Leveraging and Balancing Heterogeneous Sources of Evidence in Ontology Learning

Gerhard Wohlgenannt [1] wohlg@ai.wu.ac.at

[1] Institute f. Information Business, **Vienna Univ. of Economics (WU)**, Austria

ESWC 2015, Portoroz
June 3, 2015

Overview – What to expect

General aim: Evaluate how to best use information from **multiple and heterogeneous sources of evidence** in ontology learning – to improve **system accuracy**

- Starting with basic concepts: Ontology learning, a description of our system and the evidence sources used
- Experiments to address research questions. **Influence of**:
  - How many evidence sources used?
  - How much evidence per source?
  - Source quality
  - ...
- Conclusions

Introduction & Concepts

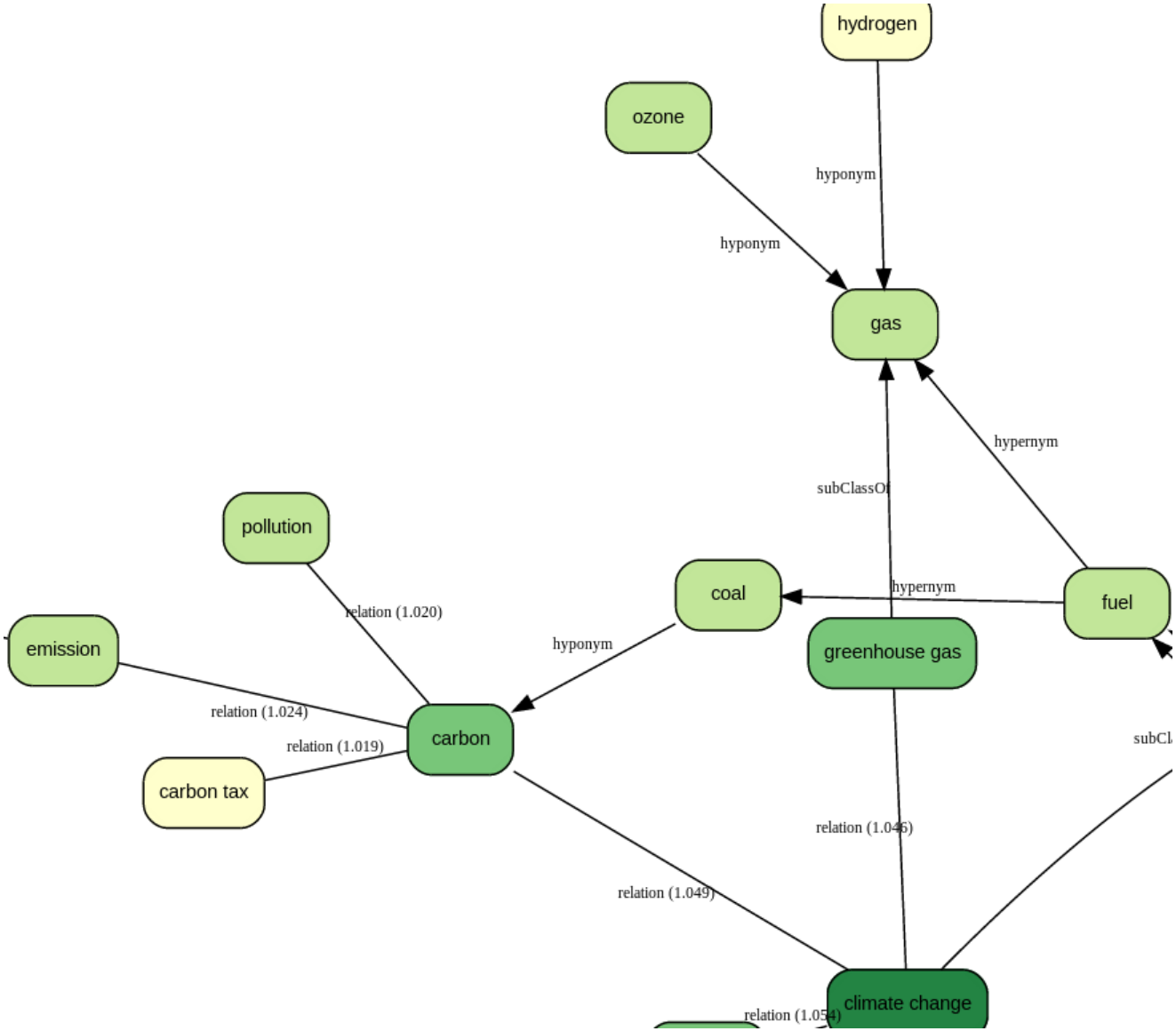
- Ontologies: Provide vocabulary used on the Semantic Web.
- Ontology construction expensive $\rightarrow$ methods such as Ontology Learning and Crowdsourcing to bootstrap $\rightarrow$ make more the process scalable and cheaper
- Ontology Learning: Use of supervised and unsupervised methods to (semi-) automatically generate an ontology from data
- Ontology Learning traditionally from **one source**, usually a domain text corpus (simplified)
- Here: Many sources, how integrate and balance them?

Our Ontology Learning System

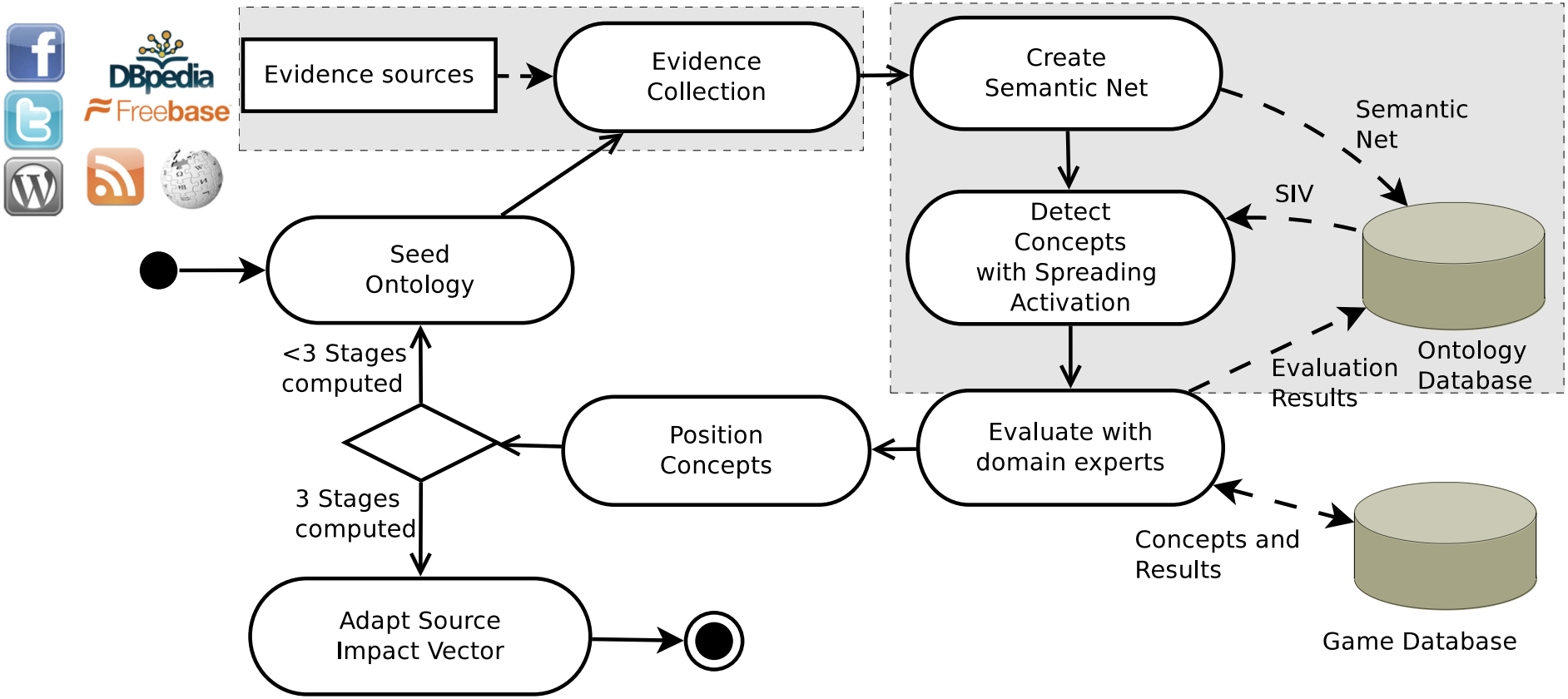
What it does: Extend ontologies

1. Seed ontology
2. Collect evidence for new concepts
3. Determine new concepts and their position
4. $\rightarrow$ extended ontology (light-weight)

Example of an Extended Ontology



System Diagram

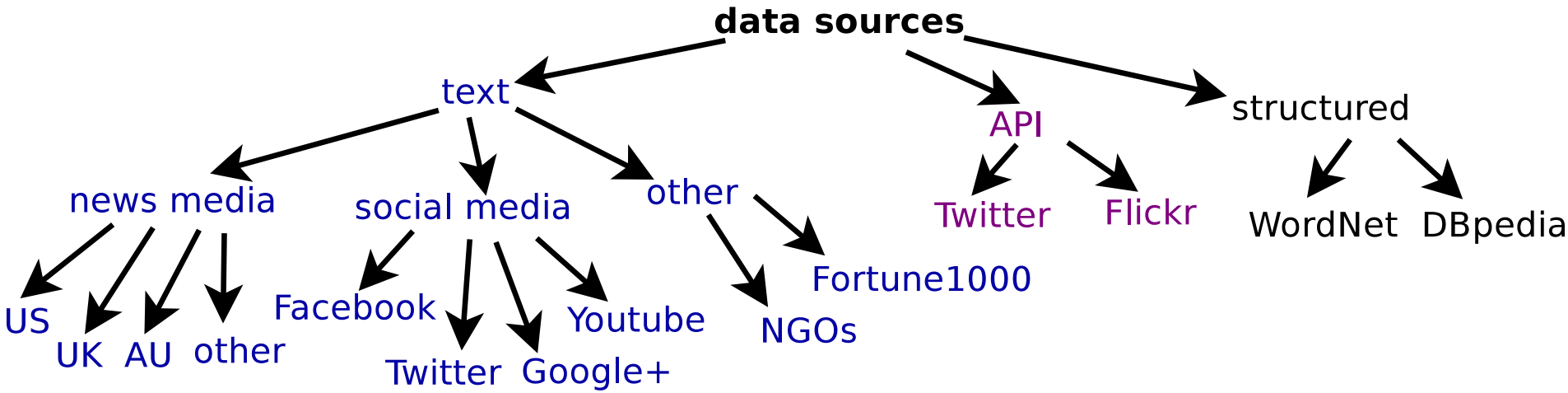


Evidence Sources

- For given **seed concepts** $\rightarrow$ provide (domain) terms and relations to the system
- In current configuration mostly based on **domain text** (keyword extraction, etc), but also **social media**, and **structured sources**

- **32 evidence sources** – of heterogeneous quality, number of evidences, and type of source

Data Sources



Data is collected (mirrored) every month to generate new ontologies from scratch (monthly).

Heterogeneous Evidence Sources – Part 1

	Method		
Data sources			
domain text from:	Keyw./page	Keyw./sent.	Hearst patterns
US news media	1	2	3
UK news media	4	5	6
AU/NZ news media	7	8	9
other news media	10	11	12
Social media: Twitter	13	-	14
Social media: Youtube	15	-	16
Social media: Facebook	17	-	18
Social media: Google+	19	-	20
NGOs Websites	21	22	23
Fortune 1000 Websites	24	25	26

Table : The 26 evidences sources used in the ontology learning process based on domain text. The data is collected from the Web to create corpora in monthly intervals.

Heterogeneous Evidence Sources – Part 2

	Method				
Data source:	hypernyms	hyponyms	synonyms	API	SPARQL
WordNet	27	28	29	-	-
DBpedia	-	-	-	-	30
Twitter	-	-	-	31	-
Flickr	-	-	-	32	-

Table : The remaining 6 evidence sources, which are based on WordNet, Social Media APIs, and DBpedia.

Example of Evidence – Keywords for “CO2”

First 12 terms found – sorted by significance . . .

Term	Significance	Term	Significance
carbon price floor	164.85	emission	110.48
sec	135.54	air	99.99
fertilisation	133.63	waste	90.17
PM10	123.45	0-62mph	89.12
environment committee	121.27	flame	86.74
member state	114.62	carbon tax	78.53

Table : Example evidence (keywords and their χ^2 co-occurrence significance) for the seed concept “CO2”.

Concept Selection

- What happens?
- All collected evidence forms a semantic network, with typically $>$ 20000 labelled links between thousands of terms
- **Next step**: Select 25 concept candidates from huge number of terms
- **How?**: Spreading activation – a technique for neural/associative networks

Research Questions

- Multiple sources provide **redundancy and complementary information**.
- General assumption: Redundancy of information in different sources represents a measure of relevance and trust (Manzano-Macho et al., 2008)
- Heterogeneous sources offer the potential for higher levels of accuracy – we mainly look at the concept detection phase
- Research questions:
  - How many sources?
  - How much evidence per sources?
  - Effect of source characteristics (quality, heterogeneity)? . . .

Method / Goal

- Lots of Experiments: . . . with different settings for number of sources, evidences per source, etc.
- **Goal:** try to find answers to (ontology) learning scenarios that can be generalized (at least to some point)
- As we use a simple and intuitive evidence integration logic (spreading activation) – except similar results with other integration logics

Evaluation setup

- 2 domains: **climate change** and **tennis**
- Try different settings for Number of sources and Evidences used per source $\rightarrow$ generated ontologies for all those settings in every month between July 2013 and November 2014.
- Assessment of accuracy done by domain experts

$$Accuracy = \frac{Relevant\ concept\ candidates\ generated}{All\ concept\ candidates\ generated} \quad (1)$$

Why balancing needed anyway??

Quantity and Quality of Evidence per Source and Seed Concept

Method:	Avg. Num. of Evid.	Top 25	Top 100	Top 500
Keywords/page	400	0.31	0.26	0.12
Keywords/sent.	200	0.27	0.19	0.10
Hearst Patterns	18		0.15	
API Twitter	70		0.10	
API Flickr	16		0.18	
WordNet (Hyper)	15		0.24	
WordNet (Hypo.)	17		0.21	
DBpedia	13		0.27	

Table : Average number of evidence (per source and concept) and evidence quality (domain relevance) per extraction method.

Experiments

- Accuracy Regarding Number of Evidences Used per source and concept: balance number of evidences per source, save computation time, but maybe loose helpful data
- Accuracy Regarding Number of Sources: How many sources need to benefit from heterogeneity / multiple sources
- Accuracy Regarding Number of Seed Concepts: similar to number of sources

Accuracy Regarding Number of Evidences Used

No. Evidences	Acc. CC	Acc. Tennis	Acc. Rand. Keyw. CC
limit=5	56.44	46.80	52.72
limit=10	64.05	55.53	56.51
limit=20	67.57	**60.27**	60.98
limit=50	**68.68**	59.87	61.64
limit=100	67.79	58.27	62.73
limit=200	67.87	58.53	65.13
limit=500	66.39	57.88	66.01
no limit	66.29	57.34	**66.29**

Table : Accuracy of concept detection (percentage of relevant concept candidates) for the domains of climate change (CC) and tennis – depending on evidences per source and concept.

Accuracy Regarding Number of Sources

%Relevant	1 (Twitter)	1 (UK-Keyw.)	5 srcs	15 srcs	32 srcs
CC limit=50	16.54	48.80	59.52	**68.28**	68.84
CC limit=200	19.85	49.78	57.48	**67.73**	67.64
Tennis limit=50	21.15	50.67	52.25	**56.88**	57.87
Tennis limit=200	23.17	52.78	54.33	**57.74**	58.33

Table : Accuracy of concept detection regarding the number of evidence sources ("srcs") used – for two limit-settings.

Accuracy Regarding Number of Seed Concepts

	Stage1 - 2 SC	**Stage2 - ca. 18 SC**	**Stage3 - ca. 35 SC**
limit=5	54.67	**61.87**	56.53
limit=50	**80.30**	69.96	55.56
limit=200	**78.83**	68.33	56.22

Table : Accuracy depending on number of seed concepts (SC) and evidence limit applied.

Details about Relevance Assessment

- More detailed look at concept candidates that were rated as non-relevant
- From 100 candidates rated non-relevant to the domain of climate change, 61% were in fact at least partly relevant to the domain, but very generic or too specific.
- $\rightarrow$ only 39% not relevant at all.
- Too generic: for example: “impact”, “mitigation”, “issue”, “policy”, etc.

Results & Conclusions

- A few thousand terms are enough evidence to leverage redundancy, evidence beyond that doesn't provide much benefit (esp. if sorted by expected quality).
- 10-15 heterogeneous evidence sources sufficient to gain benefits of redundancy.
- This information is helpful to set up new systems, or when needing to scale down some existing system.
- Balancing input from different sources in general more beneficial than raw number of evidence per source.

Future Work

Future Work

- More domains
- Other systems to support generalizability.
- System optimization by per **source impact**. Currently all sources have the same impact factor set in the learning algorithm. Preliminary results show that accuracy can be raised ca. 5-7% this way.

Thank you

- gerhard.wohlgenannt@wu.ac.at, http://www.wu.ac.at/infobiz
- Questions?
- I am thankful for remarks! :-)