Classifying the Wikipedia Articles into the OpenCyc Taxonomy

Aleksander Pohl
Jagiellonian University, Krakow

Web of Linked Entities 2012

11th of November, 2012

Agenda

Aleksander Pohl

- PhD student at the **University of Science and Technology**, Krakow, Poland
- Assistant lecturer at the Chair in Computational Linguistics, **Jagiellonian University**, Krakow, Poland
- `apohllo@o2.pl`
- `http://apohllo.pl`
- `github.com/apohllo`
- `twitter.com/AleksanderPohl`

Agenda

Classification of the Wikipedia articles

- *categorization* – what categories given article belongs to?
- *classifiction* – what is the type of the concept given article is about?

E.g. *Michael Jackson*
types:

- American
- singer
- dancer

categories:

- Burials at Forest Lawn Memorial Park
- Jackson musical family
- Michael Jackson

Classification of the Wikipedia articles

- *categorization* – what categories given article belongs to?
- *classifiction* – what is the type of the concept given article is about?

E.g. *Michael Jackson*
types:

- American
- singer
- dancer

categories:

- Burials at Forest Lawn Memorial Park
- Jackson musical family
- Michael Jackson

Classification of the Wikipedia articles

- *categorization* – what categories given article belongs to?
- *classifiction* – what is the type of the concept given article is about?

E.g. *Michael Jackson*
types:

- American
- singer
- dancer

categories:

- Burials at Forest Lawn Memorial Park
- Jackson musical family
- Michael Jackson

Classification of the Wikipedia articles

- *categorization* – what categories given article belongs to?
- *classifiction* – what is the type of the concept given article is about?

E.g. *Michael Jackson*
types:

- American
- singer
- dancer

categories:

- Burials at Forest Lawn Memorial Park
- Jackson musical family
- Michael Jackson

OpenCyc „Taxonomy”

- `http://opencyc.org` and `http://sw.opencyc.org`
- 71 thousands of types (first-order collections)
- multiple parents
- hand-crafted (more than 20 years of development)
- types defined in English and with rules
- sophisticated inference engine
- research and commercial versions available

OpenCyc „Taxonomy”

- `http://opencyc.org` and `http://sw.opencyc.org`
- 71 thousands of types (first-order collections)
- multiple parents
- hand-crafted (more than 20 years of development)
- types defined in English and with rules
- sophisticated inference engine
- research and commercial versions available

OpenCyc „Taxonomy”

- `http://opencyc.org` and `http://sw.opencyc.org`
- 71 thousands of types (first-order collections)
- multiple parents
- hand-crafted (more than 20 years of development)
- types defined in English and with rules
- sophisticated inference engine
- research and commercial versions available

OpenCyc „Taxonomy”

- `http://opencyc.org` and `http://sw.opencyc.org`
- 71 thousands of types (first-order collections)
- multiple parents
- hand-crafted (more than 20 years of development)
- types defined in English and with rules
- sophisticated inference engine
- research and commercial versions available

OpenCyc „Taxonomy”

- `http://opencyc.org` and `http://sw.opencyc.org`
- 71 thousands of types (first-order collections)
- multiple parents
- hand-crafted (more than 20 years of development)
- types defined in English and with rules
- sophisticated inference engine
- research and commercial versions available

OpenCyc „Taxonomy”

- `http://opencyc.org` and `http://sw.opencyc.org`
- 71 thousands of types (first-order collections)
- multiple parents
- hand-crafted (more than 20 years of development)
- types defined in English and with rules
- sophisticated inference engine
- research and commercial versions available

OpenCyc „Taxonomy”

- `http://opencyc.org` and `http://sw.opencyc.org`
- 71 thousands of types (first-order collections)
- multiple parents
- hand-crafted (more than 20 years of development)
- types defined in English and with rules
- sophisticated inference engine
- research and commercial versions available

Agenda

Information extraction

- relation extraction
  - types used as relation arguments constraints
  - E.g. *partOf*:
    animal – body part
    organization – (sub)organization
    time slice – time slice
- information retrieval
  - faceted search
  - E.g. CV database:
    What *programming languages* given person knows?
    What *positions* given person occupied?

Information extraction

- relation extraction
  - types used as relation arguments constraints
  - E.g. *partOf*:
    animal – body part
    organization – (sub)organization
    time slice – time slice
- information retrieval
  - faceted search
  - E.g. CV database:
    What *programming languages* given person knows?
    What *positions* given person occupied?

Information extraction

- relation extraction
  - types used as relation arguments constraints
  - E.g. *partOf*:
    animal – body part
    organization – (sub)organization
    time slice – time slice
- information retrieval
  - faceted search
  - E.g. CV database:
    What *programming languages* given person knows?
    What *positions* given person occupied?

Information extraction

- relation extraction
  - types used as relation arguments constraints
  - E.g. *partOf*:
    animal – body part
    organization – (sub)organization
    time slice – time slice
- information retrieval
  - faceted search
  - E.g. CV database:
    What *programming languages* given person knows?
    What *positions* given person occupied?

Information extraction

- relation extraction
  - types used as relation arguments constraints
  - E.g. *partOf*:
    animal – body part
    organization – (sub)organization
    time slice – time slice
- information retrieval
  - faceted search
  - E.g. CV database:
    What *programming languages* given person knows?
    What *positions* given person occupied?

Information extraction

- relation extraction
  - types used as relation arguments constraints
  - E.g. *partOf*:
    animal – body part
    organization – (sub)organization
    time slice – time slice
- information retrieval
  - faceted search
  - E.g. CV database:
    What *programming languages* given person knows?
    What *positions* given person occupied?

Agenda

Type assignment

- Detect the type of the concept in
  - Wikipedia infoboxes
  - Wikipedia categories
  - introductory sentences
  - direct Cyc-Wikiepdia mapping
- Select consistent types by comparing
  - categories vs. infoboxes
  - categories vs. introductory sentences
  - categories vs. direct Cyc-Wikiepdia mapping
- Use cross-checked categories as a last resort

Type assignment

- Detect the type of the concept in
  - Wikipedia infoboxes
  - Wikipedia categories
  - introductory sentences
  - direct Cyc-Wikiepdia mapping
- Select consistent types by comparing
  - categories vs. infoboxes
  - categories vs. introductory sentences
  - categories vs. direct Cyc-Wikiepdia mapping
- Use cross-checked categories as a last resort

Type assignment

- Detect the type of the concept in
  - Wikipedia infoboxes
  - Wikipedia categories
  - introductory sentences
  - direct Cyc-Wikiepdia mapping
- Select consistent types by comparing
  - categories vs. infoboxes
  - categories vs. introductory sentences
  - categories vs. direct Cyc-Wikiepdia mapping
- Use cross-checked categories as a last resort

Type assignment

- Detect the type of the concept in
  - Wikipedia infoboxes
  - Wikipedia categories
  - introductory sentences
  - direct Cyc-Wikiepdia mapping
- Select consistent types by comparing
  - categories vs. infoboxes
  - categories vs. introductory sentences
  - categories vs. direct Cyc-Wikiepdia mapping
- Use cross-checked categories as a last resort

Type assignment

- Detect the type of the concept in
  - Wikipedia infoboxes
  - Wikipedia categories
  - introductory sentences
  - direct Cyc-Wikiepdia mapping
- Select consistent types by comparing
  - categories vs. infoboxes
  - categories vs. introductory sentences
  - categories vs. direct Cyc-Wikiepdia mapping
- Use cross-checked categories as a last resort

Categories

(Most) categories in plural indicate the type of the concept.

Michael Jackson's categories:

- African-American *businesspeople*
- African-American *dancers*
- African-American *male singers*
- African-American *rock singers*

Procedure:

- detect the categories with words in plural
- map the detected categories into Cyc types

Categories

(Most) categories in plural indicate the type of the concept.

Michael Jackson's categories:

- African-American *businesspeople*
- African-American *dancers*
- African-American *male singers*
- African-American *rock singers*

Procedure:

- detect the categories with words in plural
- map the detected categories into Cyc types

Categories

(Most) categories in plural indicate the type of the concept.

Michael Jackson's categories:

- African-American *businesspeople*
- African-American *dancers*
- African-American *male singers*
- African-American *rock singers*

Procedure:

- detect the categories with words in plural
- map the detected categories into Cyc types

Infoboxes

The type of the infobox indicates the type of the concept.

```
{{Infobox musical artist
| birth_name = Michael Joseph Jackson
| alias = Michael Joe Jackson
| birth_date = {{Birth date|1958|8|29|mf=yes}}
| birth_place = [[Gary, Indiana|Gary]],
  Indiana, U.S.
| death_date = {{Death date and age
  |2009|6|25|1958|8|29|mf=yes}}
| death_place = Los Angeles, California, U.S.
| genre = [[Rhythm and blues|R&B]],
  [[pop music|pop]],
  [[rock music|rock]],
  [[soul music|soul]], [[dance music|dance]],
  [[funk]], [[disco]], [[new jack swing]]
```

Background information	
Birth name	Michael Joseph Jackson[1]
Also known as	Michael Joe Jackson
Born	August 29, 1958 Gary, Indiana, U.S.
Died	June 25, 2009 (aged 50) Los Angeles, California, U.S.
Genres	R&B, pop, rock, soul, dance, funk, disco, new jack swing

Procedure:

- map the infobox types into Cyc types

Infoboxes

The type of the infobox indicates the type of the concept.

```
{{Infobox musical artist
| birth_name = Michael Joseph Jackson
| alias = Michael Joe Jackson
| birth_date = {{Birth date|1958|8|29|mf=yes}}
| birth_place = [[Gary, Indiana|Gary]],
  Indiana, U.S.
| death_date = {{Death date and age
  |2009|6|25|1958|8|29|mf=yes}}
| death_place = Los Angeles, California, U.S.
| genre = [[Rhythm and blues|R&B]],
  [[pop music|pop]],
  [[rock music|rock]],
  [[soul music|soul]], [[dance music|dance]],
  [[funk]], [[disco]], [[new jack swing]]
```

Background information	
Birth name	Michael Joseph Jackson[1]
Also known as	Michael Joe Jackson
Born	August 29, 1958 Gary, Indiana, U.S.
Died	June 25, 2009 (aged 50) Los Angeles, California, U.S.
Genres	R&B, pop, rock, soul, dance, funk, disco, new jack swing

Procedure:

- map the infobox types into Cyc types

Introductory sentences

The introductory sentence in most of the cases contains the type of the concept.

Michael Joseph Jackson (August 29, 1958 – June 25, 2009) was an American *recording artist*, *entertainer* and *businessman*.

Procedure:

- detect the location of the type
- disambiguate it against Wikiepdia articles (Improved Wikipedia Miner algorithm)
- map Wikipedia type-articles into Cyc types

Introductory sentences

The introductory sentence in most of the cases contains the type of the concept.

Michael Joseph Jackson (August 29, 1958 – June 25, 2009) was an American *recording artist*, *entertainer* and *businessman*.

Procedure:

- detect the location of the type
- disambiguate it against Wikiepdia articles (Improved Wikipedia Miner algorithm)
- map Wikipedia type-articles into Cyc types

Introductory sentences

The introductory sentence in most of the cases contains the type of the concept.

Michael Joseph Jackson (August 29, 1958 – June 25, 2009) was an American *recording artist*, *entertainer* and *businessman*.

Procedure:

- detect the location of the type
- disambiguate it against Wikiepdia articles (Improved Wikipedia Miner algorithm)
- map Wikipedia type-articles into Cyc types

Direct mapping

There are many identical concepts in Wikipedia and Cyc.

`http://en.wikipedia.org/wiki/Michael_Jackson`
same as
`http://sw.opencyc.org/concept/Mx4rvxrvsZwpEbGdrcN5Y29ycA`

Procedure:

- use the mapping as it is

Direct mapping

There are many identical concepts in Wikipedia and Cyc.

`http://en.wikipedia.org/wiki/Michael_Jackson`
same as
`http://sw.opencyc.org/concept/Mx4rvxrvsZwpEbGdrcN5Y29ycA`

Procedure:

- use the mapping as it is

Direct mapping

There are many identical concepts in Wikipedia and Cyc.

`http://en.wikipedia.org/wiki/Michael_Jackson`
same as
`http://sw.opencyc.org/concept/Mx4rvxrvsZwpEbGdrcN5Y29ycA`

Procedure:

- use the mapping as it is

Cross-checking

Use OpenCyc as a source of definitive knowledge:

- positive – subsumption relation

```
(genls?  #$Person #$Animal) => T
```

- negative – disjointness relation

```
(collections-disjoint?
#$Person #$LiteraryWork) => T
```

Cross-checking

Use OpenCyc as a source of definitive knowledge:

- positive – subsumption relation

```
(genls?  #$Person #$Animal) => T
```

- negative – disjointness relation

```
(collections-disjoint?
#$Person #$LiteraryWork) => T
```

Cross-checking

Use OpenCyc as a source of definitive knowledge:

- positive – subsumption relation

```
(genls?  #$Person #$Animal) => T
```

- negative – disjointness relation

```
(collections-disjoint?
#$Person #$LiteraryWork) => T
```

Agenda

Number of classified concepts

Variant	C_t	C_c	C_v	C_i	Δ
Infoboxes	2188	1712	1471	67	**1471**
Definitions	406	247	154	60	**154**
Cyc mappings	35	25	14	5	**3**
Categories	2470	742	593	—	**593**
Total					**2221**

- C_t – total number of classified concepts
- C_c – number of classifications that were cross-validated
- C_v – number of valid classifications
- C_i – number of invalid classifications
- Δ – number of classifications included in the final result

Precision and recall of the classification

Variant	P	R	$P_{1/2}$	$R_{1/2}$	A	$C_{1/2}$	#
Infoboxes	**97.8**	77.2	90.0	78.0	**92.5**	9.7	**1471**
Definitions	93.5	69.4	**93.9**	68.6	89.0	**5.2**	154
Cyc mappings	94.0	76.4	89.1	71.5	86.1	10.8	3
Categories	81.9	**80.4**	82.1	**78.7**	90.5	10.9	593
Overall (est.)	**93.3**	**77.5**	**88.2**	**77.5**	**91.7**	**9.7**	**2221**

- P – precision for classifications with agreed answer.
- R – recall for classification with agreed answer.
- $P_{1/2}$ – precision for classifications with one uncertain answer.
- $R_{1/2}$ – recall for classifications with one uncertain answer.
- A – agreement between the subjects.
- $C_{1/2}$ – percentage of classifications that were confusing for one of the subjects.

Download

- **http://github.com/apohllo/cyc-wikipedia** – result samples
- full result available upon request

Agenda

Questions?

Why not DBpedia?

- Only classifies articles with infoboxes (approx. 1.5M)
- DBpedia ontology is relatively small (several hundreds of classes)
- The classes lack English and formal definitions

Why not YAGO?

- YAGO classes are compound.
  E.g. *AfricanAmericanBusinesspeople*
- There are inconsistent classifications.
  E.g. *Gertrude Stein* is classified as *Works by Gertrude Stein*

Why not FreeBase?

- Freebase topics are *categories*, not *types*.

Plans

- other languages (Polish, Serbian)
- better algorithm for type determination in Wikipedia categories and introductory sentences
- more fine-grained categories

Thank you!