HTW Chur
Hochschule für Technik und Wirtschaft
University of Applied Sciences

webLyzard technology

Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence

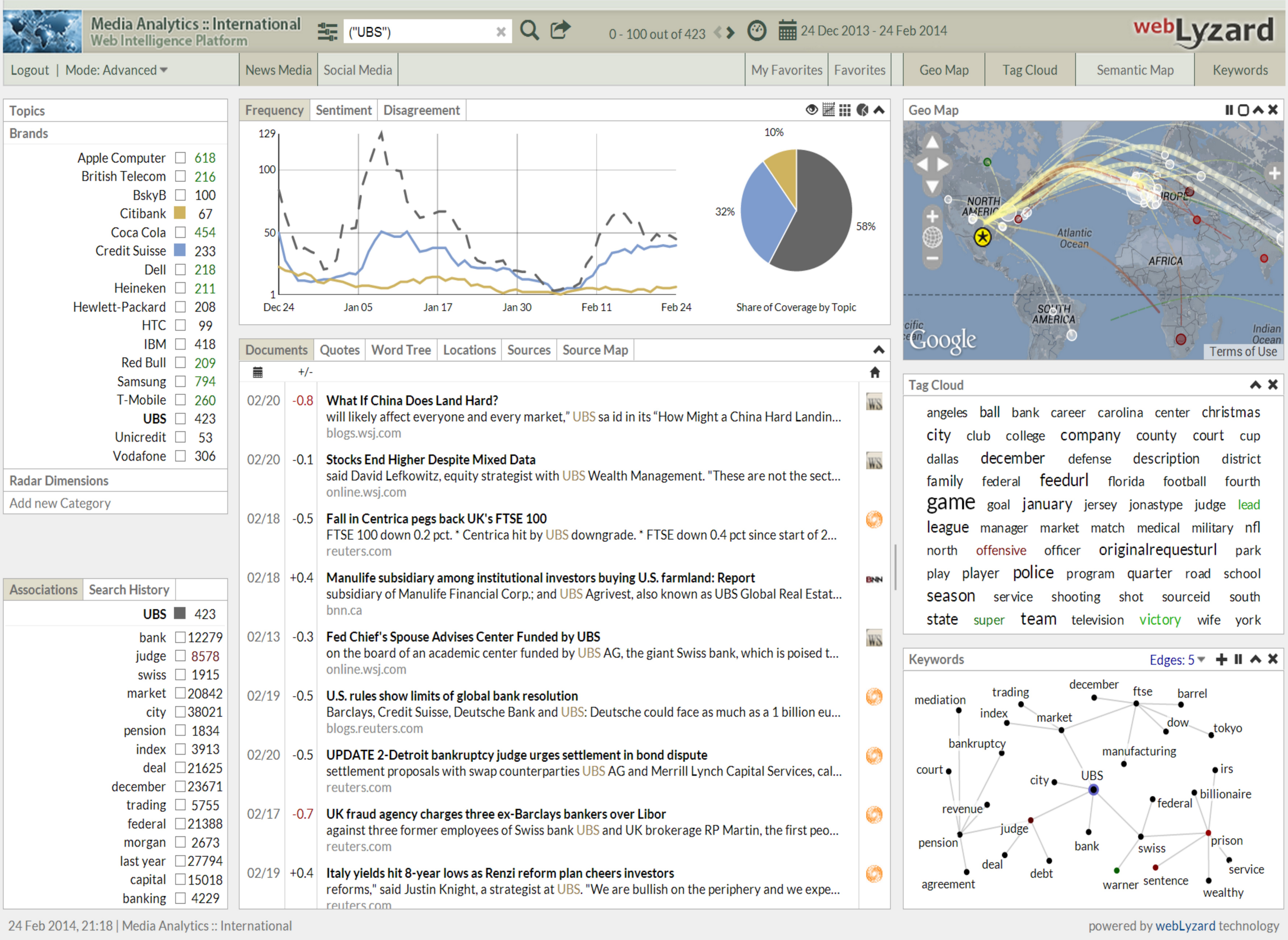
Albert Weichselbraun

Agenda

1. Introduction
2. Datasets: Linked Enterprise Data
3. Method
    - Major challenges
    - Pre-processing
    - Disambiguation
4. Evaluation
5. Outlook and conclusions

Introduction

- Named Entity Linking
- Key issues
  - extraction of company names and useful context and structural information from linked data sources
  - mention generation and assessment
  - disambiguation – locate mentions in text documents and ground them to the corresponding entities in the linked data source



Topics

Brands

Apple Computer	□	618
British Telecom	□	216
BskyB	□	100
Citibank	■	67
Coca Cola	□	454
Credit Suisse	■	233
Dell	□	218
Heineken	□	211
Hewlett-Packard	□	208
HTC	□	99
IBM	□	418
Red Bull	□	209
Samsung	□	794
T-Mobile	□	260
UBS	□	423
Unicredit	□	53
Vodafone	□	306

Radar Dimensions

Add new Category

Associations | Search History

UBS	■	423
bank	□	12279
judge	□	8578
swiss	□	1915
market	□	20842
city	□	38021
pension	□	1834
index	□	3913
deal	□	21625
december	□	23671
trading	□	5755
federal	□	21388
morgan	□	2673
last year	□	27794
capital	□	15018
banking	□	4229

02/20 -0.8 **What If China Does Land Hard?**
will likely affect everyone and every market," UBS sa id in its "How Might a China Hard Landin...
blogs.wsj.com

02/20 -0.1 **Stocks End Higher Despite Mixed Data**
said David Lefkowitz, equity strategist with UBS Wealth Management. "These are not the sect...
online.wsj.com

02/18 -0.5 **Fall in Centrica pegs back UK's FTSE 100**
FTSE 100 down 0.2 pct. * Centrica hit by UBS downgrade. * FTSE down 0.4 pct since start of 2...
reuters.com

02/18 +0.4 **Manulife subsidiary among institutional investors buying U.S. farmland: Report**
subsidiary of Manulife Financial Corp.; and UBS Agrivest, also known as UBS Global Real Estat...
bnn.ca

02/13 -0.3 **Fed Chief's Spouse Advises Center Funded by UBS**
on the board of an academic center funded by UBS AG, the giant Swiss bank, which is poised t...
online.wsj.com

02/19 -0.5 **U.S. rules show limits of global bank resolution**
Barclays, Credit Suisse, Deutsche Bank and UBS: Deutsche could face as much as a 1 billion eu...
blogs.reuters.com

02/20 -0.5 **UPDATE 2-Detroit bankruptcy judge urges settlement in bond dispute**
settlement proposals with swap counterparties UBS AG and Merrill Lynch Capital Services, cal...
reuters.com

02/17 -0.7 **UK fraud agency charges three ex-Barclays bankers over Libor**
against three former employees of Swiss bank UBS and UK brokerage RP Martin, the first peo...
reuters.com

02/19 +0.4 **Italy yields hit 8-year lows as Renzi reform plan cheers investors**
reforms," said Justin Knight, a strategist at UBS. "We are bullish on the periphery and we expe...
reuters.com

Tag Cloud
angeles ball bank career carolina center christmas
city club college company county court cup
dallas december defense description district
family federal feedurl florida football fourth
game goal january jersey jonastype judge lead
league manager market match medical military nfl
north offensive officer originalrequesturl park
play player police program quarter road school
season service shooting shot sourceid south
state super team television victory wife york
Keywords
Edges: 5
mediation
trading
december
ftse
barrel
index
market
dow
tokyo
bankruptcy
manufacturing
court
city
UBS
irs
billionaire
federal
revenue
judge
pension
bank
swiss
prison
deal
debt
service
agreement
warner
sentence
wealthy

24 Feb 2014, 21:18 | Media Analytics :: International

powered by webLyzard technology

Linked Enterprise Data

- Orell Füssli Business Information AG
  → Switzerland's largest provider of business inf.
- Linked Enterprise Repository
  - based on a number of business databases
  - comprises 2.9 million companies and background information (names, key people, products, contact information, brand names, turnover, …)
  - removal of duplicates and inactive companies
  - conversion to linked data using well known name spaces
    → 570,000 organizations; 9 million triples

Linked Enterprise Data | Example

```
# teledata database
teledata-company:775 rdf:type owl:Company.
teledata-company:775 rdfs:label "American Optical Company
                                 International AG".   teledata-
company:775 rdfs:label "Carl Zeiss Vision AG".
teledata-company:775 dbpedia-owl:numberOfEmployees "35".
teledata-company:775 dbprop-de:umsatz "4183400.0".
teledata-company:775 ofwi:company-status "active".
teledata-company:775 dbpedia-owl:industry
                       ofwi-industry:8962, ofwi-industry:7752.
teledata-company:037041 schema-org:address ofwi-address:037041.


# kompass database
ofwi-company:037041 rdf:type dbpedia-owl:Company.
ofwi-company:037041 rdfs:label "Carl Zeiss Vision Swiss AG"@de.
ofwi-company:037041 dbpedia-owl:abstract
                            "Zweck der Gesellschaft ist..."@de.
ofwi-company:037041 owl:sameAs teledata-company:775.
```

Linked Enterprise Data | Example

```
# contact and legal information
ofwi-company:037041 dbprop-de:unternehmensform
                                  dbpedia-de:Aktiengesellschaft.
ofwi-company:037041 schema-org:email "office@zeis.com".
ofwi-company:037041 schema-org:faxNumber "055254473730".
ofwi-company:037041 schema-org:telephone "0552547373".
ofwi-company:037041 schema-org:email "info.swiss@vision.zeiss.com".
ofwi-company:037041 schema-org:url "http://www.vision.zeiss.ch".

# keywords regarding the company's products and services
ofwi-company:037041 dbprop:products ofwi-productgroup:38371, ...
ofwi-company:037041 dbprop:distributor "Teflon easycare",
                                       "i.Profiler", "Carl Zeiss".

# key people
ofwi-company:037041 dbprop:keyPeople ofwi-person:Peter_Däpp_(0432);
                    dbprop:keyPeople ofwi-person:Sven_Hermann_(0341).
```

Linked Enterprise Data | Example

```
# address information
ofwi-address:037041 rdf:type schema-org:PostalAddress
ofwi-address:037041 schema-org:addressCountry "CH".
ofwi-address:037041 schema-org:addressRegion "ZH".
ofwi-address:037041 schema-org:postalCode "8714".
ofwi-address:037041 schema-org:addressLocality "Feldbach".
ofwi-address:037041 schema-org:streetAddress "Feldbacherstrasse 81".

# product groups
ofwi-productgroup:38371   rdfs:label "Optische Linsen", "Gläser",
                                     "Spiegel".
ofwi-productgroup:3837122 rdfs:label "Brillengläser".

# industry mapping
ofwi-industry:8962 rdf:label "Wholesale of photographic and .."@en;
                   rdf:label "Commercia all'ingrosso di..."@it;
                   rdf:label "Commerce de gros d'appareils ...."@fr;
                   rdf:label "Grosshandel mit Foto ..."@de.
```

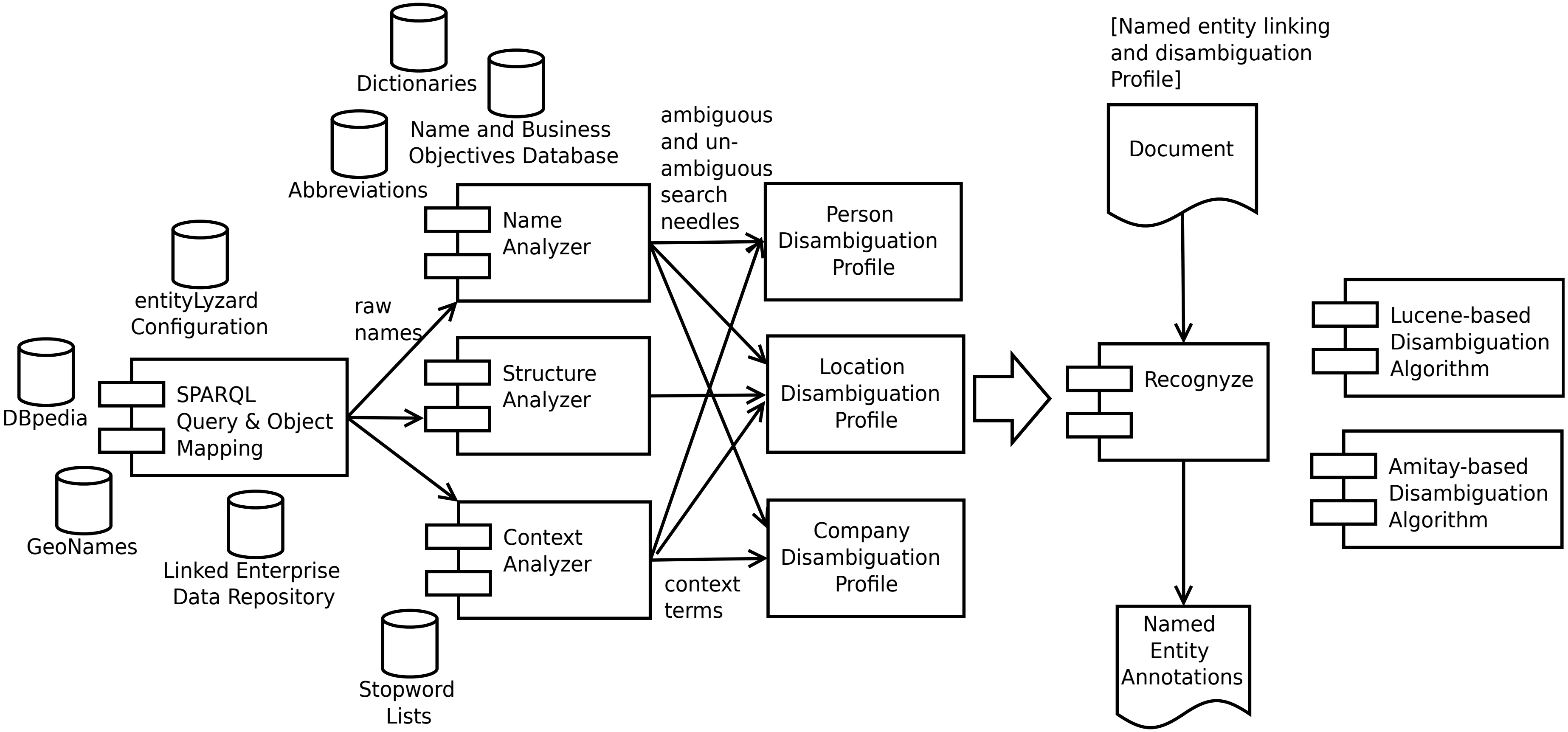
Challenges

ID	Description	Example
1	*Data quality*	
1.1.	ambiguous short names	Aktien (shares), Hell (bright), ...
1.2.	uppercase only company names	DER SA, DER HEIZER
2	*ambiguities*	
2.1.	many very small companies	1300 x Meyer, 1018 x Personalfürsorgestiftung
2.2.	legally related companies	83 x Credit Suisse, 92 x UBS
2.3.	Similar company names with no or little metadata	ABSOLUT, ABSOLUT SA, ABSOLUT COSMETICS

Challenges

ID	Description	Example
3	*Smaller data granularity*	
3.1.	Ambiguous company names	13 x IST (be); 0 x @DBpedia WEG (way)
3.2.	Ambiguous person names	Robert Frey vs. Rober Frey Consulting
4	Use of casual name forms	
4.1.	Short names	IST AG rather than Innovative Sensort Technology AG
4.2.	Use of „insider“ casual names	Sonova (Phonak Sounds AG) CS (Credit Suisse)

Method | System Architecture



Method | Name Analyzer

- Extract potential mentions from linked data
  - Entity: UBS Financial Service Basel AG
  - Possible mention Strings: UBS, UBS Financial Service, …
- Entropy-based metric

$$H(m) = H_{init}(m) + H_C(C_{ij}) + \sum_{t_k \in T_{ij}} H_t(t_k, caseSensitive(m)) + H_{compl}$$

- Entropy threshold determines minimum mention length and ensures "complete" names
- Entries below the threshold are disambiguated using prefixes and/or suffixes

Method | Context & structure analyzer

- structural information
  - related companies and subsidiaries
  - company's management
  - address information
- context information
  - products and services offered by the company
  - industry
  - revenue and number of employees

Method | Recognyze Profiles

field	value
label	eval.5.context
source	http://../r/de.dbpedia.org

query	**SELECT** ?s ?companyName ?abstract ?homepage ?foundingDate ?industry ?city ?country ?keyPeople ?tickerName
	WHERE {
	?s rdf:type dbpedia-owl:Company .
	?s rdfs:label ?companyName .
	OPTIONAL {
	?s prop-de:sitz ?city .
	}
	FILTER (**LANG**(?companyName) = 'de')
	...
	}

Method | Recognyze Profiles

field	value
entity type	Recognyze.OrganizationEntity
disambiguation algorithm	Lucene similarity
pre-processor	*binding*: ?companyName *handler*: OrganizationNameHandler
...	...
filter	*scope: name* dict=dict.C, dict.de_CH, dict_de, dict.en
filter	*scope: context* dict=stopwords.C, stopwords.de, stopwords.en, ...
...	...
affix filter (disambiguation)	Recognyze.OrganizationAffix

Method | Disambiguation

- Geo → Amitay
- Organizations → modified Lucene similarity
- ambiguous mentions are disambiguated using prefix and suffix terms

$$s(e_i, d) = f_c(mentions_e, d) \cdot |mentions_e| \sum_{t \in mentions_e} [idf(t^2) \cdot boost(t)]$$

- ranking is refined by using weights obtained from context information (number of employees and turnover)

Evaluation | Corpora

- extended AWP.ch news dataset
  - 320,000 manually annotated news messages
  - 150 randomly selected German-speaking news messages
  - annotations of *all* covered companies which have been manually confirmed by domain experts
- NZZ (Neue Zürcher Zeitung) news dataset
  - 150 randomly selected NZZ business news articles manually annotated by domain experts

Evaluation | Setting

- raw names
  → extracted names "as is"
- simple
  → tokenize names and generate standardized alternative names (e.g. I.B.M. > IBM, ...)
- advanced
  → full Recognyze name pre-processing

Evaluation | Estimated Coverage

Setting	Rescore	AWP messages R	NZZ articles R
raw names		0.52	0.13
	√	0.52	0.13
simple		0.95	0.95
	√	0.81	0.66
advanced		0.87	0.81
	√	0.83	0.76

Evaluation | Linking Performance

Setting	Rescore	AWP messages P | R | F1	NZZ articles P | R | F1
raw names		0.44 0.52 0.44	0.14 0.13 0.11
	√	0.49 0.52 0.47	0.16 0.13 0.13
simple		0.07 0.52 0.10	0.03 0.45 0.06
	√	0.09 0.61 0.14	0.04 0.55 0.07
advanced		0.36 0.71 0.41	0.38 0.75 0.46
	√	0.50 0.80 0.59	0.60 0.74 0.63

Outlook and conclusions

- Recognyze draws upon linked data sources
  → no learning step involved
  → tested with OFWI linked enterprise data and Dbpedia
- Data pre-processing considerably improves the component's performance
- Future work will focus on
  - adding support for additional named entity types (people and events)
  - improved extraction of contextual information (e.g. obtain abbreviations from Dbpedia abstracts)
  - create easy ways to create and share Recognyze profiles