Lab I

Entity Linking

Daan Odijk
University of Amsterdam

Overview of Lab I

- Entity Linking toolkits and webservices

- Hands-on Entity Linking using webservices
  - Using an entity linking webservice
  - Evaluation of webservice in news

Public Toolkits and Web Services for Entity Linking

- Wikipedia Miner
- TagMe
- DBpedia Spotlight
- Illinios Wikifier
- AIDA

- OpenCalais

Wikipedia Miner

- Open-Source Webservice in Java
  - Hadoop preprocessing pipeline
- Lexical matching
  - Wikipedia page titles and anchor text
- Machine Learning based on Wikipedia articles
  - Prior probability
  - Relatedness: overlap in incoming links with unambigiously linking articles
  - Context quality

D. Milne and I.H. Witten. **Learning to link with Wikipedia**. *CIKM'08*.

TagMe

- Approach similar to Wikipedia Miner
  - Lexical matching
  - Relatedness measure
- Voting for disambiguation
  - Based on all possible bindings
  - Heuristics to select best target
- Designed for short texts

P. Ferragina and U. Scaiella. **Tagme: on-the-fly annotation of short text fragments**. *CIKM'10*.

DBpedia Spotlight

- LingPipe Exact Dictionary-Based Chunker
- Disambiguation in local context
  - Vector-space Model using Bag-of-Words
  - Cosine similarity
- Web Service

P.N. Mendes, M. Jakob, A. García-Silva and C. Bizer. **DBpedia Spotlight: Shedding light on the web of documents**. *I-SEMANTICS'11*.

Illinois Wikifier

- Illinois NER system
- Disambiguation as weighted sum of features
  - Textual similarity
  - Global coherence based on link structure
- Available as download

L. Ratinov and D. Roth. **Design Challenges and misconceptions in named entity recognition**. *CoNLL'09*

AIDA

- Stanford NER Tagger
- Link to YAGO2
- Disambiguation in 3 variants
  - PriorOnly: link to most common target
  - Local: disambiguate individual links with local features
  - CocktailParty: collective disambiguation maximizing coherence using iterative graph-based approach
- Web application

M.A. Yosef, J. Hoffart, I. Bordino, M. Spaniol and G. Weikum. **AIDA: an online tool for accurate disambiguation of named entities in text and tables**. *PVLDB'11*.

OpenCalais

- Only on public content
  - Does not keep a copy of content
  - Keeps a copy of the metadata it extracts
- Free for up to 50,000 documents per day
- Early adopters:
  - CBS Interactive / CNET, Huffington Post, Al Jazeera, The White House
  - More than 30,000 developers, more than 50 publishers

	Programming Language	Service	Available Languages	Open Source
Wikipedia Miner	Java	Web API	EN + any WP	✔
TagMe	Java	Web API	EN, IT	✖
DBpedia Spotlight	Java	Web API	EN + any DBp	✔
Illinois Wikifier	Java	Application	EN	✔
AIDA	Java	Web Application	EN	✔
OpenCalais	?	Web API	EN, FR, SP	✖

	Matching	Target KB	Context	Comment
Wikipedia Miner	Lexical	Wikipedia	ML on Relatedness	
TagMe	Lexical	Wikipedia	Vote on Relatedness	Focus on Short texts
DBpedia Spotlight	Lexical?	DBpedia	Cosine Similarity	Structure
Illinois Wikifier	NER	Wikipedia	Global Coherence	
AIDA	NER	YAGO2	Multiple	Structure
OpenCalais	?	Calais	?	

Benchmark of Entity Linking Toolkits and Webservices

- Benchmarking framework
- Five annotated datasets:
  - News articles (AIDA/CoNLL, AQUINT, MSNBC)
  - Tweets (Meij)
  - Web pages (IITB)
- Evaluation Measures
  - Fuzzy matching
  - Different settings

M. Cornolti, P. Farragina and M. Ciaramita. **A Framework for Benchmarking Entity-Annotation Systems**. *WWW'13*.

Benchmark of Entity Linking Toolkits and Webservices

- Lexical Matching vs NER
  - NER-based system: high precision, but low recall
  - Lexical matching performs better overall
- Wikipedia as target KB performs better
- Best disambiguation differs per measure
- Training and tuning for specific setting helps
- Substantial differences in runtime

M. Cornolti, P. Farragina and M. Ciaramita. **A Framework for Benchmarking Entity-Annotation Systems**. *WWW'13*.

Overview of Lab I

- Entity Linking toolkits and webservices

- Hands-on Entity Linking using webservices
  - Using an entity linking webservice
  - Evaluation of webservice in news

AQUAINT Dataset

- 50 documents out of original AQUAINT dataset
  - English newswire
- Annotated with links to Wikipedia
  - Wikipedia-style linking: first links of most important
  - 14.5 links per document on average

D. Milne and I.H. Witten. **Learning to link with Wikipedia**. *CIKM'08*.