The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet

Christian M. Meyer

TECHNISCHE UNIVERSITÄT DARMSTADT

Elisabeth Niemann and Iryna Gurevych
International Conference on Computational Semantics
Oxford, UK, January 12–14, 2011.

UKP

Motivation

Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

Motivation

Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

✓ precise taxonomy

✗ textual information

✗ size

✗ multilingual

Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

✓	precise taxonomy	✗
✗	textual information	✓
✗	size	✓
✗	multilingual	✓

Motivation
Aligning Sense Inventories

TECHNISCHE UNIVERSITÄT DARMSTADT

Many NLP tasks rely on sense information:

Alignment on sense level

WordNet

✓	precise taxonomy	✓
✓	textual information	✓
✓	size	✓
✓	multilingual	✓

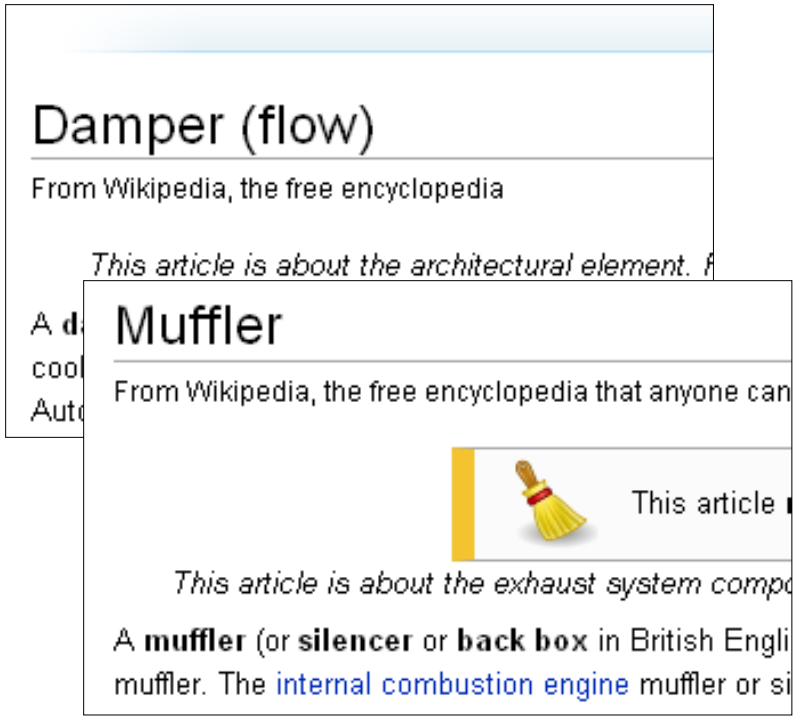
WIKIPEDIA
The Free Encyclopedia

Motivation
Alignment on Sense Level

TECHNISCHE UNIVERSITÄT DARMSTADT

Alignment on sense level

- S: (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- S: (n) **damper**, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) **damper** (a depressing restraint) *"rain put a damper on our picnic plans"*



WordNet synset

Wikipedia article

UKP

Motivation

Alignment on Sense Level

TECHNISCHE UNIVERSITÄT DARMSTADT

Alignment on sense level

- S: (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- S: (n) **damper**, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) **damper** (a depressing restraint) *"rain put a damper on our picnic plans"*

WordNet synset

Wikipedia article

Motivation

Alignment on Sense Level

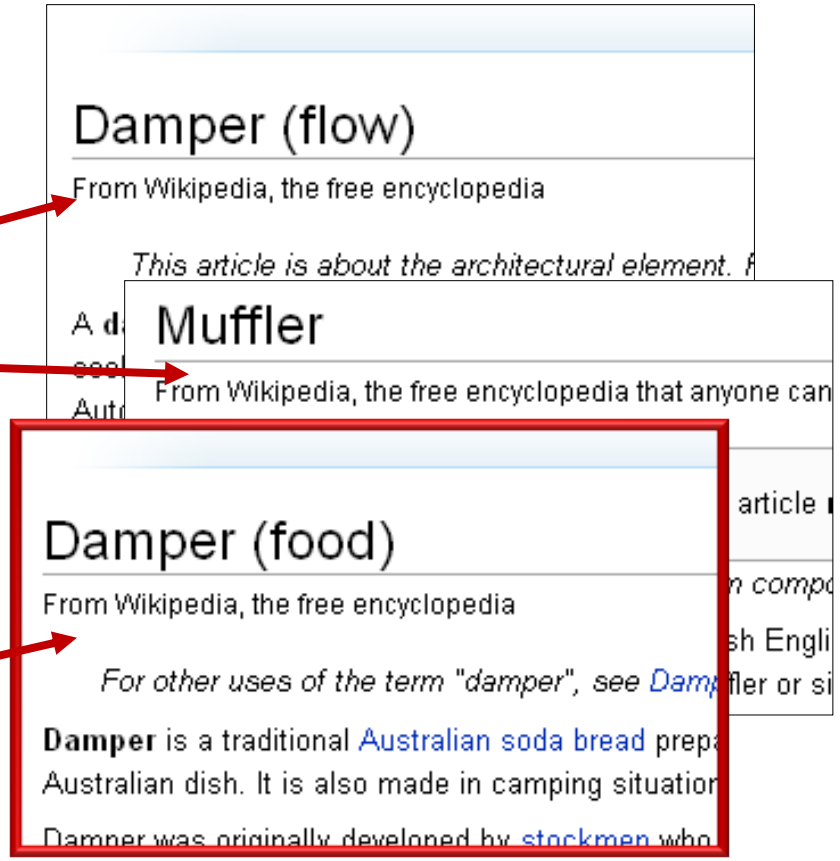
TECHNISCHE UNIVERSITÄT DARMSTADT

Alignment on sense level

- S: (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- S: (n) **damper**, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) **damper** (a depressing restraint) *"rain put a damper on our picnic plans"*

?

?



WordNet synset

UKP

Motivation

Alignment on Sense Level

TECHNISCHE UNIVERSITÄT DARMSTADT

Alignment on sense level

Two main benefits:

1. Enhanced sense representation

2. Increase of sense coverage

- S: (n) **damper** (a movable i… in a stove or chimney or fur…)
- S: (n) **damper**, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) **damper** (a depressing restraint) *"rain put a damper on our picnic plans"*

?

?

WordNet synset

UKP

Related Work

Automatic Sense Alignment of Wikipedia and WordNet

TECHNISCHE UNIVERSITÄT DARMSTADT

- Alignment of WordNet and Wikipedia's **category** system
  - (Suchanek et al., 2007); (Toral et al., 2008/2009); (Ponzetto and Navigli, 2009)
  - Category system is much smaller (0.5M vs. >3M)
  - Neglects huge amount of textual content in articles
  - Different goal: semantically enriched ontology

- Alignment of WordNet and Wikipedia **articles**
  - (Ruiz-Casado et al., 2005): Simple English Wikipedia
  - (Ponzetto and Navigli, 2010): English Wikipedia
  - Alignment based on (normalized) word overlap measure
  - Focus on 1:1 alignment

Related Work

1:1 Alignment vs. n:m Alignment

TECHNISCHE UNIVERSITÄT DARMSTADT

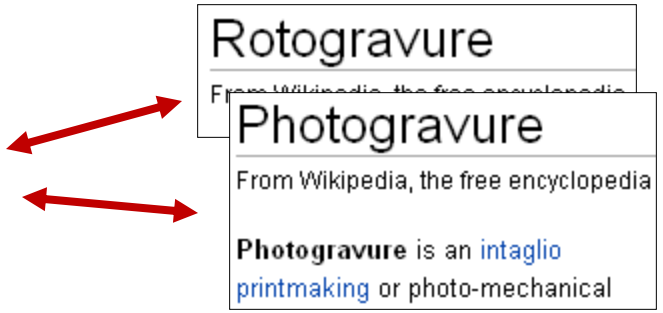
Both algorithms are modelled in a way that they always align the most likely WordNet synset for a given Wikipedia article (or vice versa):

- What if there is no Wikipedia counterpart for a given WordNet synset (or vice versa)?

S: (n) **dream** (someone or something wonderful) *"this dessert is a dream"* ⟷ **?**

- What if there is more than one Wikipedia article that can be aligned to a WordNet synset (or vice versa)?

S: (n) **photogravure**, rotogravure (using photography to produce a plate for printing)



UKP

Related Work

1:1 Alignment vs. n:m Alignment

TECHNISCHE UNIVERSITÄT DARMSTADT

Both algorithm[illegible]align the most likely WordNe[illegible]e versa):

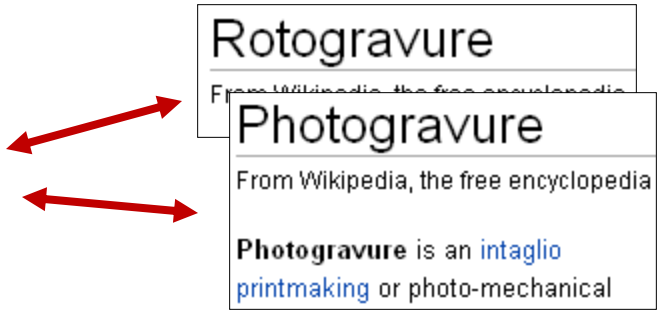
Need for n:m Alignment!

- What if there is no Wikipedia counterpart for a given WordNet synset (or vice versa)?

S: (n) **dream** (someone or something wonderful) *"this dessert is a dream"* ⟷ ?

- What if there is more than one Wikipedia article that can be aligned to a WordNet synset (or vice versa)?

S: (n) **photogravure**, rotogravure (using photography to produce a plate for printing)



UKP

Aligning Wikipedia and WordNet

Our Contributions

TECHNISCHE UNIVERSITÄT DARMSTADT

Novel Two-Step Approach for Sense Alignment

Well-Balanced Reference Dataset for Evaluation

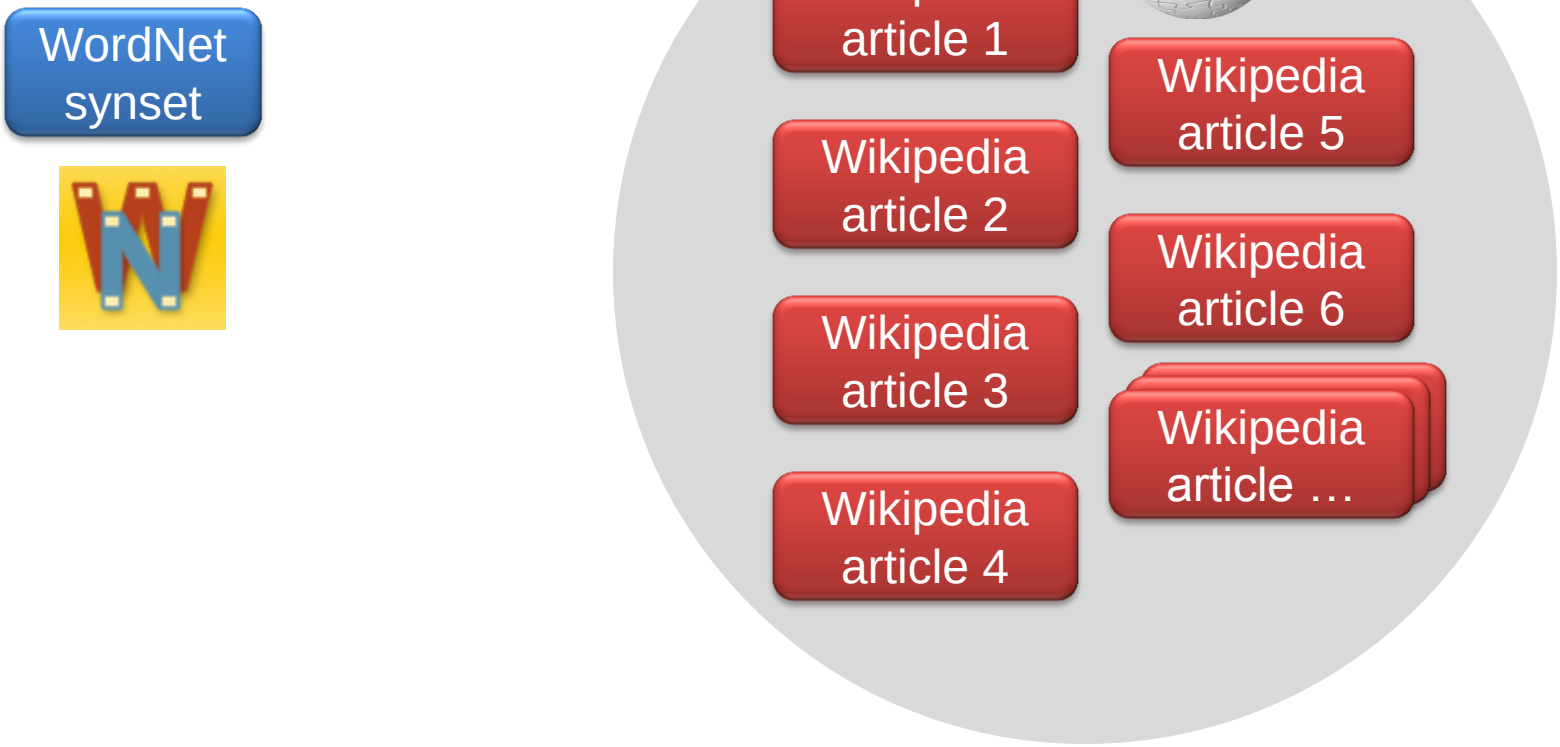
Full Alignment Publicly Available

Aligning Wikipedia and WordNet

A Two-Step Approach

TECHNISCHE UNIVERSITÄT DARMSTADT

1. *Candidate extraction*
2. *Candidate disambiguation*



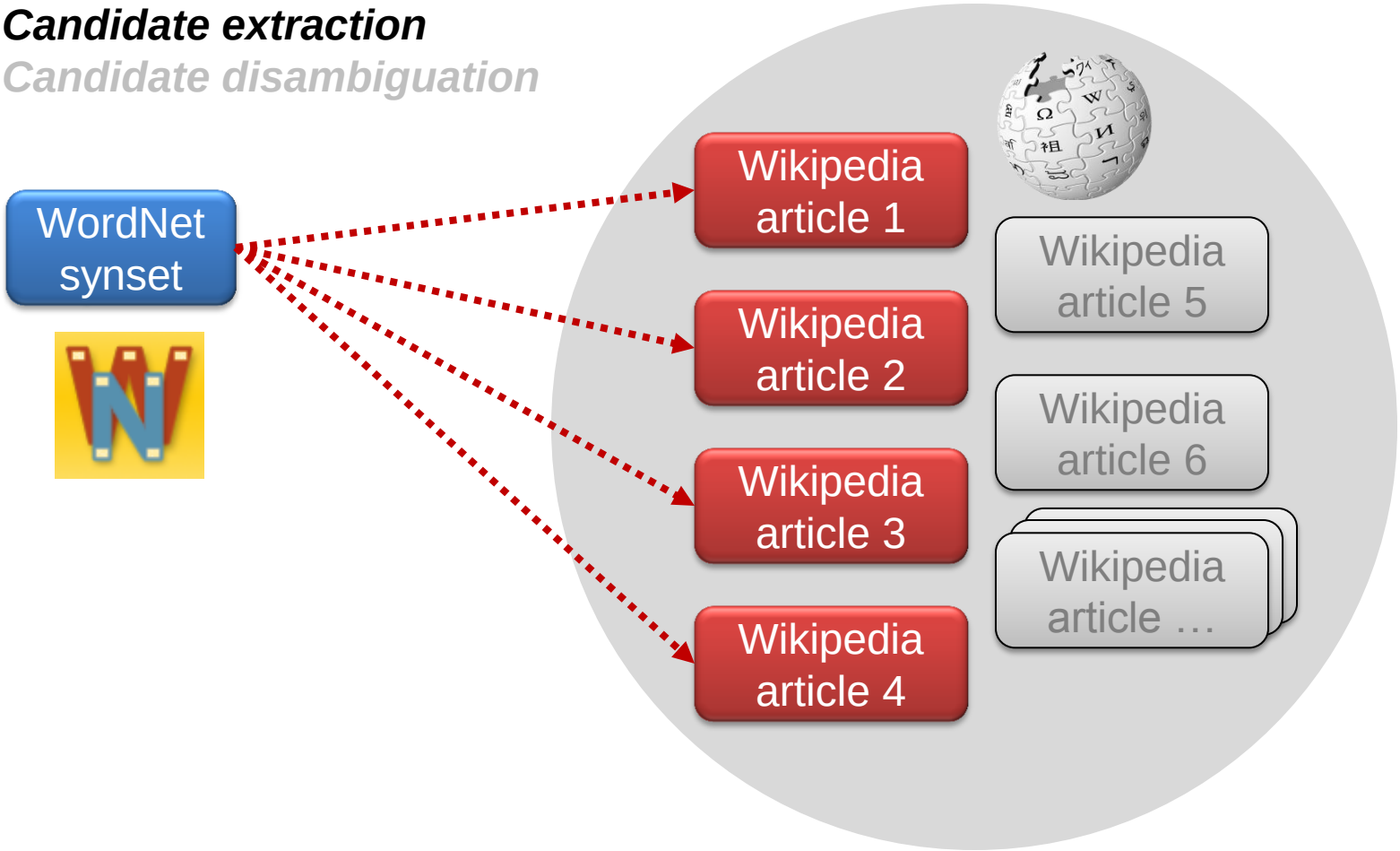
UKP

Aligning Wikipedia and WordNet

A Two-Step Approach

TECHNISCHE UNIVERSITÄT DARMSTADT

1. ***Candidate extraction***
2. ***Candidate disambiguation***



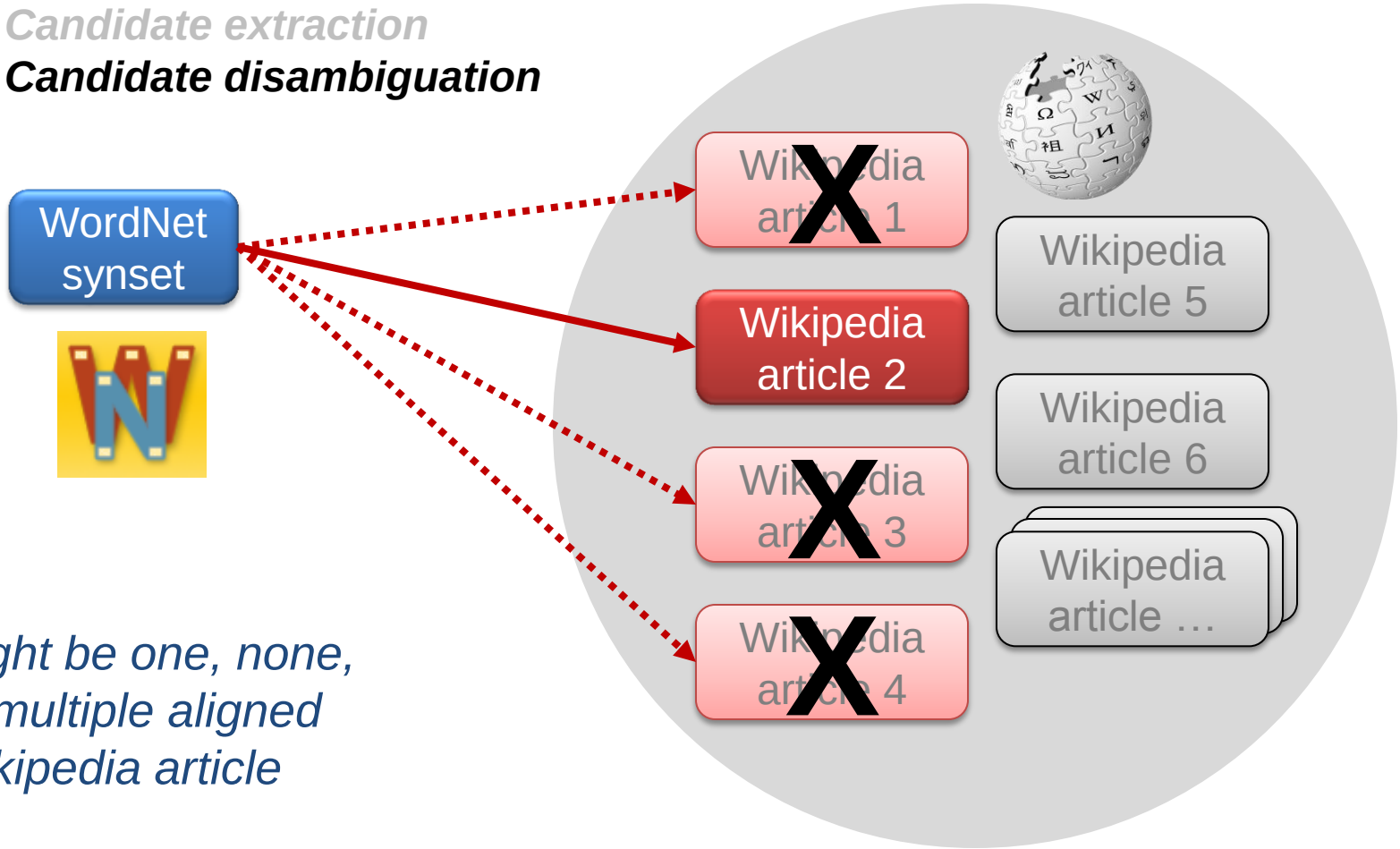
UKP

Aligning Wikipedia and WordNet

A Two-Step Approach

TECHNISCHE UNIVERSITÄT DARMSTADT

1. *Candidate extraction*
2. ***Candidate disambiguation***



Might be one, none, or multiple aligned Wikipedia article

UKP

Step 1: Candidate Extraction
Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

- For each synonymous word in the synset extract
  - Articles with the same title
  - Articles with a matching redirect
  - Articles with an inlink of the form `[[target|label]]`

- Example:
  - article *Script (typefaces)*
  - article *Script (comics)*
  - article *Penmanship* (*Handwriting* has a redirect to *Penmanship*)
  - article *Writing System* (*Arabic Alphabet* e.g. links to *Writing System*)

S: (n) **handwriting**, hand, script (something written by hand) *"she recognized his handwriting"; "his hand was illegible"*

```
The 'Arabic alphabet' is the
[[writing system|script]] used for
writing several languages of ...
```

(Wolf and Gurevych, 2010)

Step 1: Candidate Extraction
Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

- For each synonymous word in the synset extract
  - Articles with the same title
  - Articles
  - Articles

High Recall
→ High Coverage of Alignments

- Example:
  - article *Script (typefaces)*
  - article *Script (comics)*
  - article *Penmanship* (*Handwriting* has a redirect to *Penmanship*)
  - article *Writing System* (*Arabic Alphabet* e.g. links to *Writing System*)

S: (n) **handwriting**, hand, script (something written by hand) *"she recognized his handwriting"; "his hand was illegible"*

```
The 'Arabic alphabet' is the
[[writing system|script]] used for
writing several languages of ...
```

(Wolf and Gurevych, 2010)

UKP

Step 2: Candidate Disambiguation

Overview

S: (n) apple (fruit with red or yellow or green skin and sweet to tart crisp whitish flesh)

?

=

Apple

From Wikipedia, the free encyclopedia

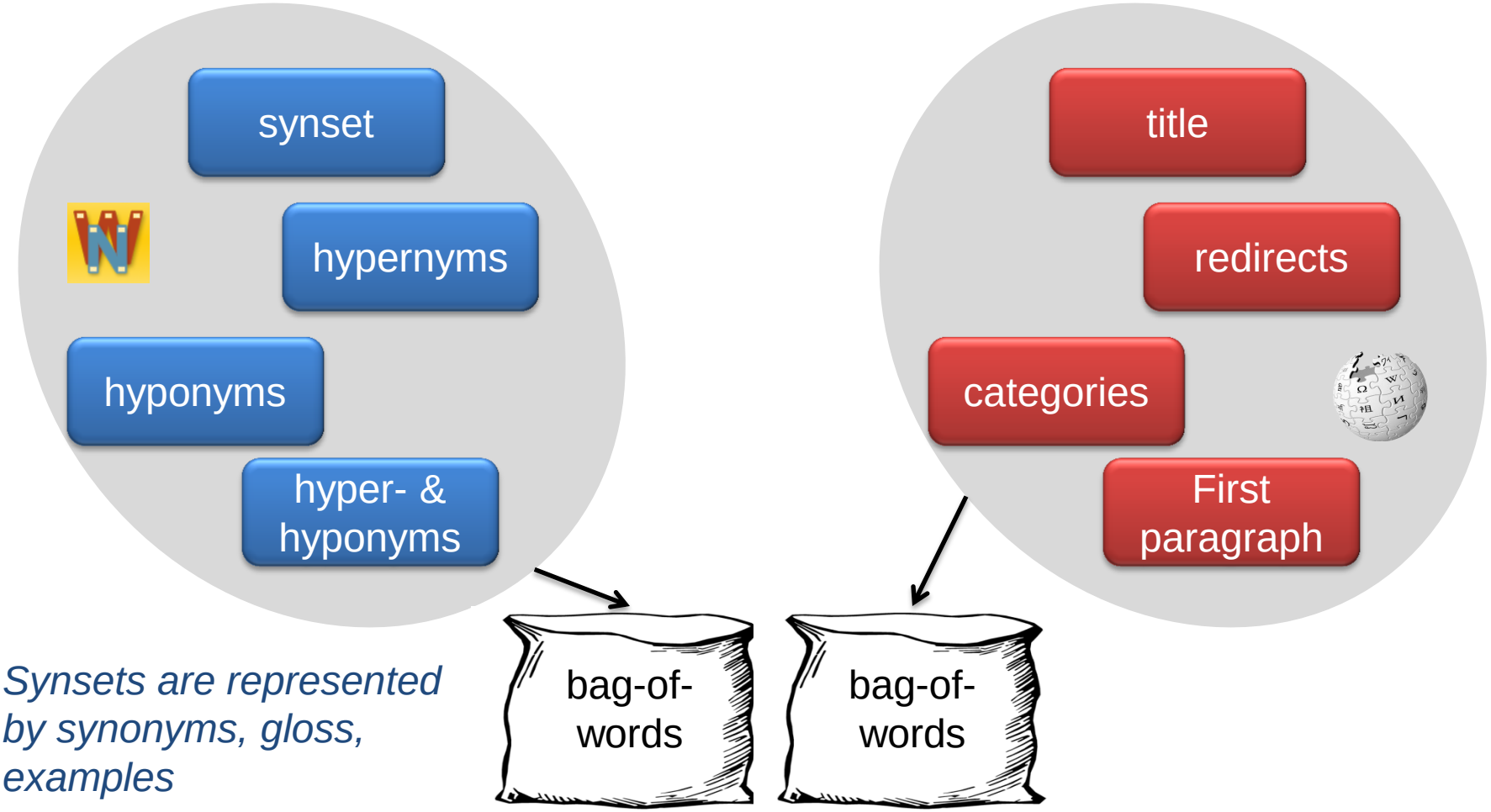
This article is about the fruit. For other uses

The **apple** is the pomaceous fruit of the apple tree, species ***Malus domestica*** in the rose family (Rosaceae), and is a perennial. It is one of the most widely cultivated tree fruits, and the most widely known of the many members of genus *Malus* that are used by humans.

a. Extract bag-of-words
b. Transform them to a vector representation
c. Calculate vector similarity scores
d. Classify each vector/sense pair as alignment or non-alignment based on a trained threshold

Step 2: Candidate Disambiguation

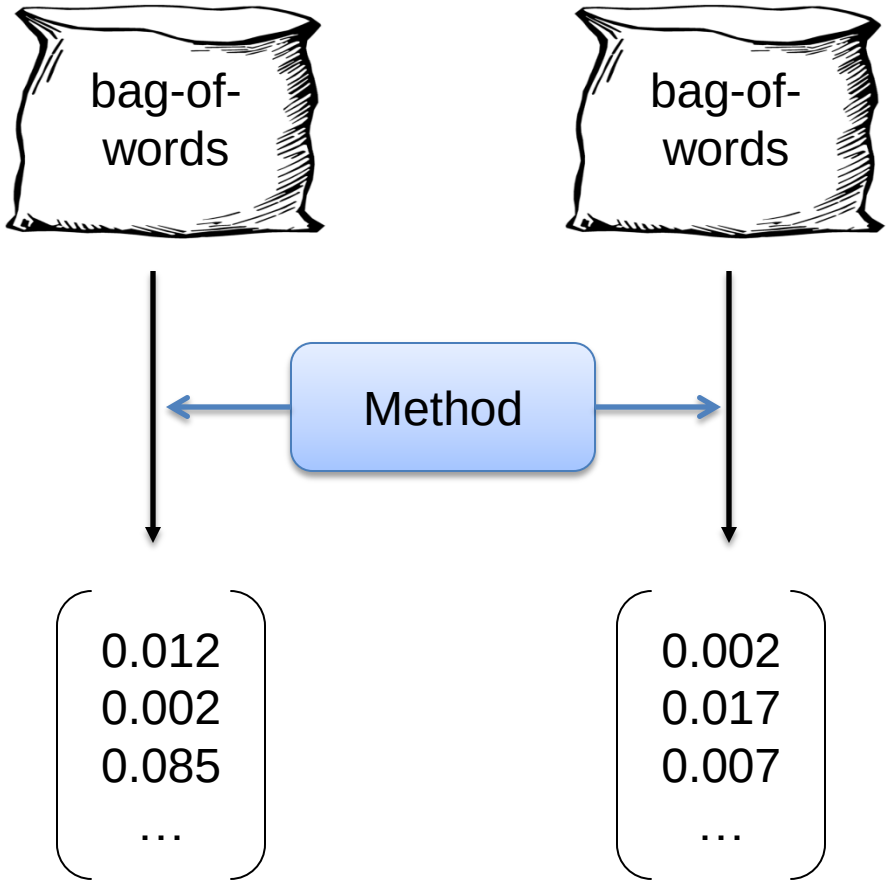
(a) Bag-of-Words



Synsets are represented by synonyms, gloss, examples

Step 2: Candidate Disambiguation

(b) Vector Representation



Step 2: Candidate Disambiguation

(c) Vector Similarity

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\text{Sim} = \cos\left(\begin{pmatrix}0.012\\0.002\\0.085\\\dots\end{pmatrix}, \begin{pmatrix}0.002\\0.017\\0.007\\\dots\end{pmatrix}\right) = 0.125$$

or

$$\text{Sim} = \text{chi}^2\left(\begin{pmatrix}0.012\\0.002\\0.085\\\dots\end{pmatrix}, \begin{pmatrix}0.002\\0.017\\0.007\\\dots\end{pmatrix}\right) = 0.117$$

Step 2: Candidate Disambiguation

(d) Alignment Classification

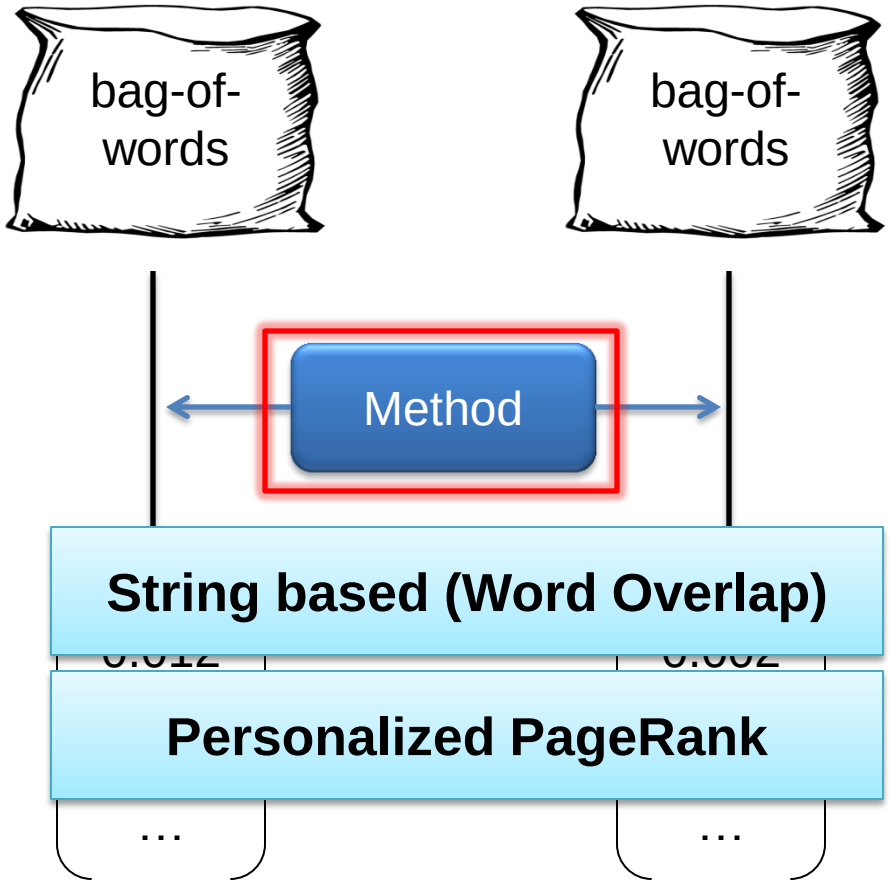
TECHNISCHE UNIVERSITÄT DARMSTADT

$$c(wn, wp) = \begin{cases} 1 & \text{if} \quad sim(wn, wp) > t \\ 0 & \text{else}\,, \end{cases}$$

- *t* is a real valued threshold
- 10-fold cross-validation to determine threshold
- use threshold that maximizes performance

Step 2: Candidate Disambiguation

(b) Vector Representation



Aligning with Personalized PageRank

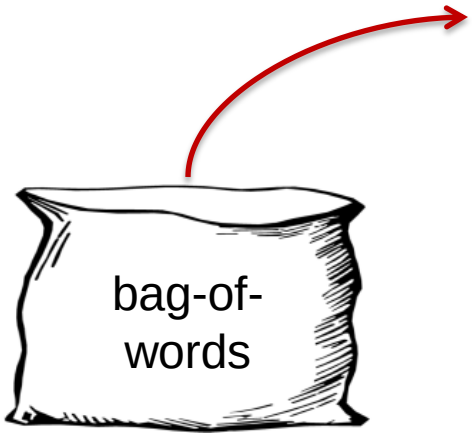
Personalized PageRank

TECHNISCHE UNIVERSITÄT DARMSTADT

- PageRank (Brin and Page, 1998) depends on transition probability *c* and random jump vector *v*
- The initial importance of a vertex can be „personalized“ using random jump vector *v* (Agirre and Soroa, 2009)
- State of the art in WSD

$$\boldsymbol{pr} = c\, M\, \boldsymbol{pr} + (1 - c)\, v$$

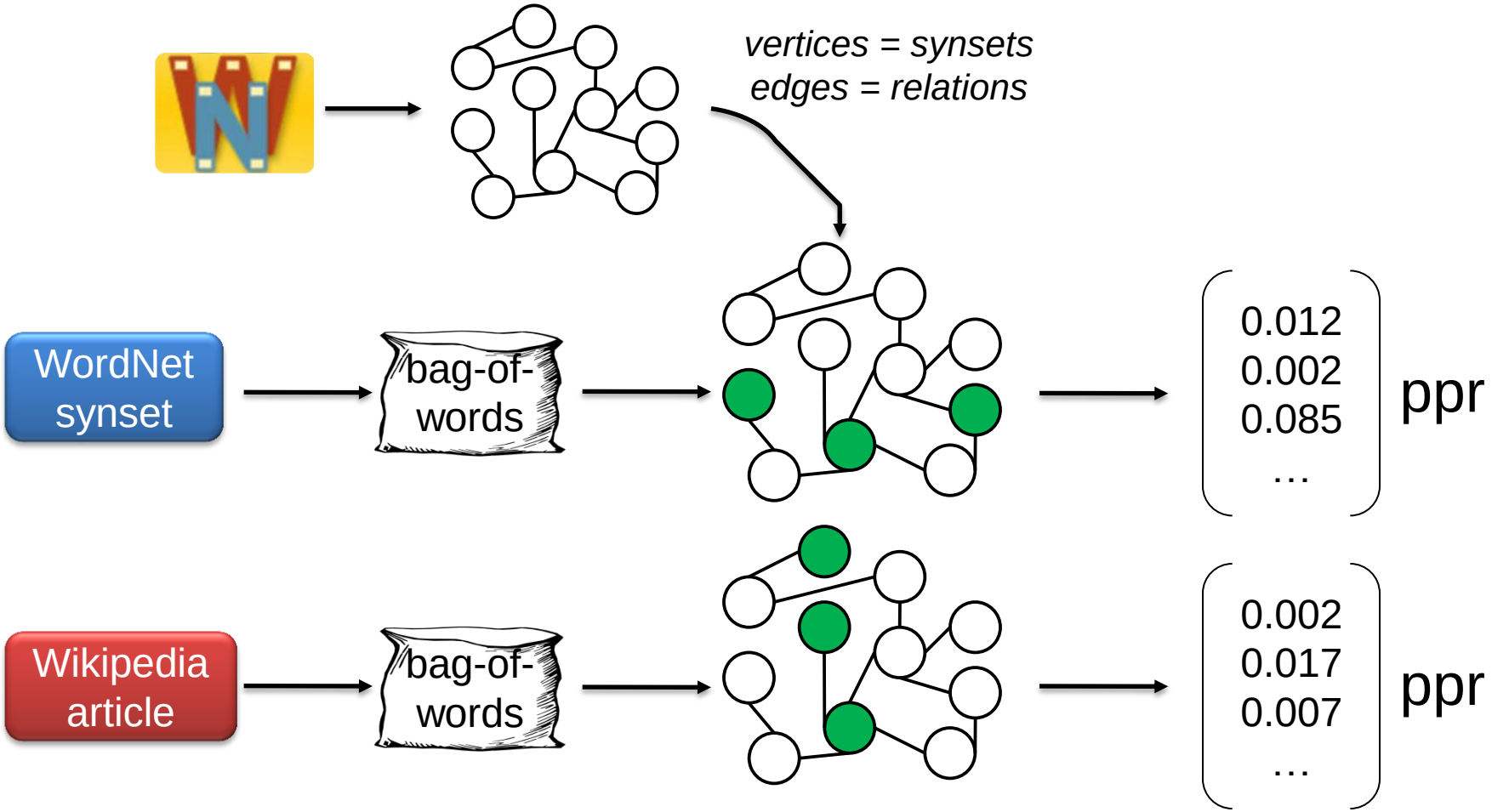
$$v_i = \begin{cases} 1 / m & \text{if } i \text{ in bag-of-words} \\ 0 & \text{otherwise} \end{cases}$$



- Personalization based on our bag-of-words
- Vertices with a word from our bag-of-words receive *1 / m* score
- *m* = number of synsets in bag-of-words

Aligning with Personalized PageRank

Our Method: ppr

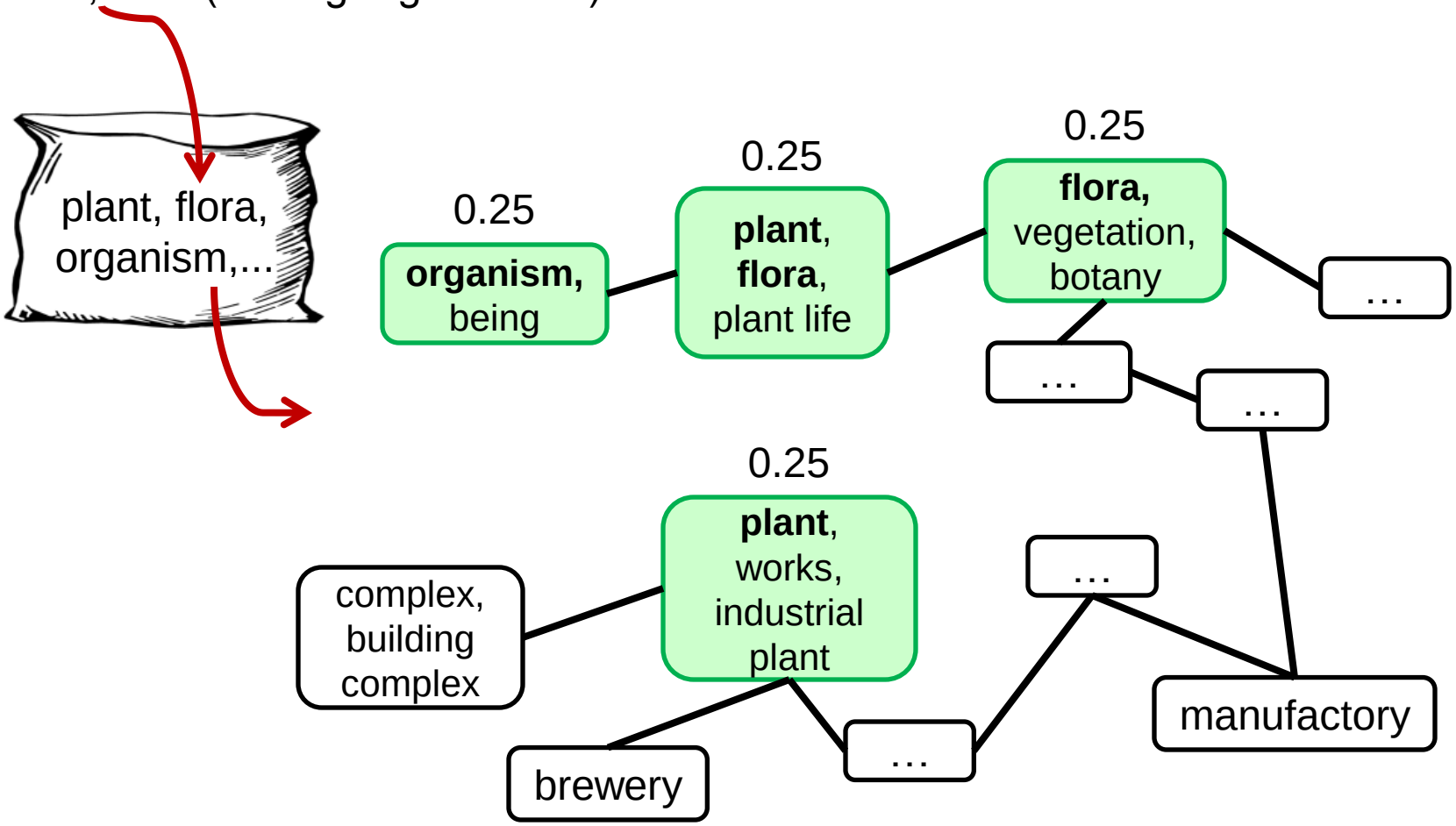


Aligning with Personalized PageRank

Our Method: ppr

TECHNISCHE UNIVERSITÄT DARMSTADT

< **plant**, flora (a living organism …) >



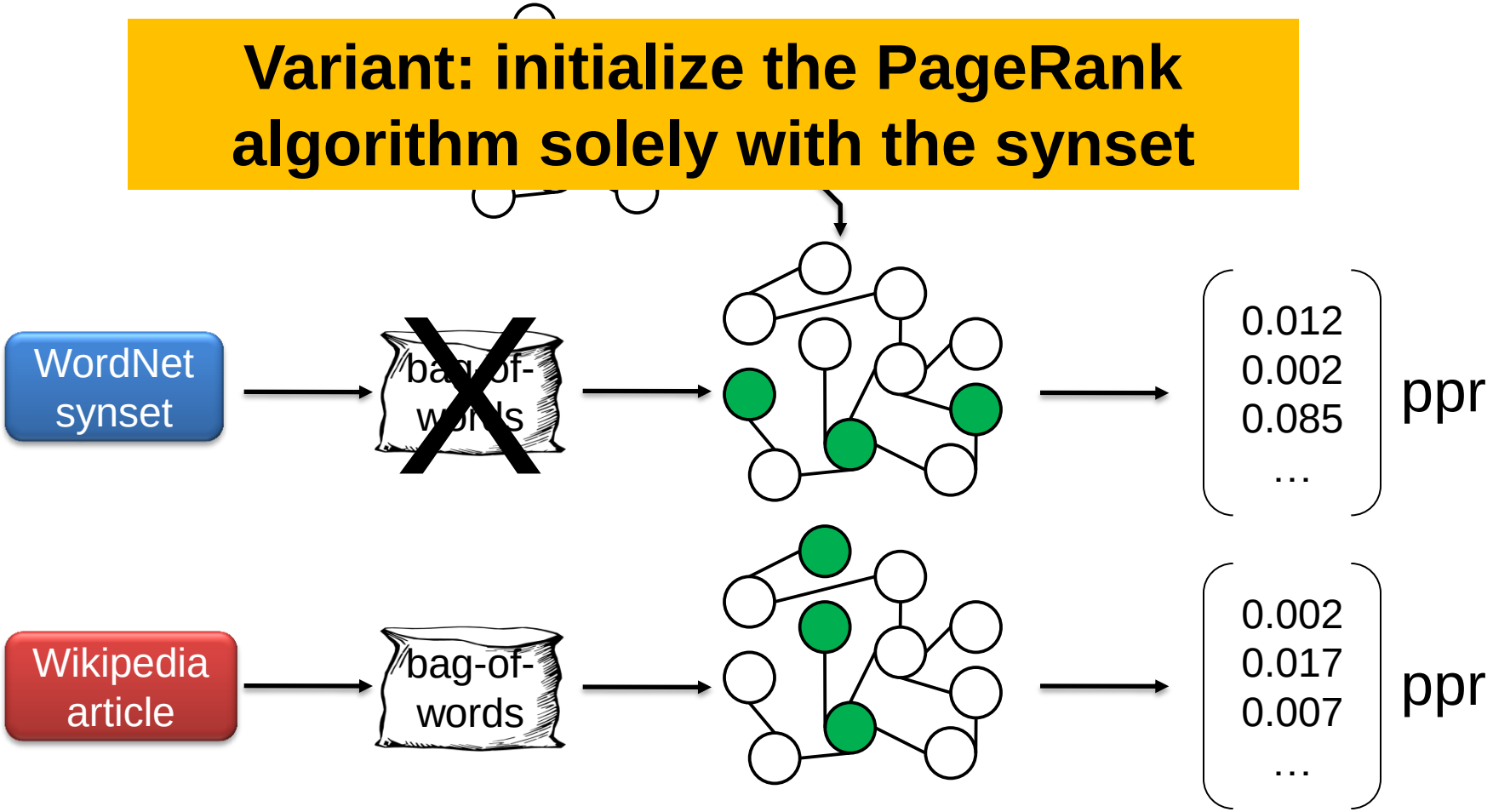
UKP

Aligning with Personalized PageRank

Our Method: ppr_d

TECHNISCHE UNIVERSITÄT DARMSTADT

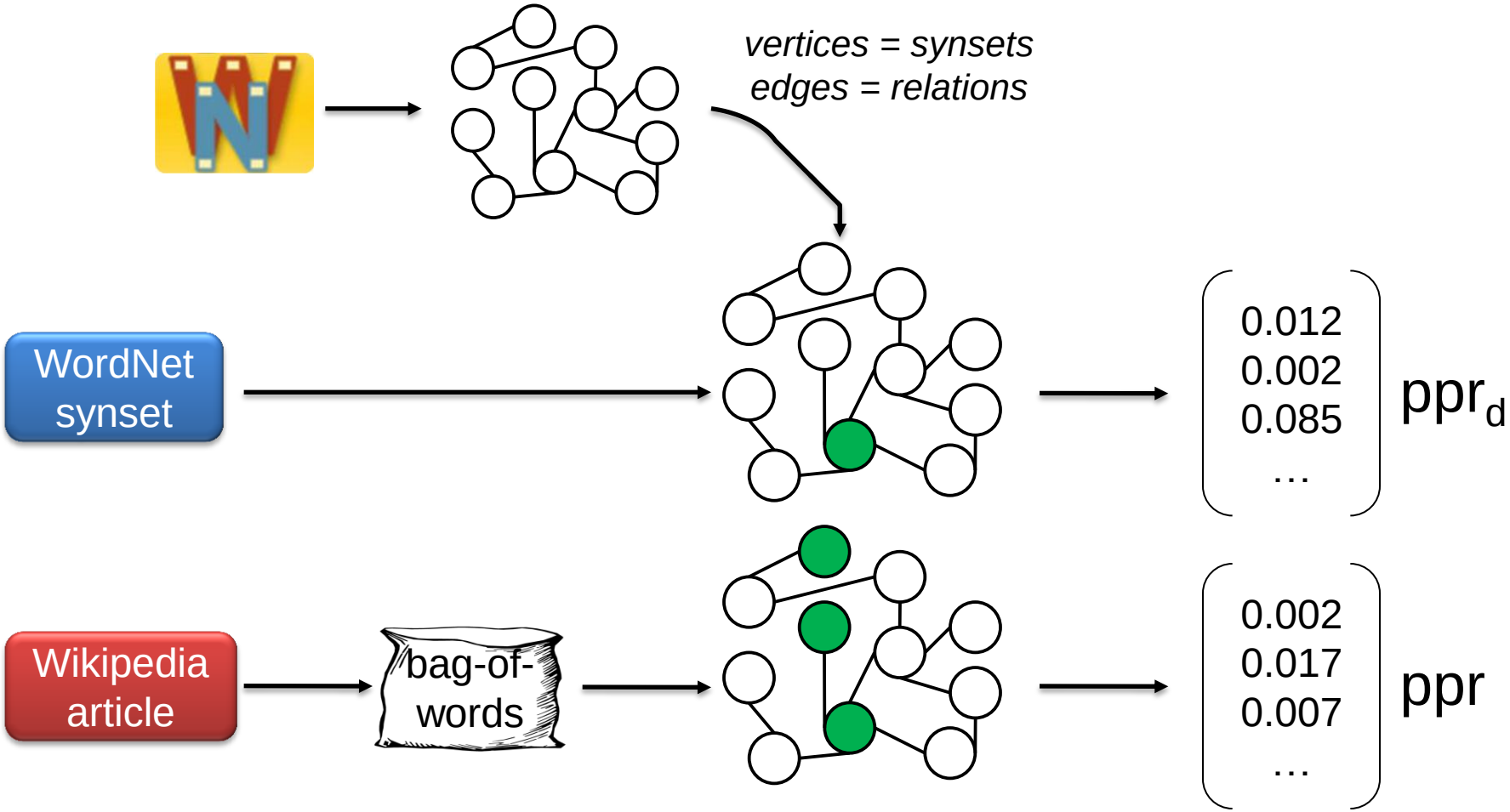
Variant: initialize the PageRank algorithm solely with the synset



UKP

Aligning with Personalized PageRank

Our Method: ppr_d



Gold Standard

Well-Balanced Reference Dataset

TECHNISCHE UNIVERSITÄT DARMSTADT

- 320 WordNet noun synsets covering:
  - Different synset sizes
  - Different shortest path lengths to root
  - Different unique beginners
  - Different number of extracted Wikipedia article candidates

- 1,815 sense alignment candidates
  - Annotated by three human annotators
  - Good pairwise annotator agreement: $\kappa = 0.866 \ldots 0.878$
  - Gold standard created using majority vote
  - 227 pairs were annotated as alignment
  - 221 synsets could be aligned to at least one Wikipedia article
  - for the remaining 99 synsets, no Wikipedia article could be aligned

Gold Standard

Well-Balanced Reference Dataset

TECHNISCHE UNIVERSITÄT DARMSTADT

- 320 WordNet noun synsets covering:
  - Different synset sizes
  - Different shortest path lengths to root
  - Different unique beginners
  - Different number of extracted Wikipedia article candidates

- 1,815 sense alignment candidates
  - Annotated by three human annotators
  - Good pairwise annotator agreement: $\kappa = 0.866 \ldots 0.878$
  - Gold standard created using majority vote
  - 227 pairs were annotated as alignment
  - 221 synsets could be aligned to at least one Wikipedia article
  - for the remaining 99 synsets, no Wikipedia article could be aligned

Evaluation
Experimental Setup

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Baselines** (1:1 alignment)

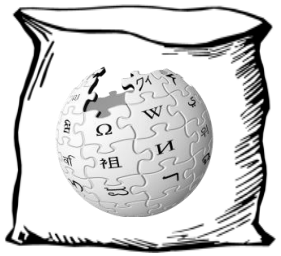
Random	*for each synset, select a random Wikipedia candidate*
MFS	*for each synset, select the most frequently linked Wikipedia article*

- **Bag of words representation – WordNet**

SYN	*synonyms, gloss & example sentence from the synset*
SYN+HYPO	*SYN plus representation of all hyponyms*
SYN+HYPER	*SYN plus representation of all hypernyms*
SYN+HYP2	*SYN plus representation of all hyponyms and hypernyms*

- **Bag of words representation – Wikipedia**

T	*Article title*
P	*First paragraph*
R	*Redirects*
C	*Categories*



Evaluation
Results (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Random baseline: 0.527
- MFS baseline: 0.534

all figures refer to F_1 measure

WordNet	Wikipedia	string	ppr_d	ppr_d + string	ppr	ppr + string

UKP

Evaluation
Results (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Random baseline: 0.527
- MFS baseline: 0.534

all figures refer to F_1 measure

WordNet	Wikipedia	string	ppr_d	ppr_d + string	ppr	ppr + string
SYN	P+T+C					
+HYPO	P+T+C					
+HYPER	P+T+C					
+HYP2	P+T+C					

Inclusion of categories (C) increases performance

Inclusion of redirects (R) decrease performance

P+T+C obtained the best results

Evaluation
Results (3)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Random baseline: 0.527
- MFS baseline: 0.534

all figures refer to F_1 measure

WordNet	Wikipedia	string	ppr_d	ppr_d + string	ppr	ppr + string
SYN	**P+T+C**	.698	**.754**		.726	
+HYPO	**P+T+C**	.702	**.739**		.722	
+HYPER	**P+T+C**	.738	.752		**.765**	
+HYP2	**P+T+C**	.732	.739		**.746**	

Personalized PageRank always outperforms string overlap approach

ppr_d outperforms ppr for SYN and +HYPO

Hypernym synsets increase performance of ppr

Evaluation
Results (4)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Random baseline: 0.527
- MFS baseline: 0.534

all figures refer to F_1 measure

WordNet	Wikipedia	string	ppr_d	ppr_d + string	ppr	ppr + string
SYN	**P+T+C**	.698	.754	**.756**	.726	**.743**
+HYPO	**P+T+C**	.702	.739	**.747**	.722	**.740**
+HYPER	**P+T+C**	.738	.752	**.765**	.765	**.781**
+HYP2	**P+T+C**	.732	.739	**.757**	.746	**.769**

Combinational approach always yields better performance (due to increasing precision)

Evaluation
Results (5)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Random baseline: 0.527
- MFS baseline: 0.534

all figures refer to F_1 measure

WordNet	Wikipedia	string	ppr_d	ppr_d + string	ppr	ppr + string
SYN	**P+T+C**	.698	.754	.756	.726	.743
+HYPO	**P+T+C**	.702	.739	.747	.722	.740
+HYPER	**P+T+C**	.738	.752	.765	.765	**.781**
+HYP2	**P+T+C**	.732	.739	.757	.746	.769

Combination of ppr and string yields best performance with

WordNet synset + hypernyms

Wikipedia article title + first paragraph + categories

Evaluation
Error Analysis

TECHNISCHE UNIVERSITÄT DARMSTADT

		automatic	
		alignment	non-alignment
manual	alignment	178	**49**
	non-alignment	**51**	1,537

- False positives due to highly related sense alignment candidates, e.g.

  (cottonseed, cottonseed oil) or *(insulin shock, insulin shock therapy)*

- False negatives due to very different sense representation, e.g.

  <payment, defrayal, defrayment: the act of paying money>
  <Payment: A payment is the transfer of wealth from one party…>

- **Future work: Include structural knowledge**

Conclusions
Lessons Learned

TECHNISCHE UNIVERSITÄT DARMSTADT

- Novel two-step approach: **Candidate Extraction** and **Disambiguation**
  - Extraction: high recall
  - Disambiguation: Combination of Personalized PageRank and Word Overlap
  - Evaluation reveals **F_1 = 0.781** on our **well-balanced reference dataset**

- With our best setting, we generated a **full alignment**
  - Not a 1:1 alignment as in previous works
  - Resources are partly complementary on sense level
  - Increased amount of knowledge for senses found in both resources

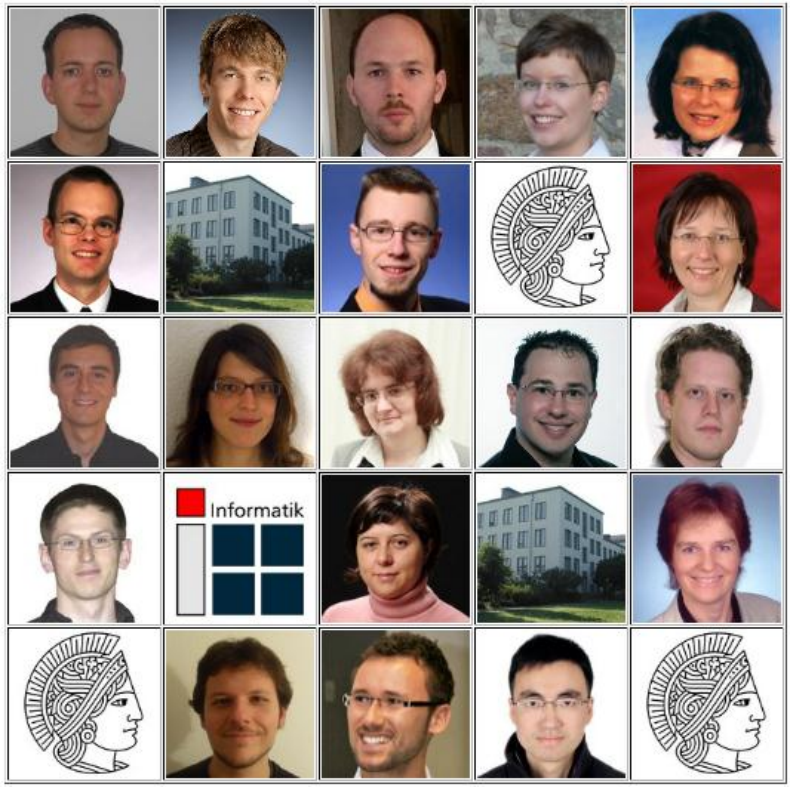
- We believe that the new resource and the enhanced knowledge therein can boost the performance of NLP tasks
  - We already started research on integrating the aligned resource in WSD tasks

Thank you for your attention!

Online Resources and Questions

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous Knowledge Processing



KTS
Klaus Tschira Stiftung gemeinnützige GmbH

e-learning center of research excellence

Bundesministerium für Wirtschaft und Technologie

VolkswagenStiftung

Deutsche Forschungsgemeinschaft DFG

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG

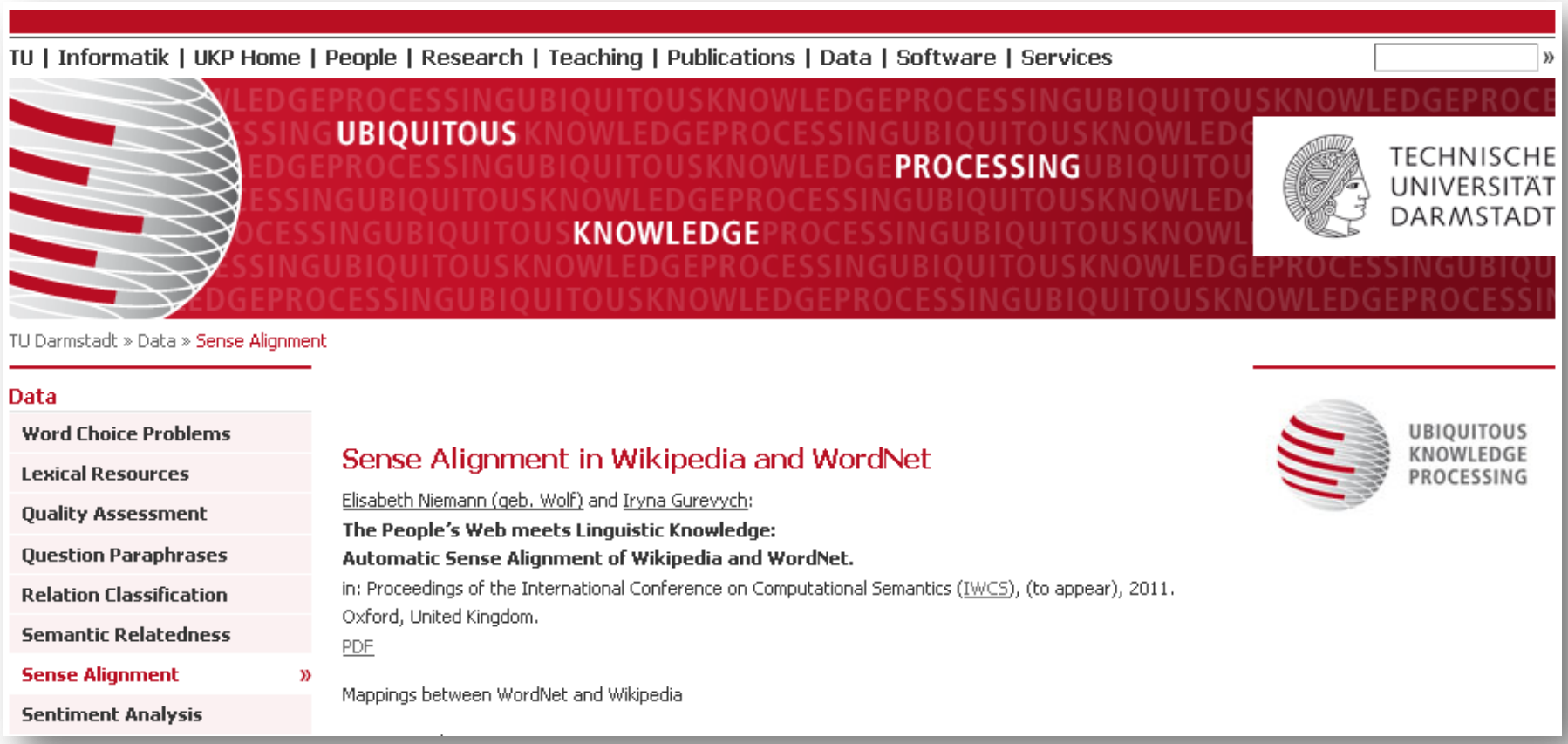
Additional Online Material:

http://www.ukp.tu-darmstadt.de/data/sense-alignment/

Thank you for your attention!

Online Resources and Questions

TECHNISCHE UNIVERSITÄT DARMSTADT



Additional Online Material:

http://www.ukp.tu-darmstadt.de/data/sense-alignment/

Kontakt / Contact

Christian M. Meyer
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

Hochschulstr. 10, 64289 Darmstadt, Germany
+49 (0)6151 16–7477
+49 (0)6151 16–5455
meyer (at) ukp.informatik.tu-darmstadt.de