Classification of Wikipedia articles based on category name patterns

klon.wzks.uj.edu.pl/wiki-types
Aleksander Smywiński-Pohl

Goal

- **Wikipedia article → Cyc type**
- *Michael Jackson* → `Singer`, `Actor`, ...
- *Warsaw* → `CapitalCityOfRegion`
- *Computational linguistics* → `FieldOfStudy`
- *Horse* → `DomesticatedAnimal`, `Herbivore`, `Mammal`, ...

Popular approaches

- Parsing of category names, e.g. YAGO
- Parsing of first sentences, e.g. Tipalo
- Mapping of infoboxes, e.g. DBpedia
- Contents-based classification, e.g. Schwa.org

Wikipedia categories

E.g. *Michael Jackson*:

- 20th-century American singers
- 20th-century American male actors
- American disco musicians
- American soul singers
- African-American choreographers
- Drug-related deaths in California
- 1958 births
- 2009 deaths
- Burials at Forest Lawn Memorial Park (Glendale)

YAGO approach

1. Parse category name
2. Identify syntactic head of the category name
3. Keep only categories with plural heads
4. Disambiguate heads against WordNet
5. Assign plural categories as types
6. Attach categorites to the WordNet hierarchy

Problems

- Stanford parser sometimes wrongly identifies heads,
  e.g. *United States **House** of Representatives elections* → *house* instead of *elections*
- Stanford parser sometimes wrongly identifies plurals,
  e.g. *University of Montana **alumni*** → singular
- Plural heads not always indicate types,
  e.g. ***Burials** at Forest Lawn Memorial Park (Glendale)* → event instead of a person

Our approach

- Use Wikipedia categories (like YAGO)
- Do not use a parser
- Identify *patterns* in category names
- Map patterns to types
- Classify articles using category name patterns

Category name patterns

Recognize names of Wikipedia articles in names of categories:

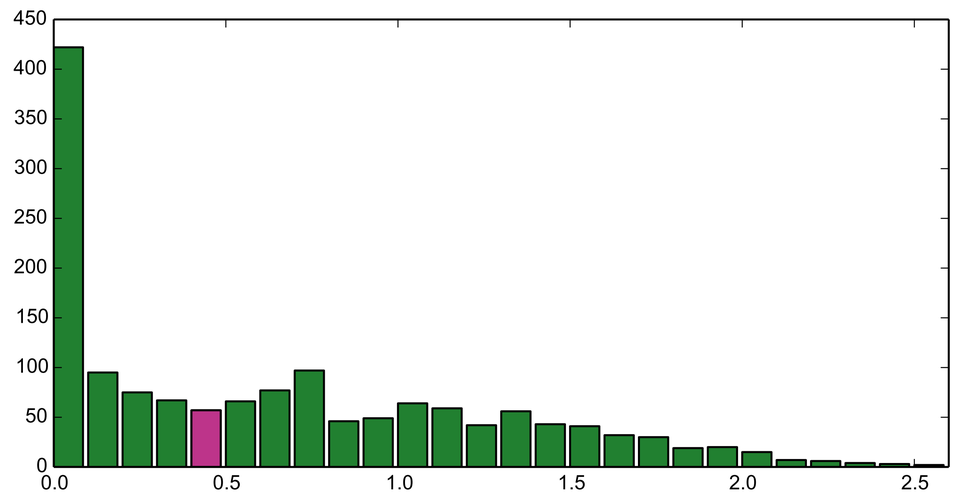
- ***United States House of Representatives** elections* → `.* elections`
- ***University of Montana** alumni* → `.* alumni`
- *Burials at **Forest Lawn Memorial Park (Glendale)*** → `Burials at .*`

Pattern mapping

- Each pattern has many matches – usually hundreds of categories
- Each category has multiple articles
- Assign articles to patterns – thousands of examples
- If there is a dominating type among the articles (e.g. determined using category name parsing) assign it to all articles in the group
- E.g. `Burials at .*` → `Person`

Dominating type

Histogram of pattern entropy



Validation

Tipalo validation dataset

Method	Precisions	Recall	F1
Tipalo (first sentence)	68.0	66.0	67.0
Syntactic head mapping	74.4	50.0	59.8
Pattern mapping	**94.7**	**60.0**	**73.5**