Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia

Simone Paolo Ponzetto
(joint work with Roberto Navigli)

Seminar für Computerlinguistik
University of Heidelberg

ponzetto@cl.uni-heidelberg.de

April 29, 2009

Encyclopedic knowledge & NLP

The crisis at **General Motors** threatens to drag down **Adam Opel**, **a storied German brand** that **GM** bought 80 years ago, on the eve of the Great Depression. Many in the industry say **Opel** has a future only if **it** can get a temporary helping hand from the German government.

But whether Chancellor Angela Merkel will make available the public financing needed to help release **Opel** from the clutches of **General Motors** now depends on a reluctant government, an influential automotive union that wants politicians to save jobs, and employees who yearn to re-establish **Opel** as **an independent German company**.

---

source: Herald Tribune Europe, March 6, 2009

Encyclopedic knowledge & NLP

The crisis at **General Motors** threatens to drag down **Adam Opel**, **a storied German brand** that **GM** bought 80 years ago, on the eve of the Great Depression. Many in the industry say **Opel** has a future only if **it** can get a temporary helping hand from the German government.

But whether Chancellor Angela Merkel will make available the public financing needed to help release **Opel** from the clutches of **General Motors** now depends on a reluctant government, an influential automotive union that wants politicians to save jobs, and employees who yearn to re-establish **Opel** as **an independent German company**.

---

source: Herald Tribune Europe, March 6, 2009

Encyclopedic knowledge & NLP

The crisis at **General Motors** threatens to drag down **Adam Opel**, **a storied German brand** that **GM** bought 80 years ago, on the eve of the Great Depression. Many in the industry say **Opel** has a future only if **it** can get a temporary helping hand from the German government.

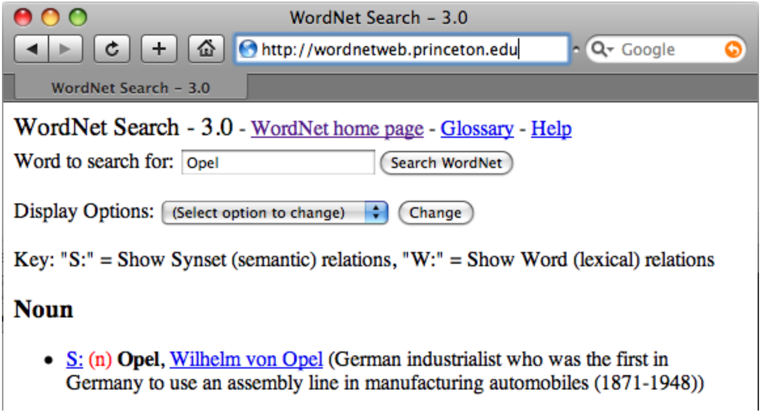
But whether Chancellor Angela Merkel will make available the public financing needed to help release **Opel** from the clutches of **General Motors** now depends on a reluctant government, an influential automotive union that wants politicians to save jobs, and employees who yearn to re-establish **Opel** as **an independent German company**.

---

source: Herald Tribune Europe, March 6, 2009

What about a widely used resource like *WordNet*?

Encyclopedic knowledge & NLP



Encyclopedic knowledge & NLP

The crisis at **General Motors** threatens to drag down **Adam Opel**, **a storied German brand** that **GM** bought 80 years ago, on the eve of the Great Depression. Many in the industry say **Opel** has a future only if **it** can get a temporary helping hand from the German government.

But whether Chancellor Angela Merkel will make available the public financing needed to help release **Opel** from the clutches of **General Motors** now depends on a reluctant government, an influential automotive union that wants politicians to save jobs, and employees who yearn to re-establish **Opel** as **an independent German company**.

---

source: Herald Tribune Europe, March 6, 2009

Encyclopedic knowledge & NLP

The crisis at General Motors threatens to drag down **Adam Opel**, **a storied German brand** that GM bought 80 years ago, on the eve of the Great Depression. Many in the industry say **Opel** has a future only if **it** can get a temporary helping hand from the German government.

But whether Chancellor Angela Merkel will make available the public financing needed to help release **Opel** from the clutches of General Motors now depends on a reluctant government, an influential automotive union that wants politicians to save jobs, and employees who yearn to re-establish **Opel** as **an independent German company**.

---

source: Herald Tribune Europe, March 6, 2009

And *Cyc*?

Encyclopedic knowledge & NLP

Encyclopedic knowledge & NLP

The crisis at **General Motors** threatens to drag down **Adam Opel**, **a storied German brand** that **GM** bought 80 years ago, on the eve of the Great Depression. Many in the industry say **Opel** has a future only if **it** can get a temporary helping hand from the German government.

But whether Chancellor Angela Merkel will make available the public financing needed to help release **Opel** from the clutches of **General Motors** now depends on a reluctant government, an influential automotive union that wants politicians to save jobs, and employees who yearn to re-establish **Opel** as **an independent German company**.

---

source: Herald Tribune Europe, March 6, 2009

Encyclopedic knowledge & NLP

The crisis at **General Motors** threatens to drag down **Adam Opel**, **a storied German brand** that **GM** bought 80 years ago, on the eve of the Great Depression. Many in the industry say **Opel** has a future only if **it** can get a temporary helping hand from the German government.

But whether Chancellor Angela Merkel will make available the public financing needed to help release **Opel** from the clutches of **General Motors** now depends on a reluctant government, an influential automotive union that wants politicians to save jobs, and employees who yearn to re-establish **Opel** as **an independent German company**.

---

source: Herald Tribune Europe, March 6, 2009

Let's check *Wikipedia* on that topic!

Wikipedia

Log in / create account

article | discussion | edit this page | history

WIKIPEDIA
The Free Encyclopedia

navigation
- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go | Search

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version

Opel

From Wikipedia, the free encyclopedia

This article is about the European car manufacturer. For the album by Syd Barrett, see Opel (album). For The former Australian telecommunications operator, see OPEL Networks.

Adam Opel GmbH (commonly known as **Opel**) is a German automaker, part of General Motors. The company was founded on 21 January 1863, and began making automobiles in 1899. Opel was acquired by General Motors Corporation in 1929 and continues as a subsidiary.[1] Opel is part of GM Europe, and is GM's largest European brand, and with Vauxhall Motors in the UK, forms GM's core European business.[2]

Contents [hide]

Adam Opel GmbH

OPEL

Type	Private company, subsidiary of General Motors Europe, (GM Europe is a division of General Motors)
Founded	1863
Headquarters	Rüsselsheim, Germany
Key people	Hans Demant, CEO
Industry	Automotive
Products	Automobile
Owner(s)	General Motors Corporation

Wikipedia

v • d • e **Opel GmbH** (Category | Vehicles) [hide]

Passenger	Agila • Antara • Astra • Corsa • GT • Insignia • Meriva • Signum • Tigra • Zafira
LCV	Combo • Movano • Vivaro
Concept	Opel Gran Turismo Concept • Aero GT • Antara GTC • CD • Diesel Rekordwagen • Eco Speedster • Frogster • Frua Diplomat • G90 • GT 2 • Insignia • Maxx • OPC X-Treme • Snowtrekker • Trixx
Historic	Admiral • Ascona • Blitz • Calibra • Campo • Commodore • Diplomat • Frontera • GT • Kadett • Kapitän • Manta • Monterey • Monza • Olympia • Omega • Rekord • Senator • Sintra • Speedster • Super Six • Vectra • "Lutzmann" • "Doktorwagen" • 4 PS "Laubfrosch" • RAK2 • P4

Parent company: General Motors Corporation, • See also: Vauxhall Motors, Holden

v • d • e **Automotive brands of General Motors and those of its affiliates and former affiliates** [show]

v • d • e **General Motors** [show]

v • d • e **Opel, a subsidiary of General Motors, road car timeline, 1947–1970s — next »** [hide]

Type	1940s	1950s	1960s	1970s
	7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9
Small family car			Kadett A, Kadett B	Kadett B, Kadett C
Small family car	Olympia	Olympia, Olympia Rekord	Olympia A	
Large family car		Rekord PI	Rekord PII	Ascona A, Ascona B
Executive car			Rekord A, Rekord B, Rekord C	Rekord C, Rekord D, Rekord E
Executive car	Kapitän	Kapitän, Kapitän, Kapitän PI / PII	Kapitän PI / PII, Commodore A	Commodore A, Commodore B, Commodore C

Wikipedia

languages

- العربية
- Aragonés
- Беларуская
- Беларуская (тарашкевіца)
- Bosanski
- Български
- Català
- Česky
- Dansk

Categories: Motor vehicle companies | Automotive companies of Germany | General Motors | Motor vehicle manufacturers of Germany | German brands | Opel | Car manufacturers | General Motors marques | Companies established in 1863

This talk

we are after a “steak and lobster” combination . . .

- ✓ manual approaches achieve *high quality for a limited coverage*
- ✓ automatic ones achieve *large coverage for a lower quality*

This talk

we are after a “steak and lobster” combination . . .

- ✓ manual approaches achieve *high quality for a limited coverage*
- ✓ automatic ones achieve *large coverage for a lower quality*

- start *manually annotated semi-structured input*
  - ➠ Wikipedia
- use a *large-coverage* taxonomy developed from Wikipedia
  - ➠ WikiTaxonomy

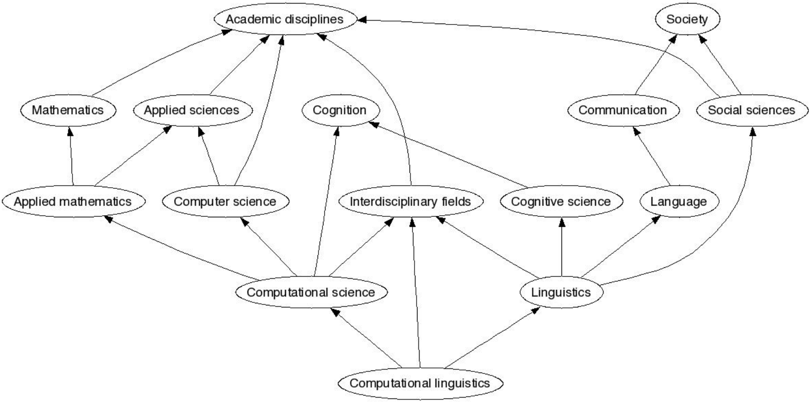
➠ <u>overcome WikiTaxonomy’s limitations by mapping it to WordNet</u>

Outline

Outline

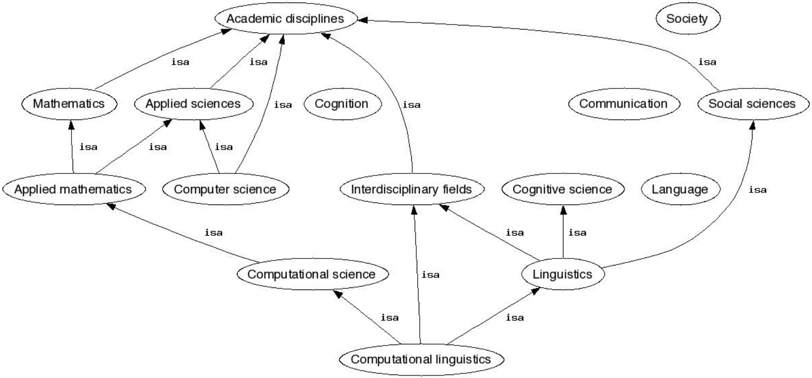
Deriving a taxonomy from Wikipedia

- start with semantic network



Deriving a taxonomy from Wikipedia

- induce **semantically-typed** relations



Deriving a taxonomy from Wikipedia

originally presented in Ponzetto & Strube (2007)

- the category network is merely a *thematic categorization* of the topics of articles

- $\boxed{\text{task}}$ label the relations between categories as *isa* and *notisa*
- $\boxed{\text{goal}}$ transform a *thematic categorization* into a **fully-fledged taxonomy**

Deriving a taxonomy from Wikipedia

- **methods**:
  - syntactic matching
  - connectivity in the network
  - lexico-syntactic patterns
- **results**:
  - we start with 337,522 categories and 743,140 links
  - we generate 335,128 *isa* relations

➠ $\boxed{\textit{large-scale}, \textit{multi-domain}\ \textbf{taxonomy}}$

Category network cleanup (1)

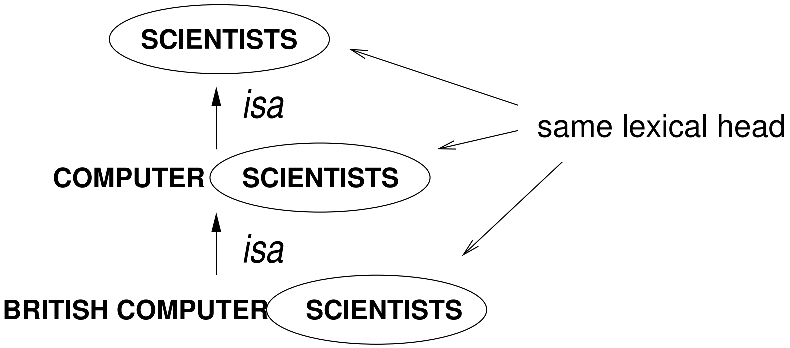
- **removal of meta-categories** used for encyclopedia management, e.g. categories under WIKIPEDIA ADMINISTRATION
- we remove all nodes whose labels contain any of the following strings: MEDIAWIKI, TEMPLATE, USER, PORTAL, CATEGORIES, ARTICLES, PAGES
- this leaves
  - 240,760 categories
  - 515,423 links

  still to be processed

Refinement link identification (2)

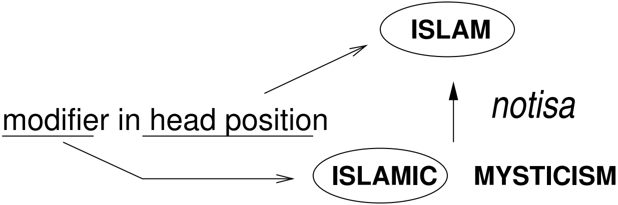
- patterns such as **y x and x by z**
- their purpose is to better structure and simplify the categorization network
- we assume this represents **is-refined-by-relations**
- this labels 126,920 category links *notisa* and leaves 388,503 relations to be analyzed

Syntax-based methods (3)



- **head matching** labels pairs of categories sharing the **same lexical head word (or lemma)**
- we identify lexical heads using the *Stanford parser* and lemmata using `morpha`

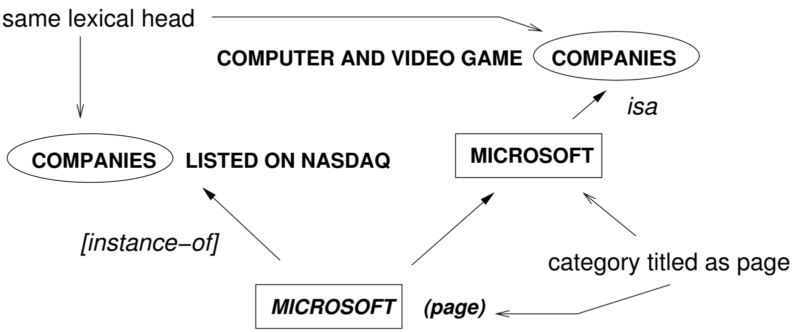
Syntax-based methods (3)



- **modifier matching** labels pairs as *notisa*, if the stem of the lexical head of one of the categories occurs in non-head position in the other category, e.g. CRIME COMICS and CRIME or ISLAMIC MYSTICISM and ISLAM

- *head* and *modifier matching* identify 141,728 *isa* relations and 67,437 *notisa* relations
  - relatively 'simple' (→ **baseline**)
  - still *large coverage*

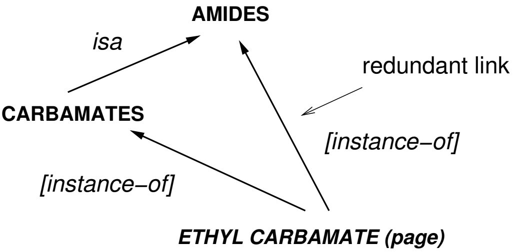
Connectivity-based methods (4)



- **instance categorization** assumes that relations between entities (Wikipedia pages) and classes (categories) can be labeled as *instance-of* (Suchanek et al., 2007)

➠ identifies 14,886 *isa* relations

Connectivity-based methods (4)

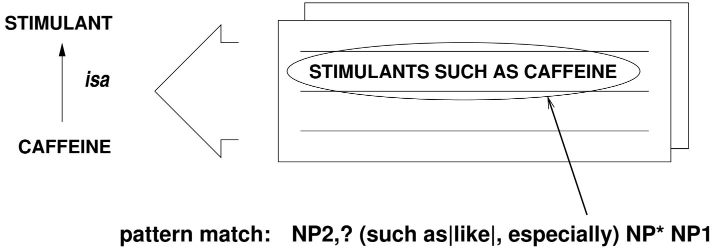


- if users **redundantly** categorize we take this as evidence for *isa* relations, e.g. ETHYL CARBAMATE

➠ identifies 16,523 *isa* relations

we are left with 147,929 unclassified relations ...

Lexico-syntactic based methods (5)



- we apply **lexico-syntactic patterns** to sentences in large text corpora to identify *isa* relations (Hearst, 1992; Caraballo, 1999)
- we assume that patterns used for identifying *meronymic relations* (Berland & Charniak, 1999) indicate that the relation **is not** an *isa* relation ➠ **notisa**

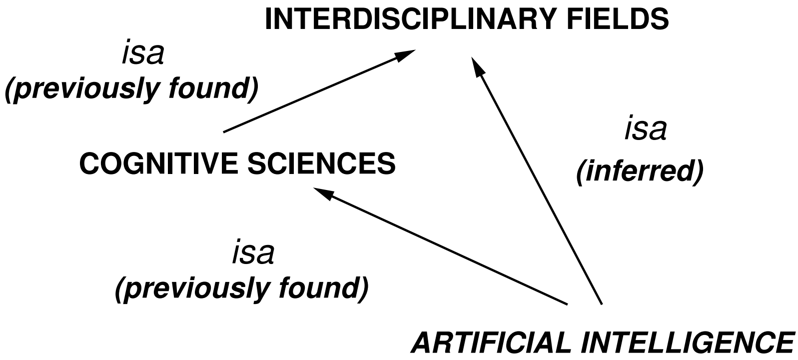
Lexico-syntactic based methods (5)

- examples of ISA patterns:
  - *NP2,?* (such as|like|, especially) *NP* NP1*
    a stimulant such as caffeine
  - *NP1 NP** (and|or|,like) other *NP2*
    caffeine and other stimulants
- examples of NOTISA patterns:
  - *NP2*'s *NP1*
    car's engine
  - *NP2* with *NP1*
    a car with an engine

Lexico-syntactic based methods (5)

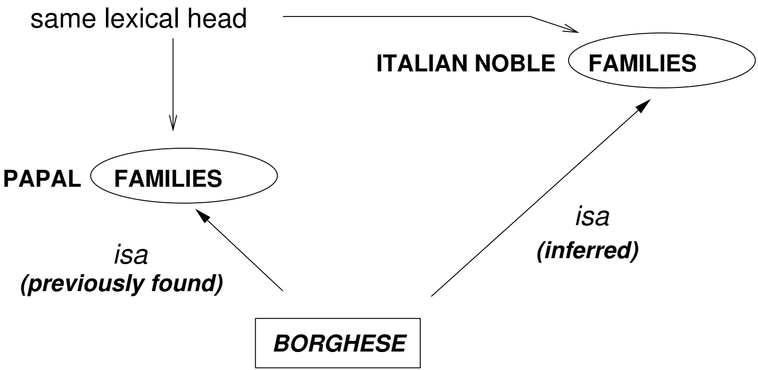
- we use the Tipster corpus (2.5×10^8 words) and the English Wikipedia itself (8×10^8 words)
- Preprocessing: tokenization, sentence splitting, POS-tagging, NP-chunking ➠ 15GB data
- *majority voting strategy* between *isa* and *notisa* patterns
- this method identifies 49,054 *isa* relations
- we apply this method also to the relations identified in step (4) and filter out 3,226 previously identified *isa* relations

Inference-based methods (6)



- *assumption*: the *isa* relation models set inclusion, and therefore **is a transitive relation**
- propagate previously found relations based on transitivity

Inference-based methods (6)



- propagate all *isa* relations to those supercategories whose head lemma matches the head lemma of a *previously identified isa supercategory*
- ➠ propagate the *isa* relation to **the sisters of the previously identified isa supercategories**

Size of the taxonomy

		ResearchCyc	WordNet	Wikipedia (sem. network)	Wikipedia (taxonomy)
# nodes	# concepts	300,000			
	# synsets		117,659		
	# categories			337,522	209,919
# edges	# assertions	3,000,000			
	# semantic pointers		285,348		
	# category links			743,140	335,128

Manual evaluation

1.106 instances evaluated manually by three judges

	R	P	F
random baseline	51.1	51.6	51.3
syntax (1-3)	17.0	95.4	28.9
connectivity (1-4, 6)	38.9	88.1	54.0
pattern-based (1-3, 5-6)	62.7	84.3	71.9
all (1-6)	69.5	81.6	75.0

. . . but is it *that* good?

manual inspection reveals that WikiTaxonomy

. . . but is it *that* good?

manual inspection reveals that WikiTaxonomy

1. includes 3,487 roots

- still a sparse set of taxonomic islands . . .

. . . but is it *that* good?

manual inspection reveals that WikiTaxonomy

1. includes 3,487 roots

- still a sparse set of taxonomic islands . . .

2. still suffers from errors (being automatically generated)

- FRUIT *isa* PLANTS

. . . but is it *that* good?

manual inspection reveals that WikiTaxonomy

1. includes 3,487 roots

   - still a sparse set of taxonomic islands . . .

   ! disambiguate the Wikipedia categories to WordNet synsets

   ➠ use WordNet as top-level taxonomy, thus *integrating* WikiTaxonomy

2. still suffers from errors (being automatically generated)

   - FRUIT *isa* PLANTS

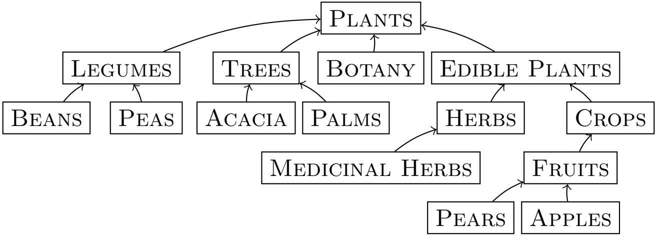
   ! align WikiTaxonomy to WordNet

   ➠ use WordNet as reference taxonomy to *restructure* WikiTaxonomy

Outline

Preliminaries

- input: WikiTaxonomy (Ponzetto & Strube, 2007)[1]



[1] www.eml-research.de/nlp/download/wikitaxonomy.php

Preliminaries

- input: WikiTaxonomy (Ponzetto & Strube, 2007)[1]

- view the taxonomy as a forest $\mathcal{F}$ of category trees T
- for each category $c \in T$ find the lexical items *heads*(c) best matching a category label in WordNet:

[1] `www.eml-research.de/nlp/download/wikitaxonomy.php`

Preliminaries

- input: WikiTaxonomy (Ponzetto & Strube, 2007)[1]

- view the taxonomy as a forest $\mathcal{F}$ of category trees T
- for each category $c \in T$ find the lexical items $heads(c)$ best matching a category label in WordNet:

- **full match**: PLANTS ➠ plant

[1] www.eml-research.de/nlp/download/wikitaxonomy.php

Preliminaries

- input: WikiTaxonomy (Ponzetto & Strube, 2007)[1]

- view the taxonomy as a forest $\mathcal{F}$ of category trees T
- for each category $c \in T$ find the lexical items $heads(c)$ best matching a category label in WordNet:

- **full match**: PLANTS ➠ plant
- **partial match**:
  ICE HOCKEY PLAYERS BY CLUB IN CANADA ➠ ice hockey player

[1] www.eml-research.de/nlp/download/wikitaxonomy.php

Preliminaries

- input: WikiTaxonomy (Ponzetto & Strube, 2007)[1]
- view the taxonomy as a forest $\mathcal{F}$ of category trees T
- for each category $c \in T$ find the lexical items $heads(c)$ best matching a category label in WordNet:
- **full match**: PLANTS ➠ plant
- **partial match**:
  ICE HOCKEY PLAYERS BY CLUB IN CANADA ➠ ice hockey player
- **head match**: EDIBLE PLANTS ➠ plant

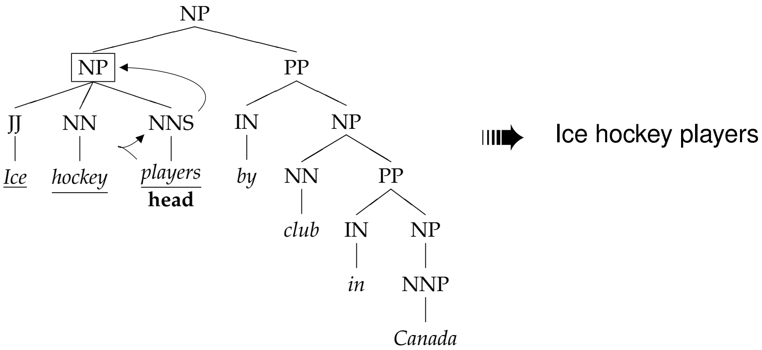
[1] www.eml-research.de/nlp/download/wikitaxonomy.php

Preliminaries

- input: WikiTaxonomy (Ponzetto & Strube, 2007)[1]
- view the taxonomy as a forest $\mathcal{F}$ of category trees T
- for each category $c \in T$ find the lexical items *heads*(c) best matching a category label in WordNet:
- **full match**: PLANTS ➠ plant
- **partial match**: ICE HOCKEY PLAYERS BY CLUB IN CANADA ➠ ice hockey player
- **head match**: EDIBLE PLANTS ➠ plant
- **coordinations**: BUILDINGS AND STRUCTURES IN GERMANY ➠ building structure

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

Finding categories' heads



! try first with a full match, if none can be found:

- parse the category label – using Klein & Manning (2003)
- find the minimal NP projection of the lexical head:
  1. start from the head terminal
  2. percolate up the tree until an NP node is found.
- else fall back to the head itself

Category disambiguation

task definition:

- for each category tree $T \in \mathcal{F}$
  - for each category $c \in T$

find a *mapping* from c to *the most appropriate synset* $\mu_T(c)$

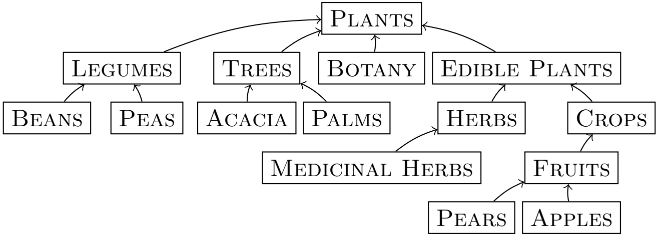
two main steps:

1. WordNet graph construction
2. disambiguation

WordNet graph construction

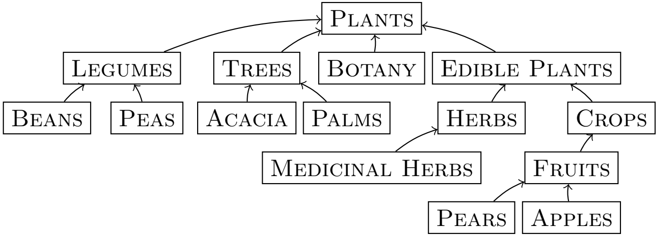
WordNet graph construction

- start with WikiTaxonomy

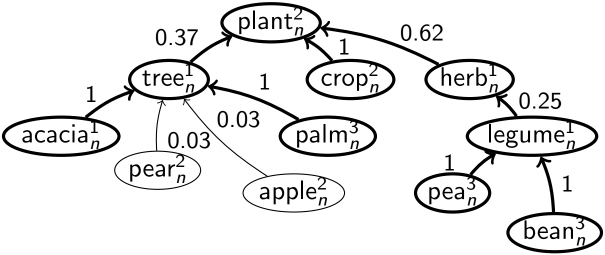


WordNet graph construction

- start with WikiTaxonomy



- create a WordNet graph



WordNet graph construction

```
empty graph G = (V, E)
for all c ∈ T do
    for all h ∈ heads(c) do
        add synsets containing h to V
```

WordNet graph construction

empty graph $G = (V, E)$
for all $c \in T$ **do**
  for all $h \in \mathit{heads}(c)$ **do**
    add synsets containing h to V
for all vertex $v_0 \in V$ **do**
  $v \leftarrow v_0$
  for all synset v', $v \sqsubseteq v'$ **do**
    if v' is root in WordNet **then**
      break
    else if $v' \in V$ **then**
      if $(v, v') \notin E$ **then**
        add (v, v') to E

WordNet graph construction

empty graph $G = (V, E)$
for all $c \in T$ **do**
    for all $h \in heads(c)$ **do**
        add synsets containing h to V
for all vertex $v_0 \in V$ **do**
    $v \leftarrow v_0$
    for all synset v', $v \sqsubseteq v'$ **do**
        if v' is root in WordNet **then**
            break
        else if $v' \in V$ **then**
            if $(v, v') \notin E$ **then**
                add (v, v') to E
            increase the edge weight $w(v, v')$

$$w(v, v') = w(v, v') + \frac{1}{2^{d_{WN}(v_0, v') - 1} \cdot 2^{d_{Wiki}(c_0, c') - 1}}$$

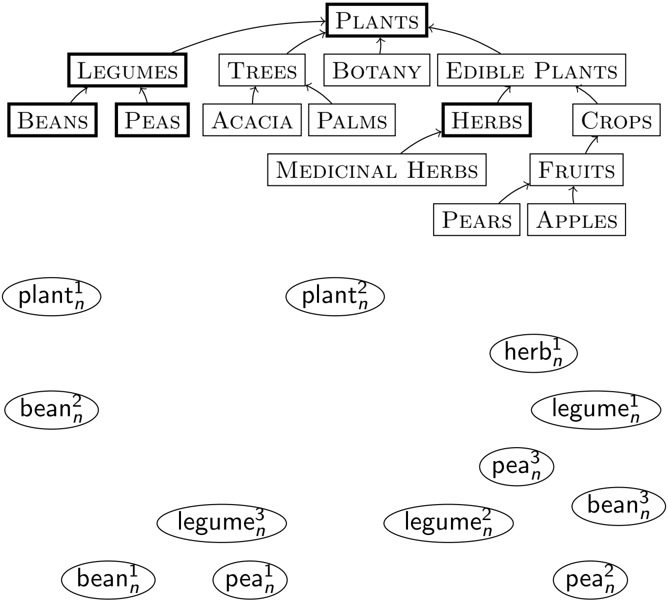
WordNet graph construction

1: empty graph $G = (V, E)$
2: **for all** $c \in T$ **do**
3:     **for all** $h \in heads(c)$ **do**
4:         add synsets containing h to V
5: **for all** vertex $v_0 \in V$ **do**
6:     $v \leftarrow v_0$
7:     **for all** synset v', $v \sqsubseteq v'$ **do**
8:         **if** v' is root in WordNet **then**
9:             **break**
10:         **else if** $v' \in V$ **then**
11:             **if** $(v, v') \notin E$ **then**
12:                 add (v, v') to E
13:             increase the edge weight $w(v, v')$

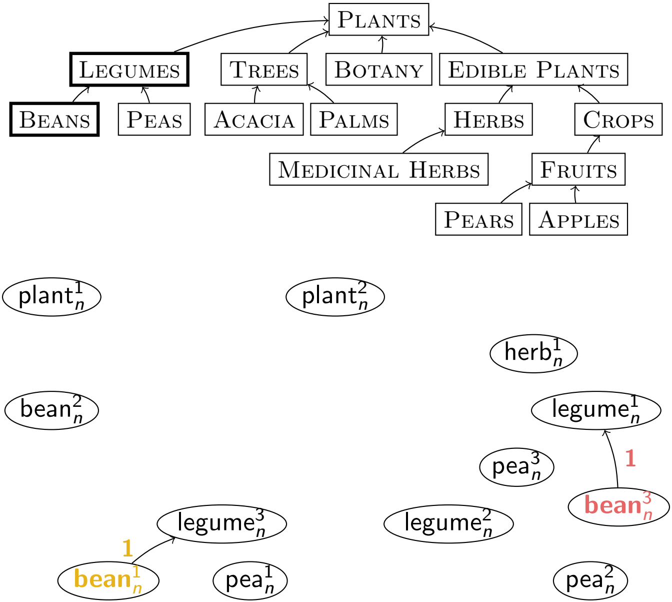
$$w(v, v') = w(v, v') + \frac{1}{2^{d_{WN}(v_0, v') - 1} \cdot 2^{d_{Wiki}(c_0, c') - 1}}$$

14:             $v \leftarrow v'$; **goto (7)**

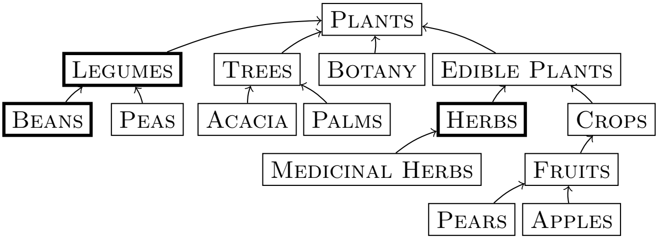
WordNet graph construction

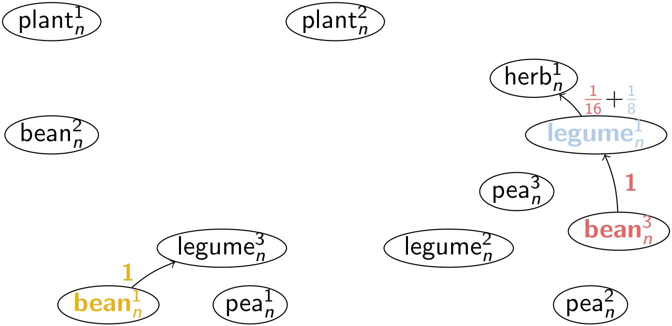


WordNet graph construction

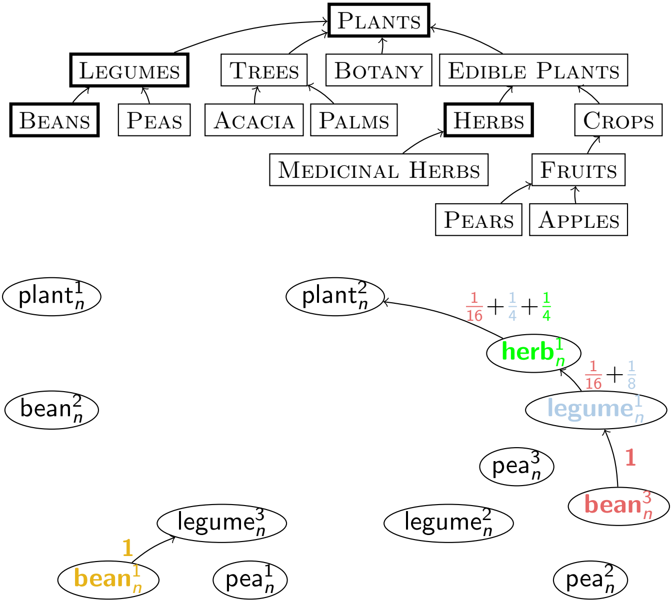


WordNet graph construction

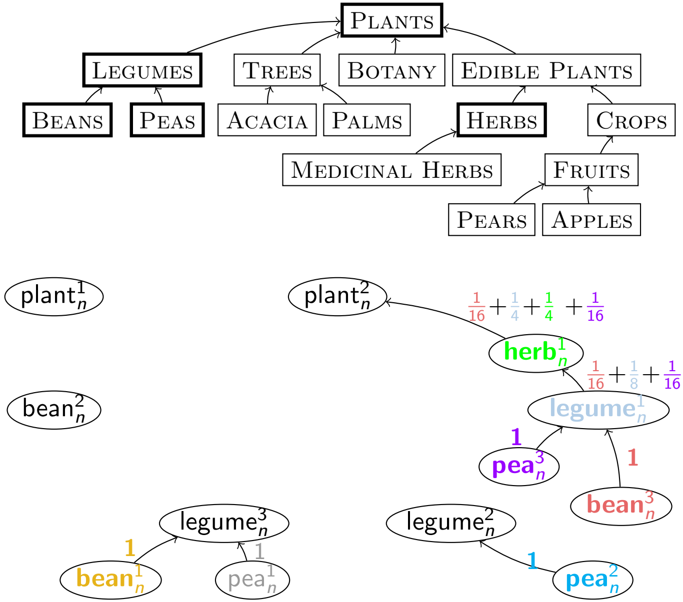




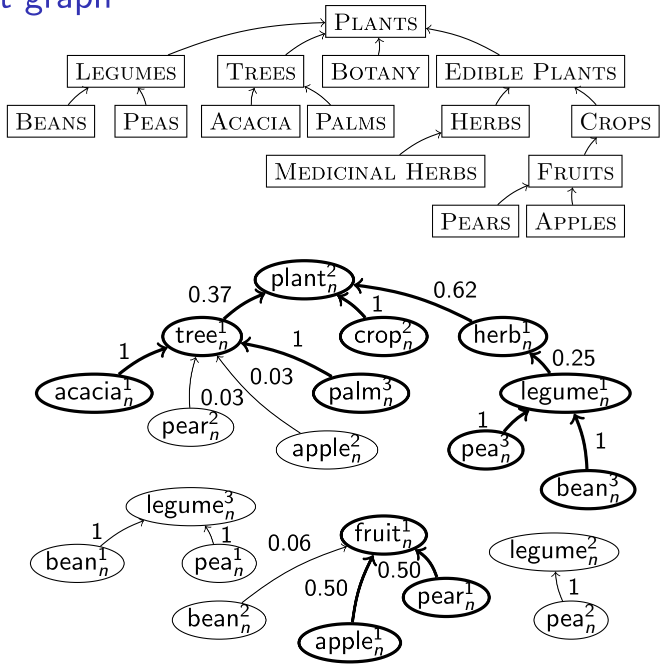
WordNet graph construction



WordNet graph construction



WordNet graph



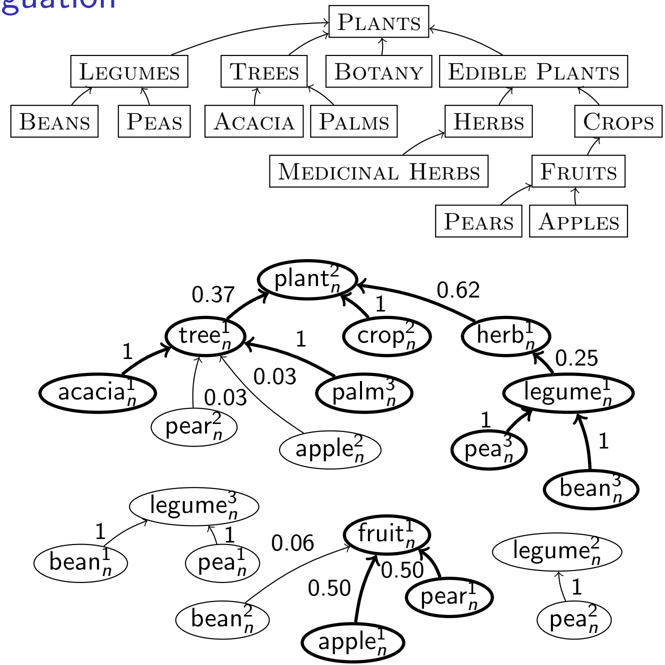
Disambiguation

! use the resulting WordNet graph to identify the most relevant synset for each Wikipedia category $c \in T$

1: sort E in decreasing order based on $w(v, v')$
2: **for all** $(v, v') \in E$ **do**
3:     **if** $\nexists\, \mu_T(c), \mu_T(c')$ **then**
4:         $\mu_T(c) = v$
           $\mu_T(c') = v'$

- in the case of ties, assign the synset which maximizes the size of the connected component of G it belongs to

Disambiguation



Taxonomy restructuring

task definition: *use the mappings* to the reference taxonomy, i.e. WordNet, to *increase the degree of alignment to it*

three main steps:

1. edge penalty weighting
2. identification of maximum penalty cuts
3. tree restructuring

Edge penalty weighting

- find the edges in WikiTaxonomy which are '*misaligned*' with the WordNet *isa* hierarchy (based on the mappings)

for all $T \in \mathcal{F}$ **do**
    for all $e \in T$ **do**
        $p(e) \leftarrow 0$
    for all $c_0 \in T$ **do**
        analyze path $c_0 \rightarrow c_1 \rightarrow \cdots \rightarrow c_n$
        for all (c_i, c_{i+1}) **do**
            if $\neg\, \mu_T(c_0)$ *isa* $\mu_T(c_{i+1})$ **then**
                update penalty p:

$$p(c_i, c_{i+1}) = p(c_i, c_{i+1}) + \frac{1}{2^{d_{Wiki}(c_0, c_{i+1}) - 1}}$$

Edge penalty weighting

- find the edges in WikiTaxonomy which are '*misaligned*' with the WordNet *isa* hierarchy (based on the mappings)
- example:
  - $\textsc{Fruits} \rightarrow \textsc{Crops} \rightarrow \textsc{Edible Plants} \rightarrow \textsc{Plants}$
  - fruit^1_n *notisa* crop^2_n
    - ➠ $p(\textsc{Fruits},\textsc{Crops}) += 1/2^0 = 1$
  - fruit^1_n *notisa* plant^2_n
    - ➠ $p(\textsc{Crops}, \textsc{Edible Plants}) += 1/2^1 = .5$
  - fruit^1_n *notisa* plant^2_n
    - ➠ $p(\textsc{Edible Plants}, \textsc{Plants}) += 1/2^2 = .25$

Identification of maximum penalty cuts

- identify those edges in T with maximal penalty:
  1. sort the edges by penalty
  2. select the subset P_α with the top α percentage of them
     - ➠ 30% based on 10% development data

Identification of maximum penalty cuts

- identify those edges in T with maximal penalty:
  1. sort the edges by penalty
  2. select the subset P_α with the top α percentage of them
     - ➠ 30% based on 10% development data
- example:
  - $P_\alpha = \{$
    $(\textsc{Botany}, \textsc{Plants})$,
    $(\textsc{Fruits}, \textsc{Crops})$,
    $(\textsc{Legumes},\textsc{Plants})$
    $\}$

Tree restructuring

- find a better attachment for each category c among the high-penalty edges $(c, c') \in P_\alpha$ within the entire forest $\mathcal{F}$

1: **for all** $c_i \in P_\alpha = \{(c_1, c'_1) \dots (c_n, c'_n)\}$ **do**
2:     **for all** $c'' \in T'$, $T' \in \mathcal{F}$ **do**
3:         **if** $\mu_T(c)$ *isa* $\mu_{T'}(c'')$ **then**
4:             remove (c, c') from T
               add (c, c'') to T'

Tree restructuring

- find a better attachment for each category c among the high-penalty edges $(c, c') \in P_\alpha$ within the entire forest $\mathcal{F}$

  for all $c_i \in P_\alpha = \{(c_1, c_1') \dots (c_n, c_n')\}$ **do**
      for all $c'' \in T'$, $T' \in \mathcal{F}$ **do**
          if $\mu_T(c)$ *isa* $\mu_{T'}(c'')$ **then**
              remove (c, c') from T
              add (c, c'') to T'

- example:
  - given $\mu_T(\text{LEGUMES}) = legume_n^1$ and $\mu_T(\text{HERBS}) = herbs_n^1$
  - we find legume$_n^1$ *isa* herb$_n^1$ in WordNet
  - ➠ we can move the subtree rooted at LEGUMES under HERBS:
    - ~~LEGUMES → PLANTS~~
      LEGUMES → HERBS

Outline

Evaluation

- evaluation of the two phases

Evaluation

- evaluation of the two phases
- ➠ two questions:
  1. **category disambiguation**: how good is the system at selecting the correct WordNet senses for the Wikipedia category labels?
  2. **taxonomy restructuring**: how good is the restructuring of the taxonomy based on the disambiguated categories?

Evaluation

- evaluation of the two phases
- two questions:
    1. **category disambiguation**: how good is the system at selecting the correct WordNet senses for the Wikipedia category labels?
    2. **taxonomy restructuring**: how good is the restructuring of the taxonomy based on the disambiguated categories?
- proposed evaluation methods:
    1. straight, *in-vitro* manual evaluation
    2. automatic, *instance-based* evaluation

Category disambiguation: manual evaluation

- random sample 2,000 categories from Wikipedia
- annotate them with WordNet synsets (one annotator), e.g.
  - THEATRES IN AUSTRIA $\rightarrow$ theatre$_n^1$
  - THEATRE IN SCOTLAND $\rightarrow$ theatre$_n^2$

Category disambiguation: manual evaluation

- random sample 2,000 categories from Wikipedia
- annotate them with WordNet synsets (one annotator), e.g.
  - THEATRES IN AUSTRIA → theatre^1_n
  - THEATRE IN SCOTLAND → theatre^2_n
- give 310 categories with the five most frequent lexical heads to a second annotator
- quantify quality and difficulty using κ (Carletta, 1996)

➠ $\kappa = 0.92$ (almost perfect agreement)

Category disambiguation: manual evaluation

- random sample 2,000 categories from Wikipedia
- annotate them with WordNet synsets (one annotator), e.g.
  - THEATRES IN AUSTRIA $\rightarrow$ theatre_n^1
  - THEATRE IN SCOTLAND $\rightarrow$ theatre_n^2
- give 310 categories with the five most frequent lexical heads to a second annotator
- quantify quality and difficulty using κ (Carletta, 1996)

➠ $\kappa = 0.92$ (almost perfect agreement)

- **two baselines:**
  1. select a sense at random
  2. select the first (i.e. most-frequent) sense
- **evaluation metric:** accuracy

Category disambiguation: manual evaluation

	tree size			
	2-9	10-100	>100	overall
category disambiguation	62.1	77.7	81.5	80.8
random baseline	36.3	44.2	46.6	46.3
most frequent sense	60.4	69.0	75.2	74.5
# trees	9	65	133	207

Taxonomy restructuring: manual evaluation

- random sample 200 restructuring moves (detachment-attachment pairs)
- check the correctness of the operation:

Taxonomy restructuring: manual evaluation

- random sample 200 restructuring moves (detachment-attachment pairs)
- check the correctness of the operation:
- **correct** if:
  - the original edge *d* is *incorrect* and the *a* is *correct*, e.g.
    ~~ARISTOTLE → CLASSICAL GREEK PHILOSOPHY~~
    ARISTOTLE → PHILOSOPHERS
  - *d* was correct and *a* specializes *d*, e.g.
    ~~BANDLEADERS → MUSICIANS~~
    BANDLEADERS → CONDUCTORS
- else **incorrect**, e.g.
  ~~MANHATTAN → NEW YORK COUNTIES~~
  MANHATTAN → COCKTAILS

Taxonomy restructuring: manual evaluation

- random sample 200 restructuring moves (detachment-attachment pairs)
- check the correctness of the operation:
- **correct** if:
  - the original edge *d* is *incorrect* and the *a* is *correct*, e.g.
    ~~ARISTOTLE → CLASSICAL GREEK PHILOSOPHY~~
    ARISTOTLE → PHILOSOPHERS
  - *d* was correct and *a* specializes *d*, e.g.
    ~~BANDLEADERS → MUSICIANS~~
    BANDLEADERS → CONDUCTORS
- else **incorrect**, e.g.
  ~~MANHATTAN → NEW YORK COUNTIES~~
  MANHATTAN → COCKTAILS
- pairs given to two annotators ($\kappa = 0.75$)

➠ we achieve **accuracy: 88.8%**

Instance-based evaluation

$\boxed{!}$ how good is the system at populating the reference taxonomy with instances?

➠ we can use instances from Wikipedia to *automatically generate two datasets for evaluation*

Instance-based evaluation

$\boxed{!}$ how good is the system at populating the reference taxonomy with instances?

➠ we can use instances from Wikipedia to *automatically generate two datasets for evaluation*

two main steps:

1. instance collection
2. dataset construction

Instance collection

1. use the heuristics from YAGO (Suchanek et al., 2007):

- for each page p of a category $c \in \mathcal{F}$:
  a. split the category label to $\langle pre, head, post \rangle$
  b. assign the relation p *instance-of* c if the lexical head *head* of c is plural.

➠ e.g. AMPHIUMA *instance-of* SALAMANDERS

Instance collection

1. use the heuristics from YAGO (Suchanek et al., 2007):

- for each page p of a category $c \in \mathcal{F}$:
  a. split the category label to $\langle pre, head, post \rangle$
  b. assign the relation p *instance-of* c if the lexical head *head* of c is plural.

➠ e.g. AMPHIUMA *instance-of* SALAMANDERS

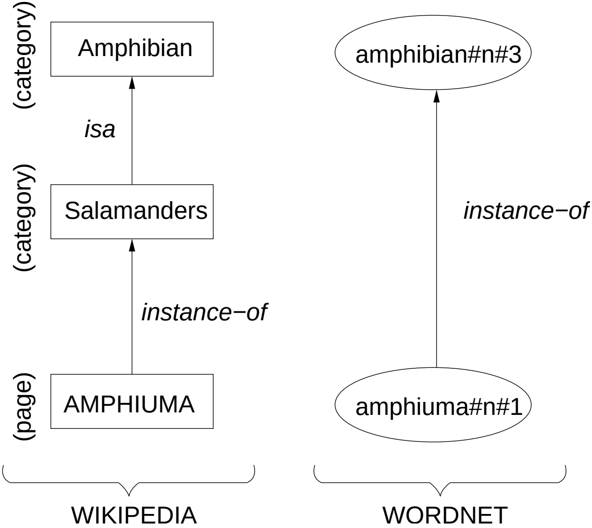
2. filter incorrect instance assignments, e.g. XYLOTHEQUE *instance-of* BOTANICAL GARDENS: check whether p occurs in HeiNER (Wentland et al., 2008)
3. retain instances which are monosemous in WordNet

Dataset construction

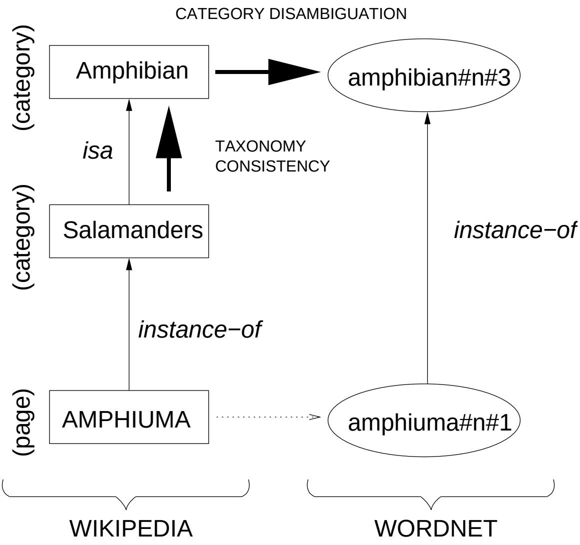
- given a Wikipedia instance i of a category c
  - AMPHIUMA *instance-of* SALAMANDERS
- given its corresponding WordNet synset $\mu_T(c) = S_{c,i}$
  - amphiuma^1_n corresponds to AMPHIUMA

1. identify the WordNet ancestors $S_{c',i}$ of $S_{c,i}$ such that some Wikipedia category c' maps to them
   - amphibian^3_n corresponds to category AMPHIBIANS

Dataset construction



Dataset construction



Instance-based evaluation: results

	before restructuring	after restructuring
category disambiguation	95.3	95.7
random baseline	63.1	63.1
most frequent sense	79.1	78.5
taxonomy consistency	38.4	44.3
# test instances	70,841	73,490

Discussion

$\boxed{!}$ we obtain **high performance figures** on all evaluations

- 80.8% on category disambiguation (*manual* evaluation)
- 88.8% on taxonomy restructuring (*manual* evaluation)

Discussion

! we obtain **high performance figures** on all evaluations

- 80.8% on category disambiguation (*manual* evaluation)
- 88.8% on taxonomy restructuring (*manual* evaluation)

! instance-based evaluation provides a way to automatically build a dataset for evaluating **how good WordNet can be populated with instances from Wikipedia**

- up to 95.7% on category disambiguation (instance-based evaluation)
- ➠ we populate WordNet synsets with Wikipedia instances with **high accuracy**

Discussion

- ! we obtain **high performance figures** on all evaluations
  - 80.8% on category disambiguation (*manual* evaluation)
  - 88.8% on taxonomy restructuring (*manual* evaluation)
- ! instance-based evaluation provides a way to automatically build a dataset for evaluating **how good WordNet can be populated with instances from Wikipedia**
  - up to 95.7% on category disambiguation (instance-based evaluation)
  - ➠ we populate WordNet synsets with Wikipedia instances with **high accuracy**
- ! taxonomy restructuring improves the degree of alignment of WikiTaxonomy to WordNet, thus **recovering from errors**
  - +0.4% on category disambiguation (*instance-based*)
  - +5.9% on taxonomy consistency (*instance-based*)

Outline

Conclusions

- we proposed a **knowledge-rich approach for disambiguating Wikipedia categories to WordNet synsets**
- this mapping can be used to **link the system of categories in Wikipedia to WordNet**

Conclusions

- we proposed a **knowledge-rich approach for disambiguating Wikipedia categories to WordNet synsets**
- this mapping can be used to **link the system of categories in Wikipedia to WordNet**
  - use WordNet as upper-level taxonomy for the Wikipedia category network

Conclusions

- we proposed a **knowledge-rich approach for disambiguating Wikipedia categories to WordNet synsets**
- this mapping can be used to **link the system of categories in Wikipedia to WordNet**
  - use WordNet as upper-level taxonomy for the Wikipedia category network
  - populate WordNet with instances from Wikipedia

Conclusions

- we proposed a **knowledge-rich approach for disambiguating Wikipedia categories to WordNet synsets**
- this mapping can be used to **link the system of categories in Wikipedia to WordNet**
  - use WordNet as upper-level taxonomy for the Wikipedia category network
  - populate WordNet with instances from Wikipedia
  - ➠ get the best of both worlds:
    - fine-grained classification of instances (Wiki)
    - better structured abstract concepts (WordNet)
  - ➠ 'sort-of' WikiTaxonomy 2.0

The big picture . . .

The big picture . . .

Strube & Ponzetto (2006):

- use the category network as a conceptual network

Ponzetto & Strube (2007):

- generate a taxonomy from the network

Ponzetto & Navigli (2009):

- link that network the WordNet

The big picture . . .

Strube & Ponzetto (2006):

- use the category network as a conceptual network

Ponzetto & Strube (2007):

- generate a taxonomy from the network

Ponzetto & Navigli (2009):

- link that network the WordNet

what's next?!

- our approach is resource-independent
  - ➠ apply to other resources, e.g. Cyc
- the backbone of Wikipedia are the articles
  - ➠ disambiguate the pages (cf. Wikification)
- Wikipedia is multilingual
  - ➠ do it for many languages
- find applications
  - ➠ knowledge-lean QA

Thanks!

Acknowledgments

- Roberto Navigli
- Anette and NLP group at SCL
- Michael and NLP group at EML Research

Check out

- ongoing work and papers at
  `http://www.cl.uni-heidelberg.de/~ponzetto`

Berland, Matthew & Eugene Charniak (1999).
Finding parts in very large corpora.
In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, Md., 20–26 June 1999, pp. 57–64.

Caraballo, Sharon A. (1999).
Automatic construction of a hypernym-labeled noun hierarchy from text.
In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* College Park, Md., 20–26 June 1999, pp. 120–126.

Carletta, Jean (1996).
Assessing agreement on classification tasks: The kappa statistic.
Computational Linguistics, 22(2):249–254.

Hearst, Marti A. (1992).
Automatic acquisition of hyponyms from large text corpora.
In *Proceedings of the 15th International Conference on Computational Linguistics,* Nantes, France, 23-28 August 1992, pp. 539–545.

Klein, Dan & Christopher D. Manning (2003).
Fast exact inference with a factored model for natural language parsing.
In Suzanna Becker, Sebastian Thrun & Klaus Obermayer (Eds.), *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, pp. 3–10. Cambridge, Mass.: MIT Press.

Ponzetto, Simone Paolo & Roberto Navigli (2009).
Large-scale taxonomy mapping for restructuring and integrating Wikipedia.
In *Proceedings of the 21th International Joint Conference on Artificial Intelligence,* Pasadena, Cal., 14–17 July 2009.

Ponzetto, Simone Paolo & Michael Strube (2007).
Deriving a large scale taxonomy from Wikipedia.
In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence,* Vancouver, B.C., Canada, 22–26 July 2007, pp. 1440–1445.

Strube, Michael & Simone Paolo Ponzetto (2006).
WikiRelate! Computing semantic relatedness using Wikipedia.
In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July 2006, pp. 1419–1424.

Suchanek, Fabian M., Gjergji Kasneci & Gerhard Weikum (2007).
YAGO: A core of semantic knowledge. unifying WordNet and Wikipedia.
In *Proceedings of the 16th World Wide Web Conference*, Banff, Canada, 8–12 May, 2007, pp. 697–706.

Wentland, Wolodja, Johannes Knopp, Carina Silberer & Matthias Hartung (2008).
Building a multilingual lexical resource for named entity disambiguation, translation and transliteration.
In *Proc. of LREC '08.*