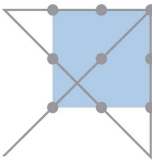


Exploiting Wikipedia for Coreference Resolution

Simone Paolo Ponzetto

EML Research gGmbH, Heidelberg

The Road to the JHU WS 2007

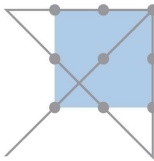


Basic idea (Strube & Ponzetto, AAAI '06):

Use the Wikipedia category system as a semantic network to compute semantic compatibility of mentions

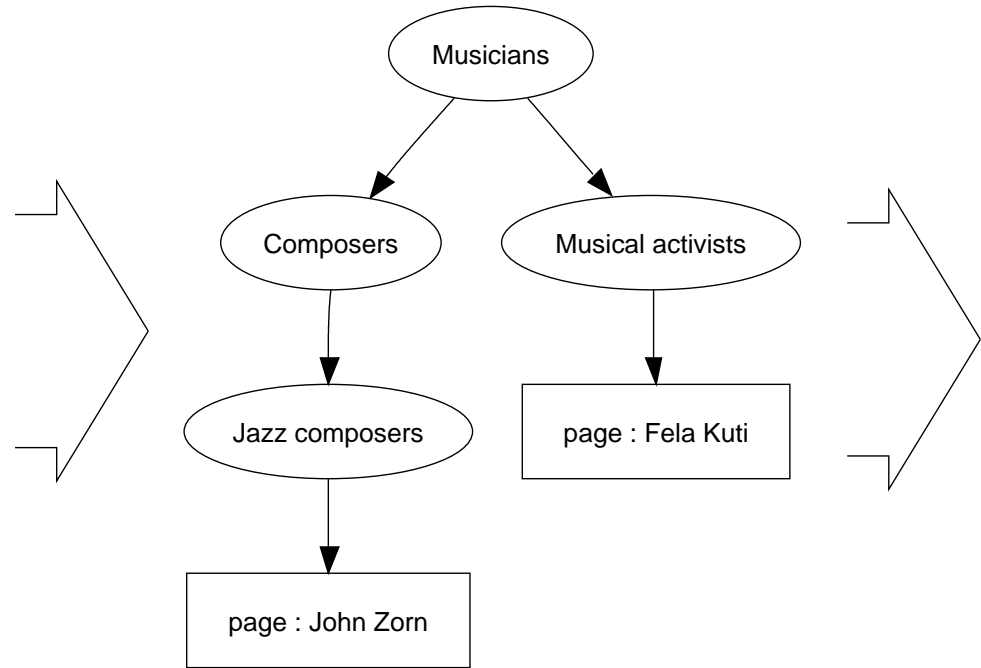
- almost every article is categorized
- **categories** are related in a taxonomic fashion and *can be used as a semantic network* a la WordNet
- ⇒ include semantic distance in the network as an indicator of semantic compatibility of mentions

The Road to the JHU WS 2007



The top screenshot shows the Wikipedia page for John Zorn. In the 'Categories' section, 'Jazz composers' is circled. The bottom screenshot shows the Wikipedia page for Fela Kuti. In the 'Categories' section, 'Musical activists' is circled.

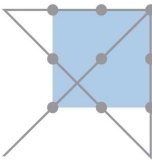
page query and retrieval
category extraction



search for a connecting path
along the category tree

relatedness measure(s) computation

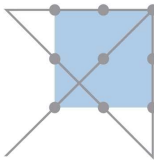
The Road to the JHU WS 2007



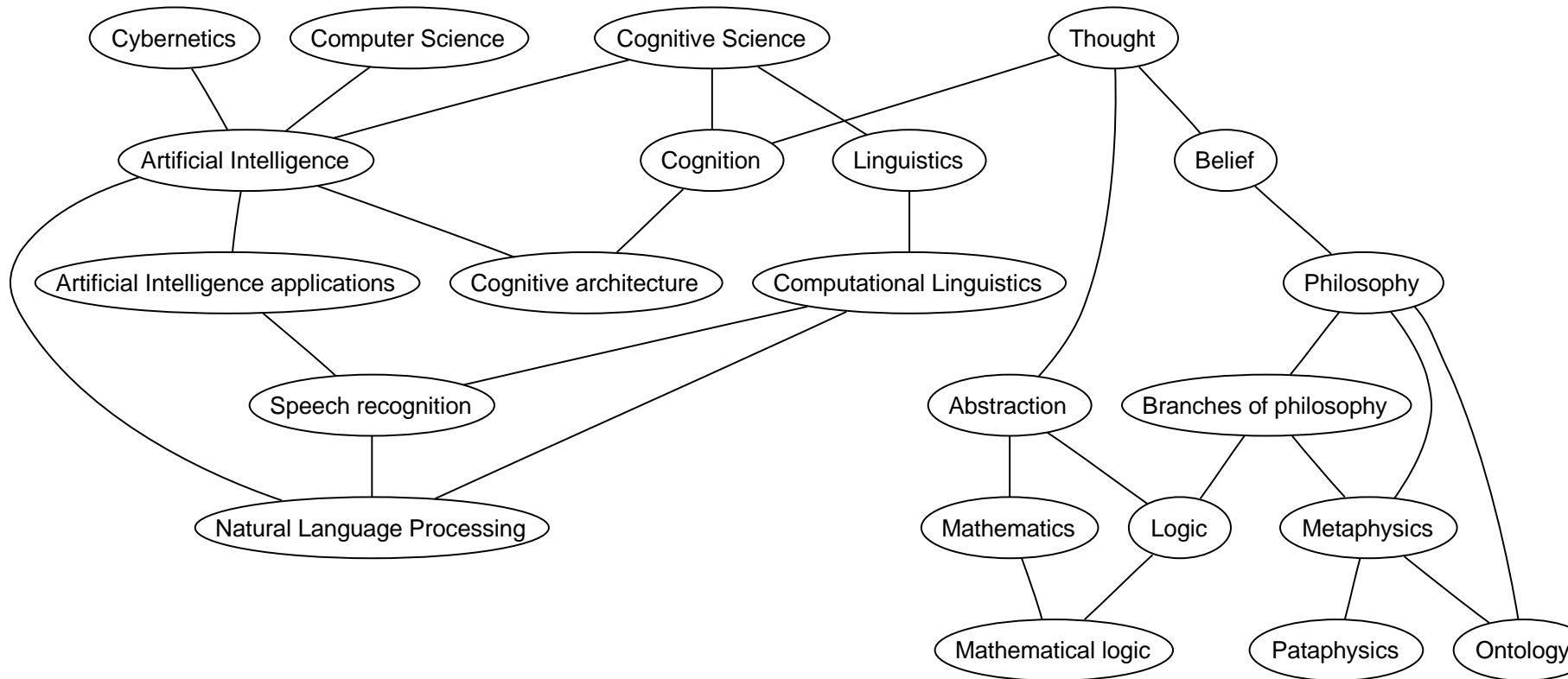
- ▣ we successfully applied Wikipedia-based semantic compatibility to coreference resolution (Ponzetto & Strube, HLT-NAACL '06)
- BUT evaluation was on semi-perfect mentions
- Q: What does happen in a realistic evaluation setting?
- A: Our method simply breaks down :-)
- ▣ things are simply too connected in the Wikipedia category system – e.g. a short connection between **President Clinton** and **Dead Kennedys**

We have to be smarter ...

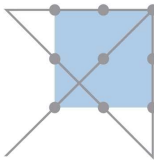
Deriving a taxonomy from Wikipedia



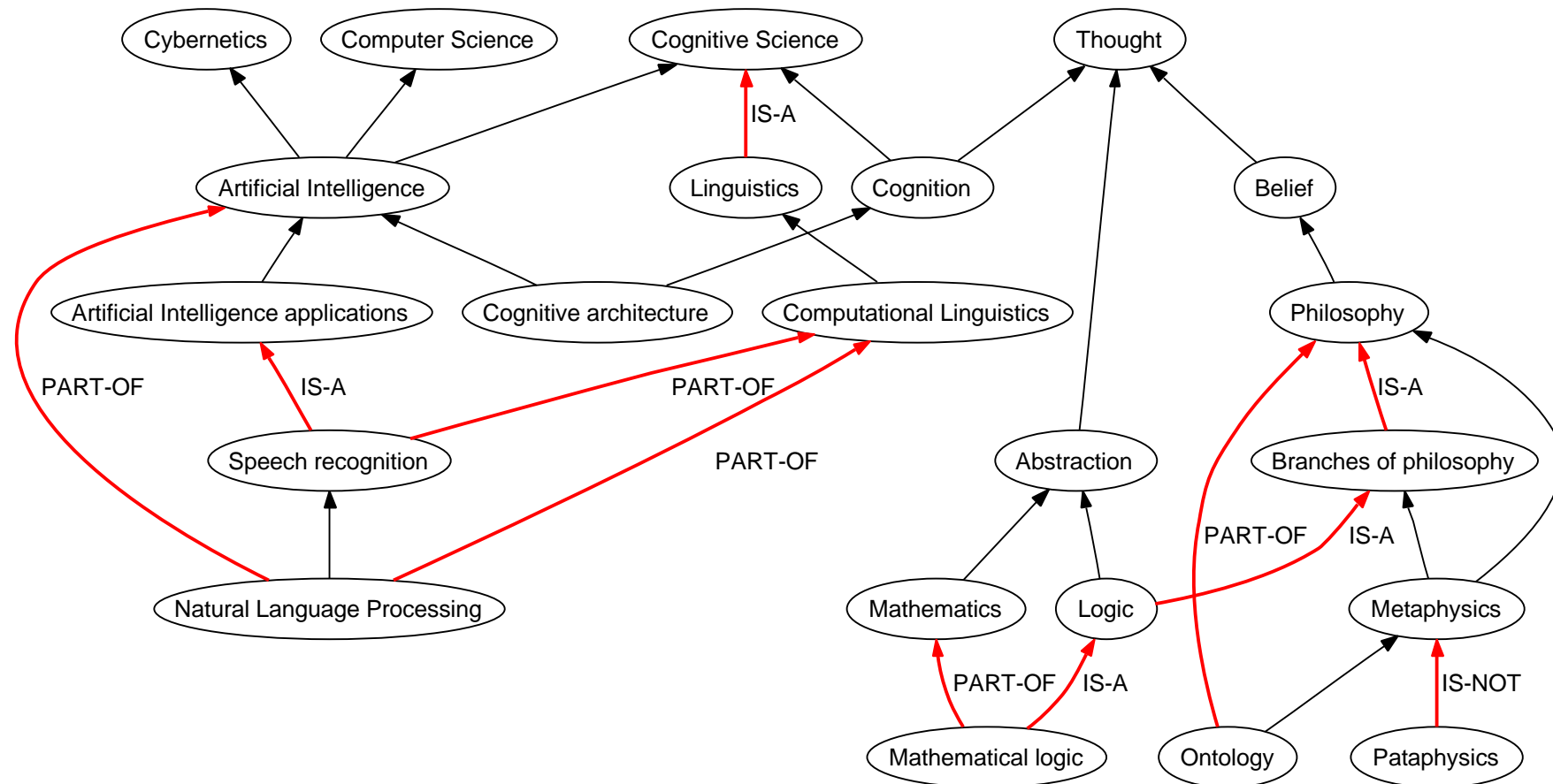
- induce **semantically-typed** category links between the categories in Wikipedia (Ponzetto & Strube, AAAI 2007)



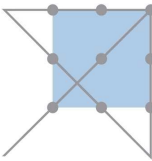
Deriving a taxonomy from Wikipedia



- induce **semantically-typed** category links between the categories in Wikipedia (Ponzetto & Strube, AAAI 2007)



Deriving a taxonomy from Wikipedia



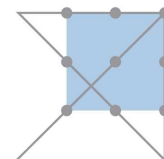
- induce **semantically-typed** category links between the categories in Wikipedia (Ponzetto & Strube, AAAI 2007)
- **connectivity-based methods**: exploit the structure and connectivity in the network
- **lexico-syntactic based methods**: use is-a (Hearst, 1992) and part-of (Berland & Charniak, 1999) surface pattern



performance competitive with WN

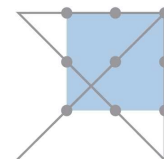
- using the generated taxonomy
- removing semantically coarse-grained categories via PageRank-based filtering

Evaluation of Wiki features (ACE-02)



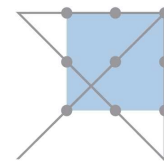
		R	P	F ₁
BNEWS	baseline	51.2	69.9	59.1
	+Wikipedia	56.8	68.0	61.9

Evaluation of Wiki features (ACE-02)



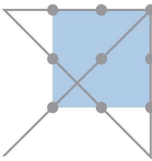
		R	P	F ₁
BNEWS	baseline	51.2	69.9	59.1
	+Wikipedia	56.8	68.0	61.9
NWIRE	baseline	49.9	68.5	57.8
	+Wikipedia	57.6	63.5	60.4

Evaluation of Wiki features (ACE-02)

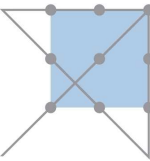


		R	P	F ₁
BNEWS	baseline	51.2	69.9	59.1
	+Wikipedia	56.8	68.0	61.9
NWIRE	baseline	49.9	68.5	57.8
	+Wikipedia	57.6	63.5	60.4
NPAPER	baseline	54.9	73.3	62.7
	+Wikipedia	59.7	68.4	63.7

Discussion



- the results are very similar to the ones from Jason
- we mostly capture **synonyms** – i.e. mentions taking to the *same Wikipedia page* (possibly via cross-disambiguation)
 - **Republican** and **GOP**
 - **CBS** and **The Tiffany Network**
- interestingly, we almost never use the taxonomy
- we have in the end an **Extended Alias** feature



Future directions

- we started by applying WikiRelate!-based semantic compatibility to a realistic setting
- the semantic compatibility features proved too noisy
- noise filtering was obtained by using our recently developed taxonomy resource

we filtered too much in the end!!!

- ➡ we are using the taxonomy as a shortcut
- ➡ we are losing the lexical knowledge that can be derived from taxonomical relations