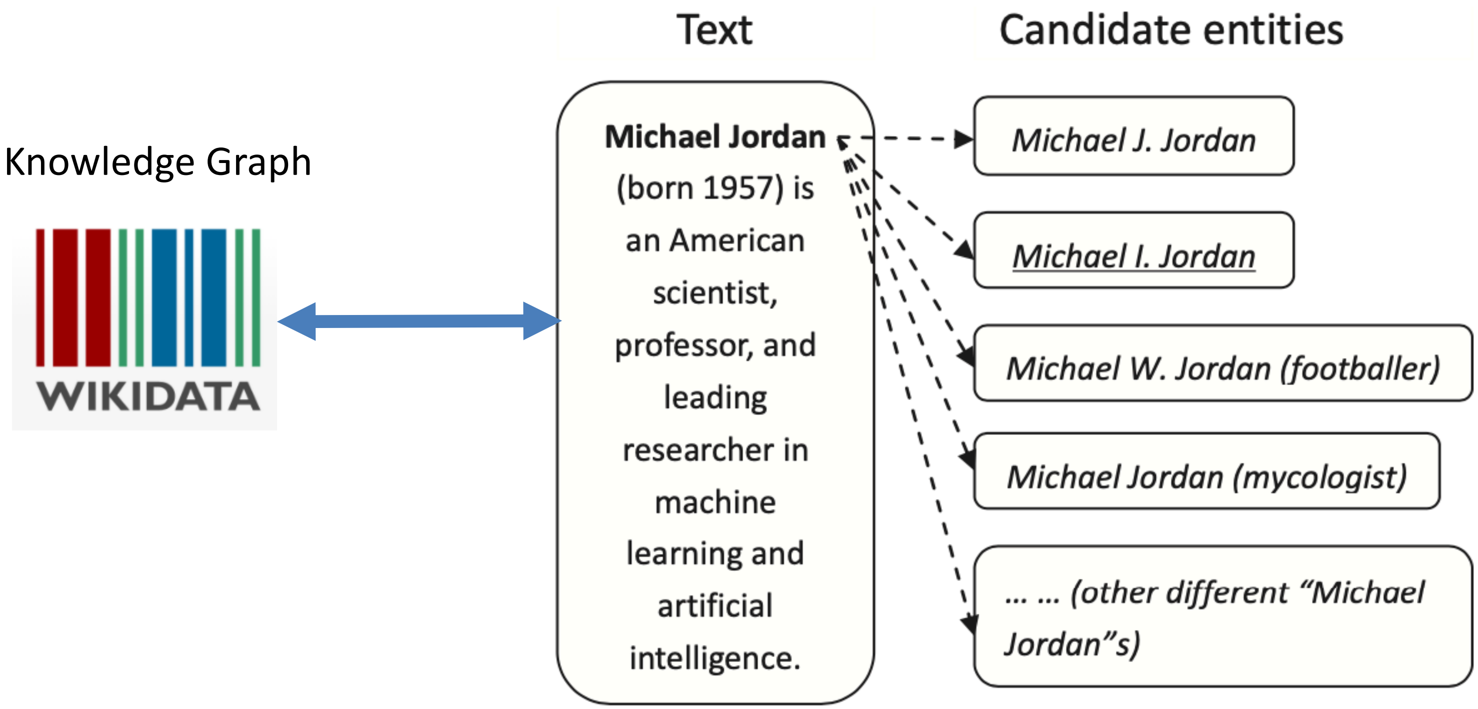
ENTITY LINKING WITH A KG

Craig Knoblock

DSCI 558/CSCI 663

What is entity linking?



Knowledge Graph

WIKIDATA

Text

Michael Jordan (born 1957) is an American scientist, professor, and leading researcher in machine learning and artificial intelligence.

Candidate entities

Michael J. Jordan

Michael I. Jordan

Michael W. Jordan (footballer)

Michael Jordan (mycologist)

… … (other different "Michael Jordan"s)

Sometimes called Named Entity Disambituation (NED)

From Shen et al., Entity Linking with a Knowledge Base

Task Description

- Given a KB with a set of entities E and a text collection with a set of entity mentioned M:

  goal is to map each m ϵ M to the corresponding e ϵ E

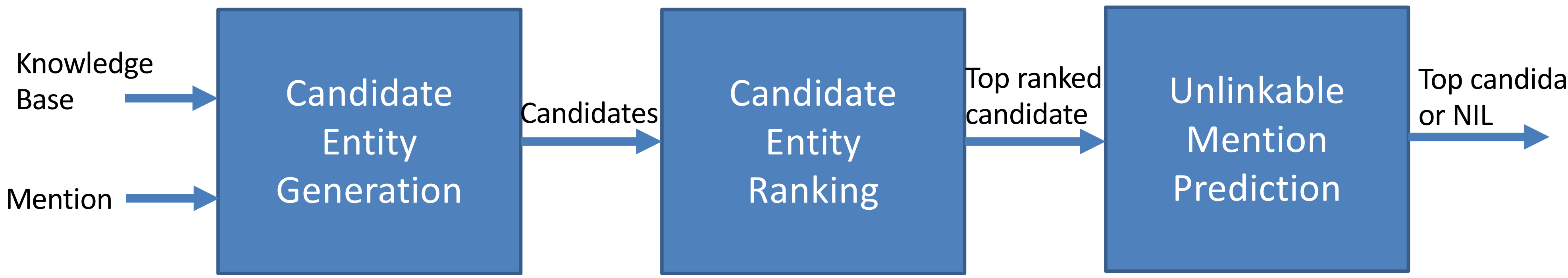
- Some mentions may be unlinkable because the corresponding e is not in the KB
  - In that case m should be labeled as NIL

- This task is typically preceded by Named Entity Recognition to find the entity mentions
  - Commonly implemented using a conditional random field (CRF) using in information extraction
  - Publicly available tools for NER: Stanford NER, OpenNLP, and LingPipe

Relationship of Entity Linking to Other Methods

- Entity coreference resolution
  - Determining the entity mentions within a document and across documents
  - Reduces to this case if there is no knowledge base
  - Entity linking is able to leverage the other information in the KB
- Word sense disambiguation
  - Determine the sense of a word from a catalog of senses e.g., Wordnet
  - Assumes the catalog of senses is complete
    - Not the case in entity linking
- Record linkage
  - Compare different records using a set of attributes using similarity scores
  - Entity linking there may be no other attributes or attributes missing and/or difficult to extract

General Approach to Entity Linking



Applications

- Information extraction
- Information retrieval
- Content analysis
- Question answering
- Knowledge base population

Candidate Entity Generation

- Name dictionary built from Wikipedia:
  - Entity pages – one for each entity
  - Redirect pages – aliases for an entity
  - Disambigutation pages – lists all of the entities with the same name
  - Bold phrases -- in the first paragraph are often aliases
  - Hyperlinks – use the anchor text to linked articles
- Exact matching and partial matching against any of these names
  - Some methods are do some form of spell checking
- Surface form expansion
  - E.g., DOD and Department of Defense both appear in the same document, DOD would get expanded
- Google search – select top Wikipedia pages

Candidate Entity Ranking

	Independent ranking	Collective ranking	Collaborative ranking
Unsupervised methods			
Supervised methods			

Entity ranking features (context independent)

- Name string comparison
  - Whether the entity mention exactly matches the candidate entity name.
  - Whether the candidate entity name starts with or ends with the entity mention.
  - Whether the candidate entity name is the prefix of or postfix of the entity mention.
  - Whether the entity mention is wholly contained in the candidate entity name, or vice-versa.
- Entity popularity
  - Count of the entity references
  - In-degree, out-degree, and length of page
- Entity type
  - Extracted from the NER

Entity ranking features (context dependent)

- Bag of words – uses TF-IDF
- Concept vector – using cosine similarity or embeddings
- Coherence between mapping entities
  - E.g., the link structure of Wikipedia

Supervised entity ranking

- Binary classification
  - uses ML method such as SVM
  - Multiple entities could be positive
- Learning to rank
  - learns a total order on all pairwise comparisons
- Probabilistic methods
  - Using graphical models
- Graph-based approaches
  - Graph-based collective entity linking
- Ensemble methods
  - Any combination of the above

Unsupervised entity ranking

- Vector-space-model (VSM) based methods
  - Compares the vector model of the entity article with the candidate entity
- IR-based methods
  - Similar to what search engines do
- Graph embeddings
  - Exploits the collective structure

Unlinkable Mention Prediction

- Approaches
  - Ignore the problem -- assume all entities are linkable
  - Only return NIL if no candidates are found
  - Assign scores to a match and require the match to be above a threshold
  - Learn a binary classifier to decide if an entity link is correct
    - Typically employ an SVM across a variety of features
  - Incorporate into the entity ranking process where NIL is a choice
    - If NIL is the top-ranked choice, the entity is unlinkable

Evaluation

$$precision = \frac{|\{correctly\ linked\ entity\ mentions\}|}{|\{linked\ mentions\ generated\ by\ system\}|}$$

$$recall = \frac{|\{correctly\ linked\ entity\ mentions\}|}{|\{entity\ mentions\ that\ should\ be\ linked\}|}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Summary

- Entity linking is the problem of identifying the corresponding entity in a KG
- Researchers have linked to Wikipedia, DBPedia, Wikidata, Yago, etc
- Typically, three steps
  - Candidate entity generation
    - Performed with a dictionary or google search
  - Candidate entity ranking
    - Range of unsupervised and supervised methods for performing these task
  - Unlinkable mention prediction
    - Typically done with a threshold or learned classifier
- Useful in a variety of applications
  - From content analysis to KG population