Distinguishing between Instances and Classes in the Wikipedia Taxonomy

Cäcilia Zirn, Vivi Nastase, Michael Strube

EML Research gGmbH
Heidelberg, Germany
A Wikipedia Ontology?

[Diagram of a Wikipedia ontology showing categories and relationships such as IS-A, CONTAINS, and IS-LOCATED-IN connections between Geography of Spain, Subdivisions of Spain, Landforms of Spain, Autonomous communities of Spain, Canary Islands, Balearic Islands, Volcanoes of Spain, Islands of Spain, Islands of the Canary Islands, Sport in the Canary Islands, Tenerife, Gran Canaria, Sport in Tenerife, Santa Cruz de Tenerife province, Las Palmas de Gran Canaria, CD Tenerife, and Municipalities in Santa Cruz de Tenerife.]
Wikipedia Ontology

The big goal:

Deriving an ontology from Wikipedia automatically

Necessary steps:

1. derive a **taxonomy** from Wikipedia (identify ISA relations), Ponzetto & Strube (AAAI 2007)

2. distinguish between **instances** and **classes** (work presented now)

3. interpret remaining **relations**, Nastase & Strube (AAAI 2008)
Outline

1. Deriving a taxonomy from Wikipedia
2. Instances and classes
3. Methods
4. Evaluation
5. Conclusions
Prerequisites: Category Network
Deriving a taxonomy

Geography of Spain

Subdivisions of Spain

Autonomous communities of Spain

Canary Islands

Sport in the Canary Islands

Sport in Tenerife

CD Tenerife

Santa Cruz de Tenerife province

Municipalities in Santa Cruz de Tenerife

Tenerife

Gran Canaria

Landforms of Spain

Volcanoes of Spain

Islands of Spain

Islands of the Canary Islands

Canary Islands

Balearic Islands

Islands of the Canary Islands

Tenerife

Gran Canaria

Sport in Tenerife

Santa Cruz de Tenerife province

Las Palmas de Gran Canaria
Outline

1. Deriving a taxonomy from Wikipedia
2. Instances and classes
3. Methods
4. Evaluation
5. Conclusions
Instances and Classes

Instances

TENERIFE, TEIDE, 2008
- are unique entities in the world
- in reasoning, they are mapped to objects

Classes

MUNICIPALITIES IN SANTA CRUZ DE TENERIFE, VOLCANOES OF SPAIN
- concepts that subsume classes or individuals
- in reasoning, they are mapped to predicates
Distinction between instances and classes... can be found in WordNet and Cyc was done manually there agreement coefficient on this task on WordNet data $\kappa = 0.75$ (Miller & Hristea, Computational Linguistics 2006) ➤ high cost!
Distinction between instances and classes... can be found in WordNet and Cyc was done manually there agreement coefficient on this task on WordNet data $\kappa = 0.75$ (Miller & Hristea, Computational Linguistics 2006) high cost! develop heuristics to distinguish between instances and classes **fully automatically**
Outline

1. Deriving a taxonomy from Wikipedia
2. Instances and classes
   3. Methods
4. Evaluation
5. Conclusions
Methods

• development of 5 methods
  • Structure-based method
  • NER (Named entity recognition)
  • Capitalization
  • Plural
  • Page

• all are **heuristics**
• use NLP techniques
• based on category network
Methods

- development of 5 methods
  - Structure-based method
  - NER (Named entity recognition)
  - Capitalization
  - Plural
  - Page
Only classes can have instances and classes.

**TENERIFE, TENERIFE NORTH AIRPORT**
Only classes can have instances and classes.

**TENERIFE, TENERIFE NORTH AIRPORT**

- if a category has hyponyms, it has to be a class
- count hyponyms (incoming ISA-links)
• If a category has **more than one** hyponym: the Category is labeled as **Class**

• If a category has **no** hyponym: the Category is labeled as **Instance**
Only classes can have instances and classes.

TENERIFE, TENERIFE NORTH AIRPORT
Only classes can have instances and classes.

**Tenerife, Tenerife North Airport**

- labeling of the ISA-links has been done automatically
- possible that links are classified erroneously
- tolerate one erroneous link
Structure-based method (4)

- If a category has **exactly one** hyponym:
  - If the hyponym **has a hyponym** itself: 
    - the Category is labeled as **Class**
  - If the hyponym **has no hyponym**:
    - the Category is labeled as **Instance**
Methods

• development of 5 methods
  • Structure-based method
  • NER (Named entity recognition)
  • Capitalization
  • Plural
  • Page
Instances correspond to unique entities in the world and are therefore named entities.

**TENERIFE, CD TENERIFE**
Instances correspond to unique entities in the world and are therefore named entities.

TENERIFE, CD TENERIFE

Idea: use a named entity recognizer
Utility: Named Entity Recognizer

- input: noun phrase
- output: named entity tags
  - *Person, Location, Organization* for named entities
  - *Other* for the rest
- we use CRFClassifier (Stanford)
Instances correspond to unique entities in the world and are therefore named entities.

**TENERIFE, CD TENERIFE**
Instances correspond to unique entities in the world and are therefore named entities.

**Tenerife, CD Tenerife**

- Some names consist of complex noun structures: *Autonomous Communities of Spain*
  - only lexical heads are passed to named entity recognizer
  - lexical heads are extracted using Stanford Parser
Utility: Parser

- analyzes the grammatical structure of the input
- outputs a parse tree

```
ROOT
  NP
  NP
  JJ autonomous
  NNS communities
  PP of
  IN NP
    NNP Spain
```
Utility: Lexical head finder

- lexical heads: determine the syntactic properties of a phrase
- in a noun phrase: the noun

```
autonomous communities of Spain
```
Method: NER (3)

- If the named entity recognizer returns one of the labels: *Person, Location, Organization*:
  - the Category is labeled as **Instance**
- If the named entity recognizer returns the label *Other*:
  - the Category is labeled as **Class**
Method: NER (3)

- If the named entity recognizer returns one of the labels: *Person, Location, Organization:*
  - the Category is labeled as **Instance**
- If the named entity recognizer returns the label *Other*
  - the Category is labeled as **Class**

the parser sometimes returns several heads

- If the majority of returned labels is *Other:*
  - the Category is labeled as **Class**
- otherwise
  - the Category is labeled as **Instance**
Methods

• development of 5 methods
  • Structure-based method
  • NER (Named entity recognition)
  • Capitalization
  • Plural
  • Page
Method: Capitalization (1)

Content words belonging to a named entity are capitalized.

Convention for Wikipedia titles.

TENERIFE LADIES OPEN
and
AUTONOMOUS COMMUNITIES OF SPAIN
Content words belonging to a named entity are capitalized.

Convention for Wikipedia titles.

**TENERIFE LADIES OPEN**

and

**AUTONOMOUS COMMUNITIES OF SPAIN**

- Bunescu & Paşca (2006) developed heuristic to process Wikipedia page titles:
  “If all content words of a page title are capitalized, it corresponds to a named entity”

► We apply this heuristic to category titles
1. preprocess first word
   • first word is always capitalized
     ➤ pass it to CRFClassifier
   • if it is not recognized as a named entity: lowercase the word

2. filter out function words

3. analyze remaining words:
   • If all words are capitalized
     ➤ the Category is labeled as **Instance**
   • otherwise
     ➤ the Category is labeled as **Class**
Methods

- development of 5 methods
  - Structure-based method
  - NER (Named entity recognition)
  - Capitalization
  - Plural
  - Page
Instances are unique generally used in singular form.

TENERIFE, SPAIN

and

AUTONOMOUS COMMUNITIES OF SPAIN
Method: Plural (1)

Instances are unique generally used in singular form.

TENERIFE, SPAIN
and
AUTONOMOUS COMMUNITIES OF SPAIN

• Exceptions: “The Millers are coming to our party” not to be expected in Wikipedia category titles
Instances are unique generally used in singular form.

TENERIFE, SPAIN
and
AUTONOMOUS COMMUNITIES OF SPAIN
Instances are unique generally used in singular form.

TENERIFE, SPAIN

and

AUTONOMOUS COMMUNITIES OF SPAIN

- the grammatical number of the lexical head is the same as the number of the category title

- we parse the category title with the Stanford Parser, obtaining:
  - the lexical head(s)
  - the part-of-speech tags
Utility: Part-of-speech tagging

- assigns each word its part of speech
- tags of interest:
  - NNPS = noun, proper, plural
  - NNS = noun, common, plural

Autonomous/JJ communities/NNS of/IN Spain/NNP
Method: Plural (3)

- If the lexical head of a phrase is tagged as plural noun (NNS, NNPS)
  - the Category is labeled as **Class**
- otherwise
  - the Category is labeled as **Instance**
Methods

• development of 5 methods
  • Structure-based method
  • NER (Named entity recognition)
  • Capitalization
  • Plural
  • Page
Articles should be placed in categories with the same name.

Advice for authors in Wikipedia.

SPAIN, TENERIFE
Articles should be placed in categories with the same name.

**Spain, Tenerife**

- a number of articles have homonymous categories
- most articles refer to unique entities

Heuristic: a category containing a page with the same name is an instance.
Method: Page (2)

- If a page with homonymous title exists
  ➤ the Category is labeled as **Instance**
- otherwise
  ➤ the Category is labeled as **Class**
Outline

1. Deriving a taxonomy from Wikipedia
2. Instances and classes
3. Methods
4. Evaluation
5. Conclusions
Use ResearchCyc as **gold standard**.

ResearchCyc

- distinguishes between #$Individual and #$SetOrCollection
- distinction is done manually

overlap Wikipedia / ResearchCyc:
- 7860 concepts
  - 44.35%(3486)#$Individual
  - 55.65%(4374)#$SetOrCollection
Use ResearchCyc as **gold standard**.
Measures (1)

<table>
<thead>
<tr>
<th></th>
<th>$T_{\text{instances}}$</th>
<th>$F_{\text{classes}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{\text{instances}}$</td>
<td>$T_{\text{classes}}$</td>
<td></td>
</tr>
</tbody>
</table>

$$\text{Prec}_{\text{instances}} = \frac{T_{\text{instances}}}{T_{\text{instances}} + F_{\text{instances}}}$$

- $T_{\text{instances}}$: Instance in Wiki & individual in Cyc
- $F_{\text{instances}}$: Instance in Wiki but not individual Cyc
- $T_{\text{classes}}$: Class in Wiki & SetOrCollection Cyc
- $F_{\text{classes}}$: Class in Wiki but not SetOrCollection in Cyc
Measures (2)

<table>
<thead>
<tr>
<th>$T_{\text{instances}}$</th>
<th>$F_{\text{classes}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{\text{instances}}$</td>
<td>$T_{\text{classes}}$</td>
</tr>
</tbody>
</table>

$$\text{Prec}_{\text{classes}} = \frac{T_{\text{classes}}}{T_{\text{classes}} + F_{\text{classes}}}$$

$T_{\text{instances}}$: Instance in Wiki & Individual in Cyc

$F_{\text{instances}}$: Instance in Wiki but **not** Individual in Cyc

$T_{\text{classes}}$: Class in Wiki & SetOrCollection in Cyc

$F_{\text{classes}}$: Class in Wiki but **not** SetOrCollection in Cyc
### Measures (3)

<table>
<thead>
<tr>
<th>$T_{instances}$</th>
<th>$F_{classes}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{instances}$</td>
<td>$T_{classes}$</td>
</tr>
</tbody>
</table>

**Accuracy** = \[
\frac{T_{instances} + T_{classes}}{T_{instances} + F_{instances} + T_{classes} + F_{classes}}
\]

- **$T_{instances}$**: Instance in Wiki & individual in Cyc
- **$F_{instances}$**: Instance in Wiki but **not** individual Cyc
- **$T_{classes}$**: Class in Wiki & SetOrCollection Cyc
- **$F_{classes}$**: Class in Wiki but **not** SetOrCollection in Cyc
Evaluate every method separately

<table>
<thead>
<tr>
<th>Method</th>
<th>Prec$_{\text{instances}}$</th>
<th>Prec$_{\text{classes}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER</td>
<td>85.23</td>
<td>76.84</td>
</tr>
<tr>
<td>page</td>
<td>66.1</td>
<td>91.5</td>
</tr>
<tr>
<td>capitalization</td>
<td>85.99</td>
<td>82.44</td>
</tr>
<tr>
<td>plural</td>
<td>66.44</td>
<td>87.99</td>
</tr>
<tr>
<td>structure</td>
<td>56.17</td>
<td>87.21</td>
</tr>
</tbody>
</table>
Evaluate every method separately

<table>
<thead>
<tr>
<th>Method</th>
<th>Prec$_{\text{instances}}$</th>
<th>Prec$_{\text{classes}}$</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER</td>
<td>85.23</td>
<td>76.84</td>
<td>79.69</td>
</tr>
<tr>
<td>page</td>
<td>66.1</td>
<td>91.5</td>
<td>75.74</td>
</tr>
<tr>
<td>capitalization</td>
<td>85.99</td>
<td>82.44</td>
<td>83.82</td>
</tr>
<tr>
<td>plural</td>
<td>66.44</td>
<td>87.99</td>
<td>75.24</td>
</tr>
<tr>
<td>structure</td>
<td>56.17</td>
<td>87.21</td>
<td>64.71</td>
</tr>
</tbody>
</table>
Final setting

Classification schemes

A) Accuracy scheme

• method with best accuracy: capitalization
• (regard method as baseline)
Final setting

Classification schemes

B) Precision scheme

- order methods according to their precision (\(\text{Prec}_{\text{instances}}\) or \(\text{Prec}_{\text{classes}}\))

1. page ➔ class
2. plural ➔ class
3. structure ➔ class
4. capitalization ➔ instance
5. remaining categories ➔ class
Final setting

Classification schemes

C) Voting scheme

1. page & plural ➔ class
2. capitalization & NER ➔ instance
3. remaining categories ➔ precision scheme
Final setting

Classification schemes

A) Accuracy scheme
B) Precision scheme
C) Voting scheme

Special form of cross-validation:
- 5 rounds of binary random splits
- maintain the #$Individual / #$SetOrCollection distribution
- evaluate on the resulting 10 data sets
## Final results

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision\textsubscript{\textit{instances}}</th>
<th>Precision\textsubscript{\textit{classes}}</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A) \textit{Accuracy sc.}</td>
<td>85.99±0.54</td>
<td>82.44±0.63</td>
<td>82.82±0.5</td>
</tr>
<tr>
<td>B) \textit{Precision sc.}</td>
<td>90.92±0.41</td>
<td>77.36±0.52</td>
<td>81.64±0.42</td>
</tr>
<tr>
<td>C) \textit{Voting sc.}</td>
<td>89.21±0.46</td>
<td>81.82±0.52</td>
<td>84.52±0.34</td>
</tr>
</tbody>
</table>
Discussion

- Preprocessing errors, e.g. wrong parsing results
  \[\text{AND YOU WILL KNOW US BY THE TRAIL OF DEAD ALBUMS}\]

- Recognizing named entities:
  \textsc{Bee Train}
  If components of a named entity are not named entities, it is not recognized

- Concepts in Cyc:
  Inter-agreement between judges is not 100% different possible judgements
1. Deriving a taxonomy from Wikipedia
2. Instances and classes
3. Methods
4. Evaluation
5. Conclusions
Conclusions

• automatic distinction between instances and classes is possible with a high accuracy (84.52%)
• combining the methods with machine learning could improve performance even more
• next step: introducing distinction between instances and classes to Wikipedia articles
• methods can easily be applied to other languages
Thanks!

Acknowledgements

- Simone Ponzetto for his work in deriving the taxonomy
- Klaus Tschira Foundation

Check out

... the results (RDF Schema)

www.eml-research.de/nlp/download/wikitaxonomy.php

... more papers on Wikipedia

www.eml-research.de/~strube