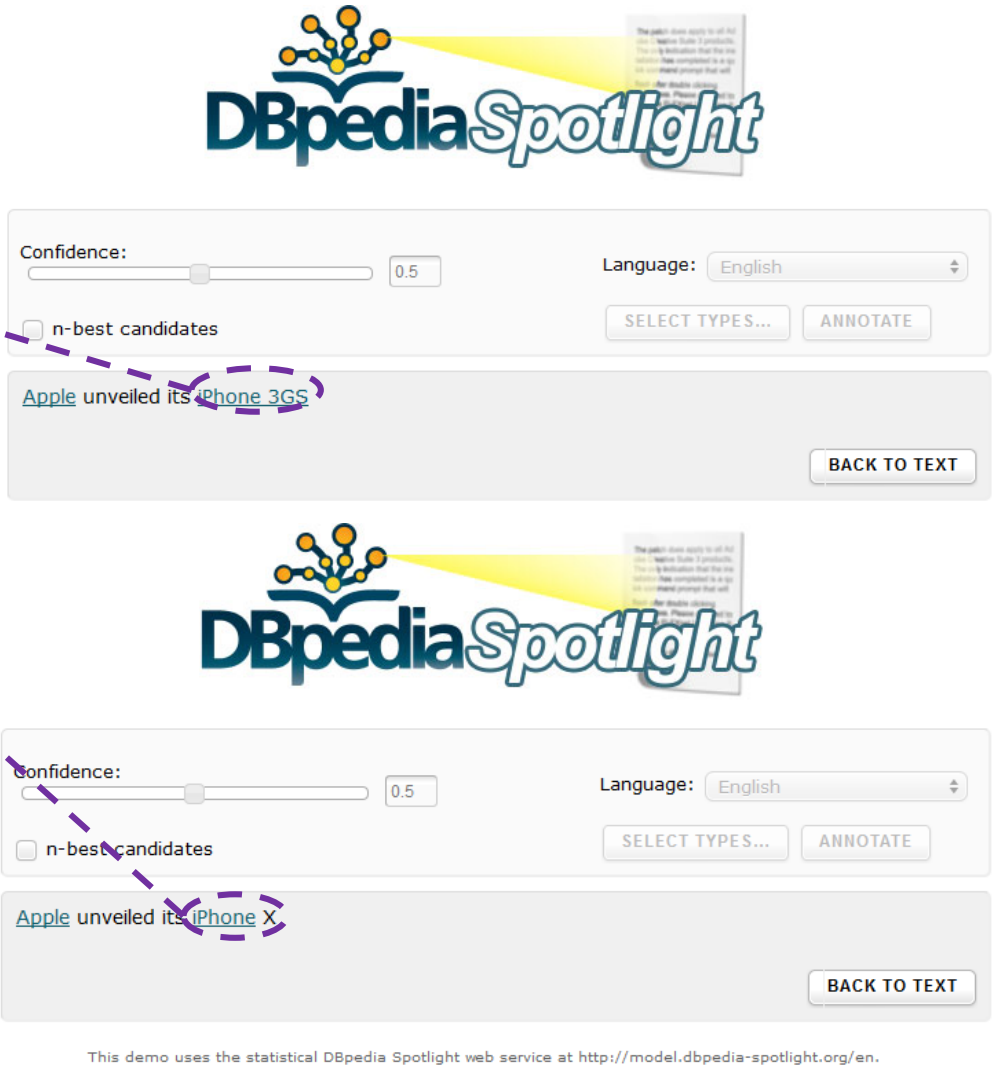
DBpedia Spotlight recognizes only resources defined in DBpedia.

Manuel Fiorelli is not recognized as a mention of a Person, because there isn't a corresponding resource in DBpedia

Some remarks - freshness

iPhone 3GS is recognized as a mention of an Apple product

iPhone X is not recognized as a mention of an Apple product.
Its page on Wikipedia was created on 10 September 2017, while the current version of DBpedia is based on a dump of Wikipedia generated on October 2016. Therefore, the iPhone X does not currently have a resource in DBpedia. DBpedia Live would allow to sidestep this problem; however, there would nonetheless be a problem with the models used by Spotlight, which are not up-to-date.



Manuel Fiorelli fiorelli@info.uniroma2.it
http://art.uniroma2.it/fiorelli

Some remarks – available background knowldge

http://dbpedia.org/resource/IPhone_3GS

vs

http://dbpedia.org/page/Samsung_Galaxy_S_II

- In both cases, *dbp:{successor,after}* *dbp:{predecessor,before}* hold, respectively, the preceding and subsequent models
- In the description of the Galaxy S II we find *dbp:brand* pointing to the family; whereas, in the description of the iPhone 3GS we don't find it. In fact, the sole reference to iPhone is a category (not a resource), which is also used for things that are not smartphones
- Apple is the *dbp:developer* of the iPhone 3GS, the *dbp:manufacturer* of which is Foxconn. Differently, Samsung Electronics is the *dbp:manufacturer* of the Galaxy S II, while its developer is not indicated.

Useful references

- Slides on IE and NER:

  https://web.stanford.edu/class/cs124/lec/Information_Extraction_and_Named_Entity_Recognition.pdf

- Introduction to Information Retrieval

  https://nlp.stanford.edu/IR-book/