

What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage



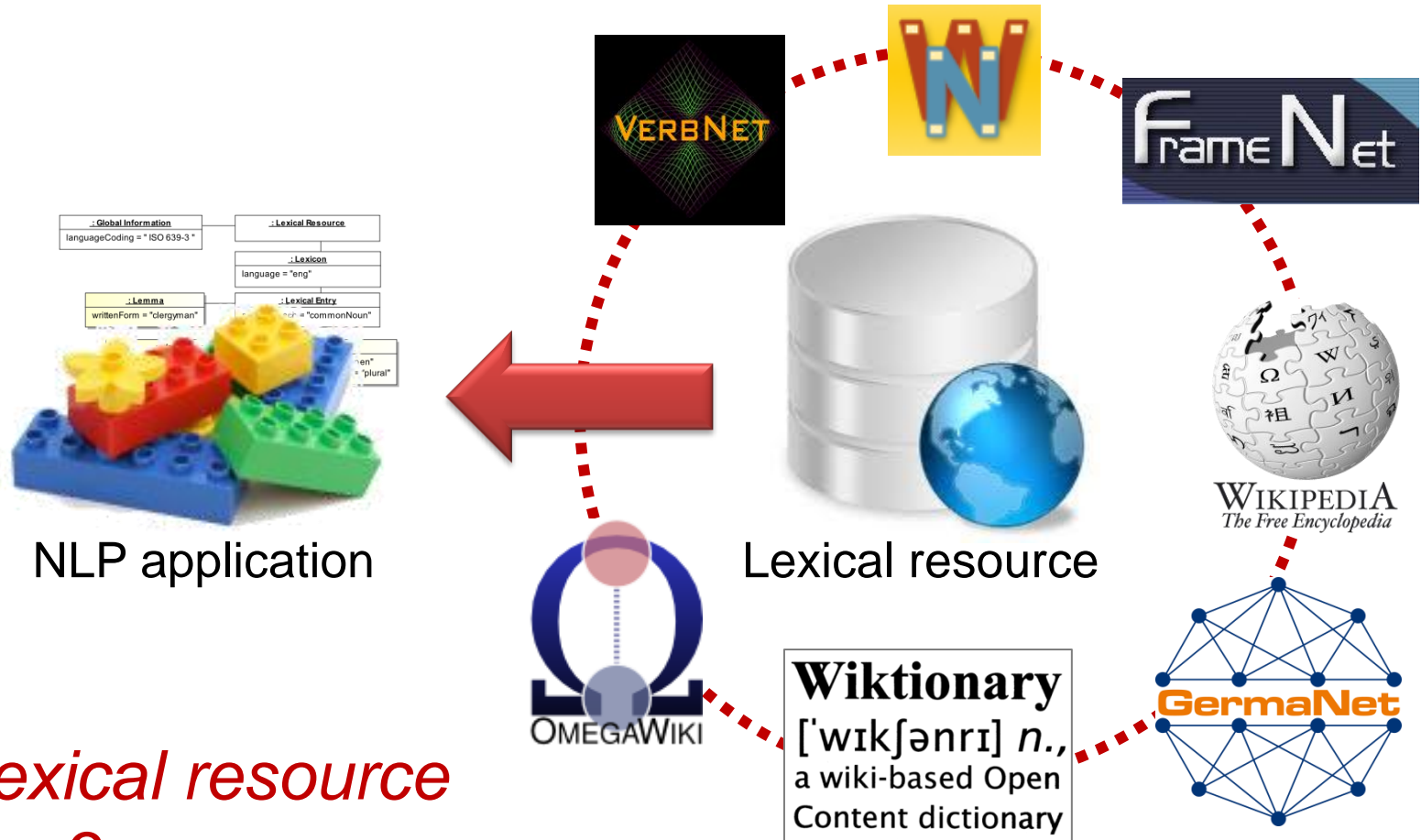
TECHNISCHE
UNIVERSITÄT
DARMSTADT

Christian M. Meyer

Christian M. Meyer and Iryna Gurevych

5th International Joint Conference on Natural Language Processing
Chiang Mai, Thailand, November 8–13, 2011.

Motivation

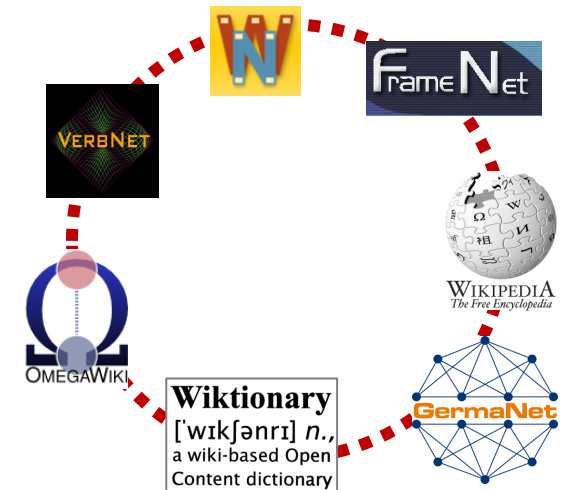


*Which lexical resource
to choose?*

Motivation

Resources are largely different

- Different coverage of words/word senses
- Different types of information
 - Encyclopedic vs. linguistic knowledge
 - Syntactic vs. semantic knowledge
 - ...



This can significantly influence the performance of your system! – Instead of choosing only one (best performing):

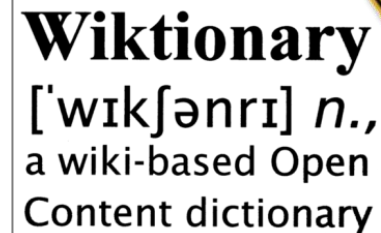
Why not combine multiple resources and benefit from all their knowledge?

Motivation

Aims:

We present a word sense alignment between Wiktionary and WordNet that comes with

- (1) Increased coverage
- (2) Enriched sense representations



Wiktionary
[ˈwɪkʃənɹɪ] *n.*,
a wiki-based Open
Content dictionary



Scope:

This work is part of the larger initiative Ubiqtionary (Uby) that aims at:

- (1) Providing a standardized model for lexical resources (using LMF)
- (2) Making available a large number of lexical resource in this format
- (3) Densely interlinking them by means of resource alignments

Lexical Resources

Wiktionary: Online lexicon that is collaboratively constructed by a community of Web users

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

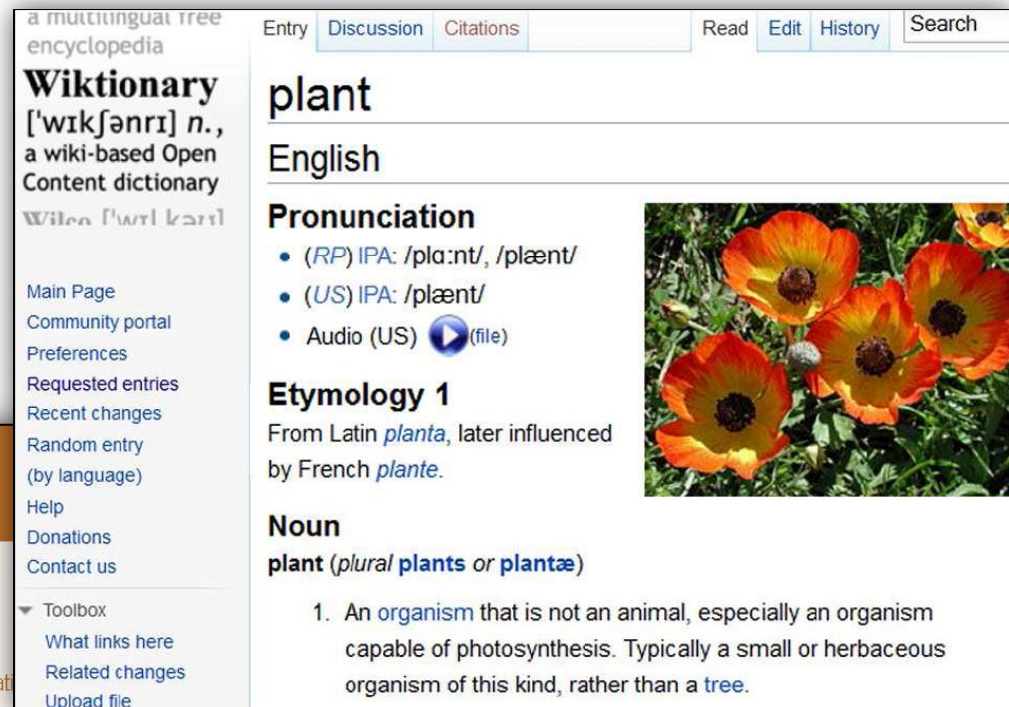
Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- **S: (n) plant**, [works](#), [industrial plant](#) (buildings for carrying on industrial labor) "*they built a large plant to manufacture automobiles*"
- **S: (n) plant**, [flora](#), [plant life](#) ((botany) a living organism lacking the power of locomotion)
- **S: (n) plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- **S: (n) plant** (something planted secretly for discovery by another) "*the police used a plant to trick the thieves*"; "*he claimed that the evidence against him was a plant*"



The screenshot shows the Wiktionary entry for 'plant'. It includes a sidebar with navigation links like 'Main Page', 'Community portal', and 'Preferences'. The main content area has tabs for 'Entry', 'Discussion', and 'Citations'. The word 'plant' is displayed in large font, followed by its English classification. The 'Pronunciation' section lists IPA for RP and US, and an audio file. The 'Etymology 1' section explains the word's origin from Latin 'planta' and French 'plante'. The 'Noun' section defines 'plant' as an organism capable of photosynthesis, with a list of numbered definitions. An image of orange and yellow flowers is shown on the right side of the page.

WordNet: Semantic network created by psycholinguists at Princeton University (Fellbaum, 1998)

Contributions

Contribution 1:

Publicly available!

**Sense Alignment between the
entire English Wiktionary and WordNet**

Contribution 2:

Publicly available!

Evaluation dataset annotated by 10 human raters

Contribution 3:

**Analysis of our aligned resource:
How it can benefit NLP tasks**

Aligning Wiktionary and WordNet

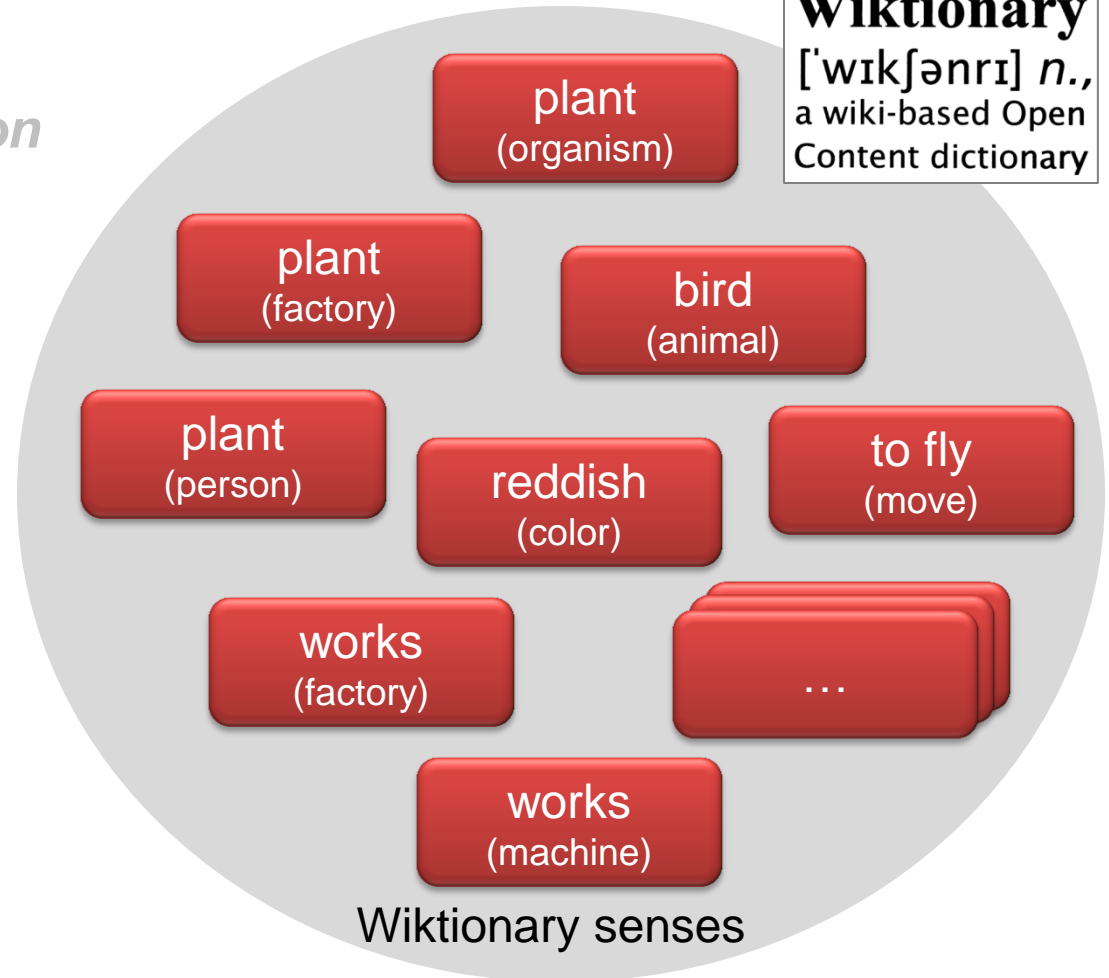
A two-step approach:

1. Candidate extraction
2. Candidate disambiguation



WordNet synsets

Wiktionary
[ˈwɪkʃənri] *n.*,
a wiki-based Open
Content dictionary



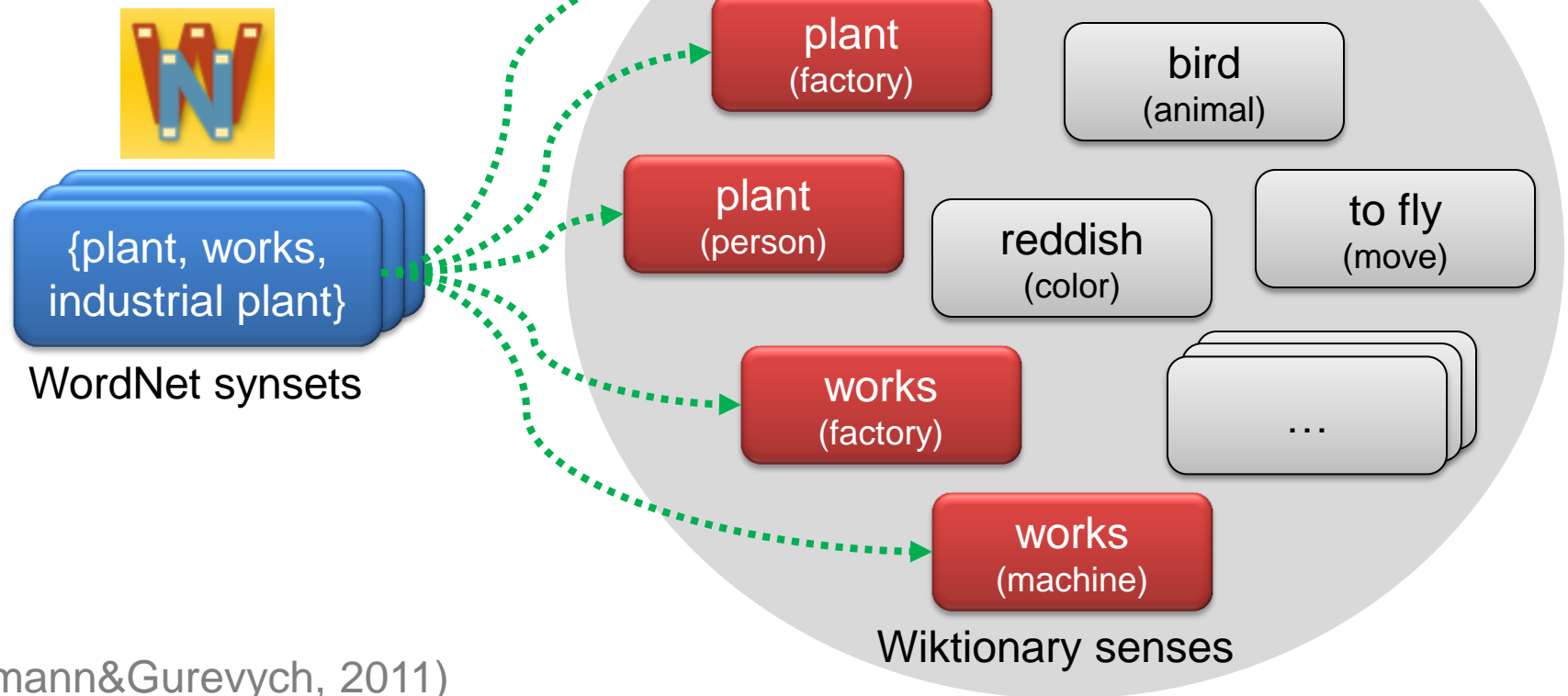
(Niemann&Gurevych, 2011)

Aligning Wiktionary and WordNet

A two-step approach:

1. Candidate extraction
2. *Candidate disambiguation*

Wiktionary
[ˈwɪkʃənri] *n.*,
a wiki-based Open
Content dictionary



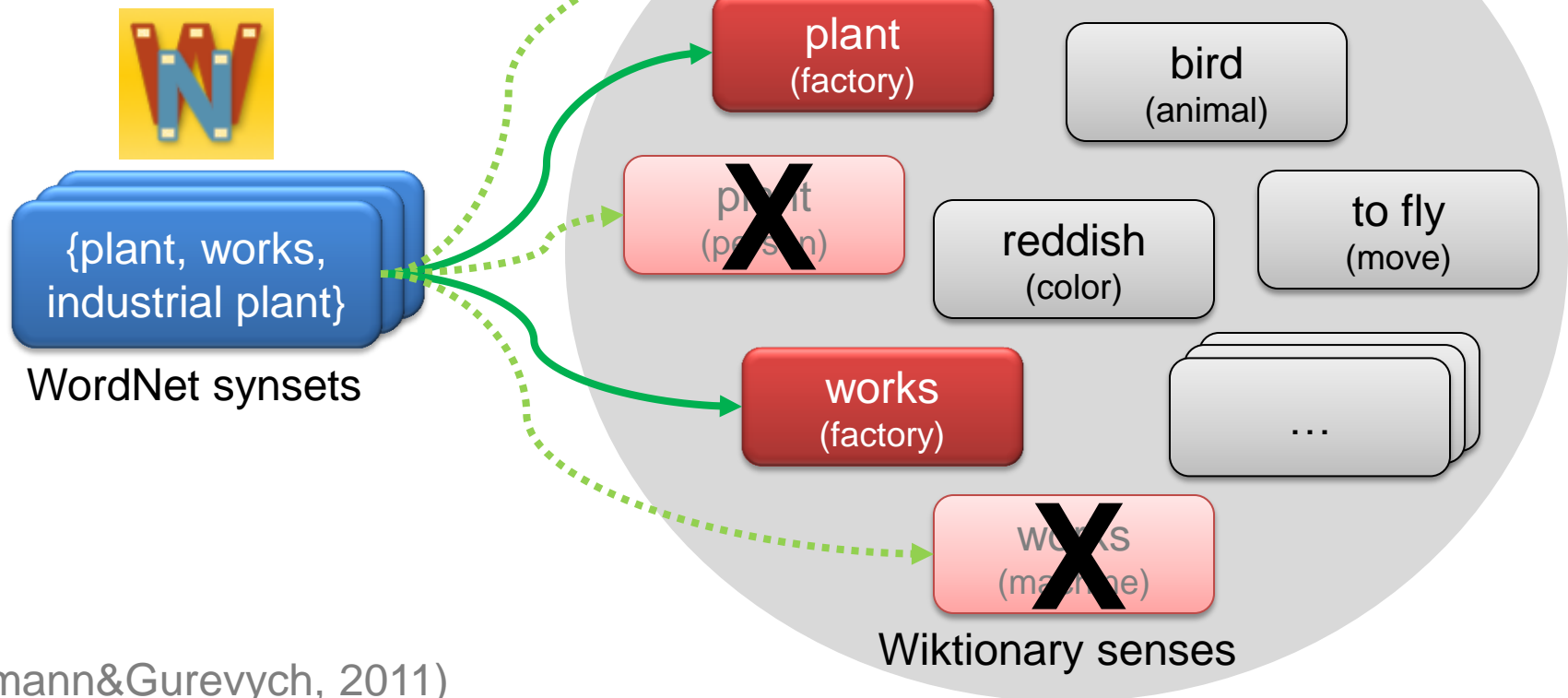
(Niemann&Gurevych, 2011)

Aligning Wiktionary and WordNet

A two-step approach:

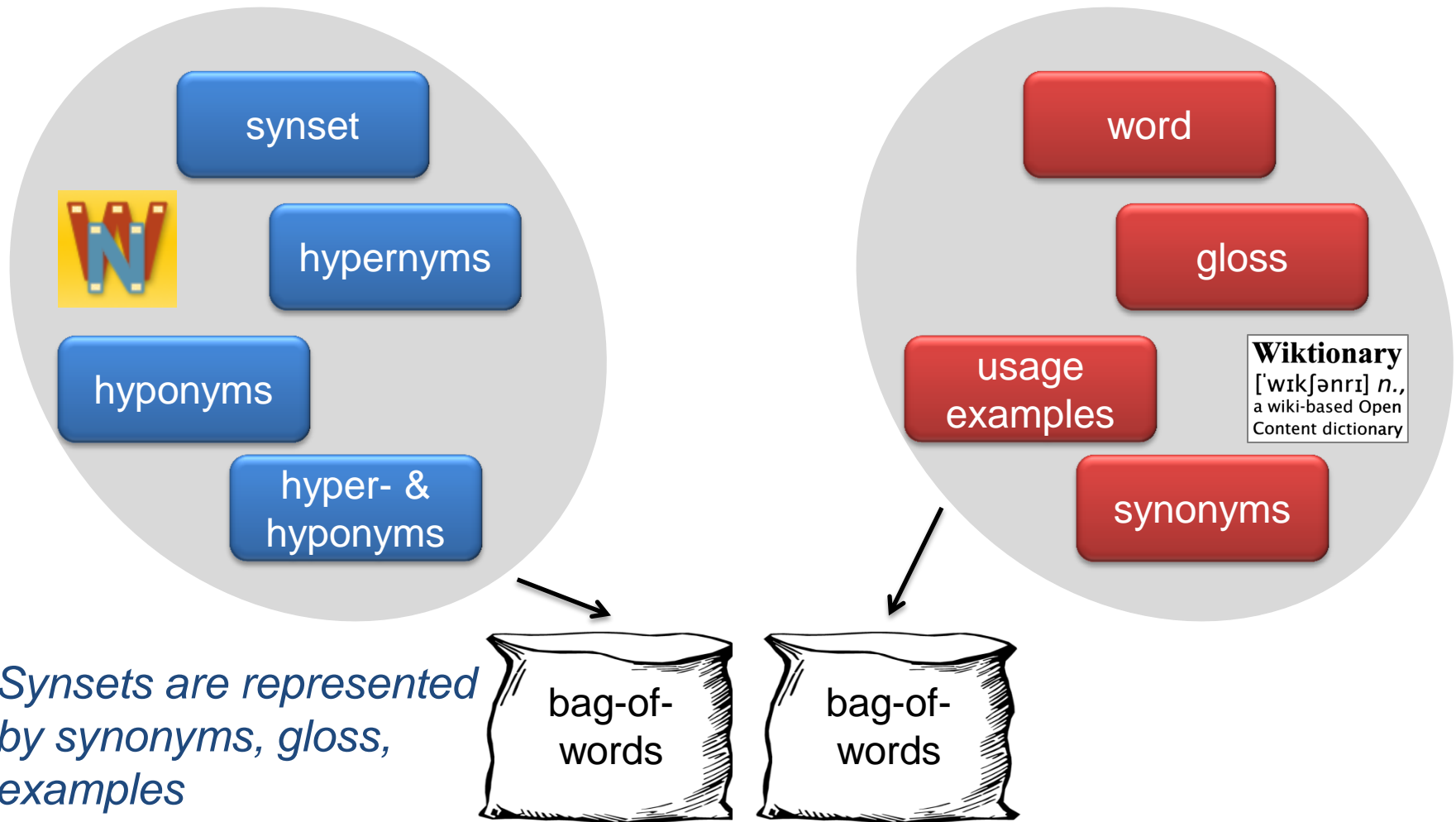
1. *Candidate extraction*
2. **Candidate disambiguation**

Wiktionary
[ˈwɪkʃənri] *n.*,
a wiki-based Open
Content dictionary



(Niemann&Gurevych, 2011)

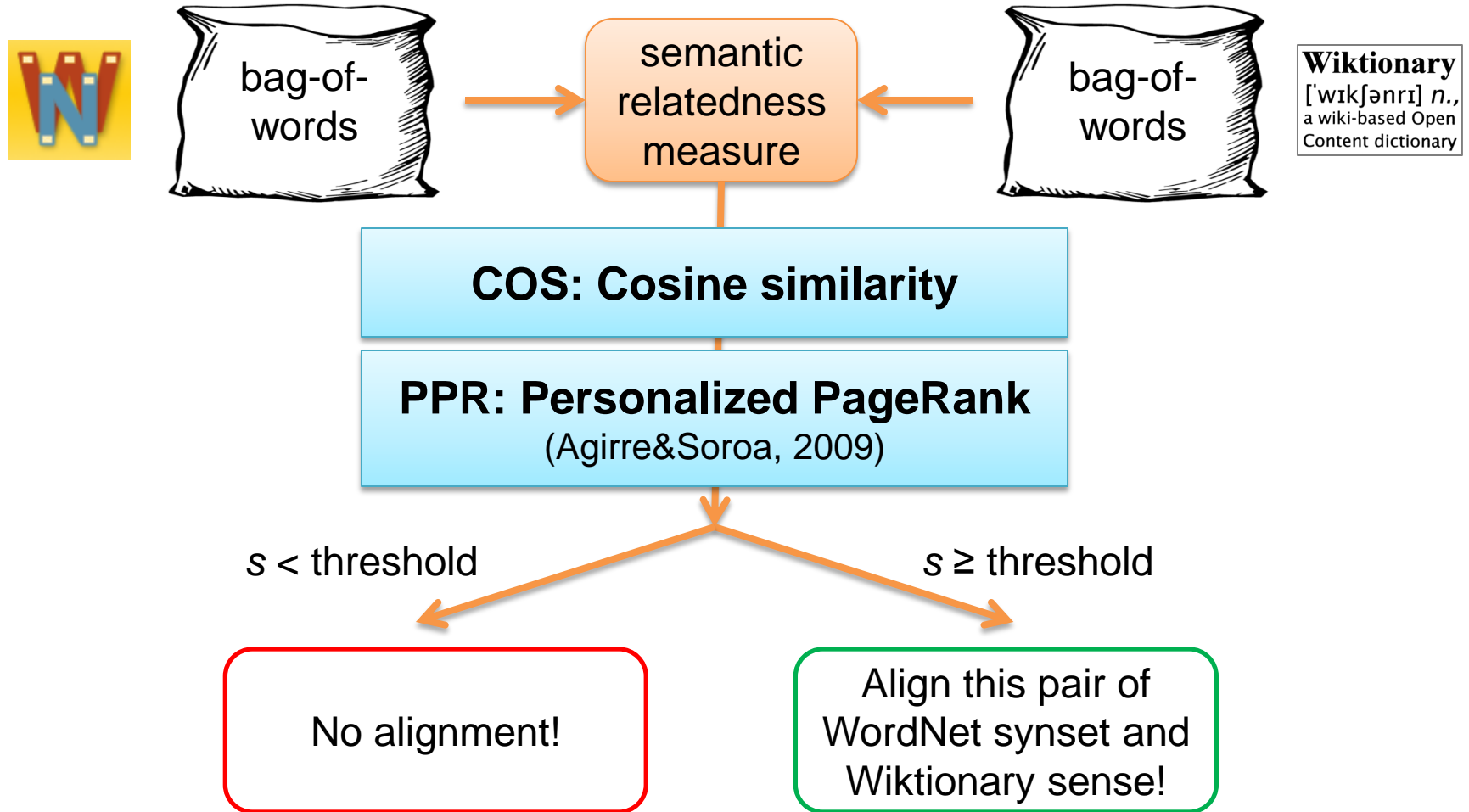
Disambiguation: BoW Representation



Disambiguation: Alignment Classification



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Evaluation Dataset



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Publicly available!

Dataset creation:

- No previous alignments = no other evaluation datasets
- We created a new dataset with 2,423 sense pairs
- 10 human raters (students/researchers from CS, math, linguistics)
- Annotate each pair as “same meaning” or “different meaning”

Dataset reliability:

- Inter-rater agreement: $A_o = .93$, $\kappa = .70$
- Removing two biased raters: $A_o = .94$, $\kappa = .74$

Gold standard:

- Majority vote of the 8 raters, additional tie breaker

Evaluation Results

- RAND: Random baseline
- MFS: Baseline aligning always the first sense (\approx most frequent sense)

Method	A	P	R	F1
RAND	.662	.212	.594	.313
MFS	.802	.329	.508	.399
COS only	.901	.598	.703	.646
PPR only	.915	.684	.636	.659
COS&PPR	.914	.674	.649	.661

- Our approach significantly outperforms the baseline (at 1% level)
- COS highest recall; PPR highest precision; COS&PPR highest F1
- Significant difference of PPR, COS&PPR over COS (at 1% level)
- No significant difference between PPR and COS&PPR

110 false negatives:

“same meaning, but was not aligned”

- Very different wording
 - *“good discernment” vs. “ability to notice what others might miss”*
- Similar senses but slightly below threshold
 - *“plants of the genus Centaurea” vs. “common weeds of the genus Centaurea”*
- References to other senses
 - *pacification: “the process of pacifying”*

98 false positives:

“different meaning, but have been aligned”

- Highly related, but slightly different specifications
 - *“a computer that provides client stations with access to files and printers as shared resources to a computer network” vs. “any computer attached to a network”*
- Erroneous interpretation or domain-specific vocabulary
 - *“any organization that provides resources and facilities for a function or event”*

Increased Coverage: Parts of Speech

- Our alignment: 56,970 sense pairs
- Final resource contains 488,988 word senses
- Substantial increase in the coverage of senses
- Wiktionary is not restricted to nouns/verbs/adjectives: proverbs, idioms, collocations, particles, determiners, inflected forms, etc.

	Wiktionary AND WordNet	Only Wiktionary	Only WordNet
Nouns	34,464	158,085	47,651
Verbs	8,252	29,119	5,515
Adj./Adv.	14,236	60,977	7,541
Other POS	0	16,778	0
Inflected Forms	0	106,328	0

Increased Coverage: Domains

	Wiktionary AND WordNet	Only Wiktionary	Only WordNet
Biology	4,465	4,067	12,869
Chemistry	2,561	8,260	2,268
Engineering	1,108	940	1,080
Geology	2,287	2,898	2,479
Humanities	4,949	2,700	5,060
IT	439	3,032	557
Linguistics	1,249	1,011	1,576
Math	615	2,747	483
Medicine	3,613	3,728	3,058
Military	574	426	585
Physics	1,246	2,835	1,252
Religion	733	1,154	781
Social Sciences	3,745	2,907	4,458
Sport	905	2,821	807

Enriched Sense Representation

Our resource can use „the best of both worlds“:



Synonyms

Gloss

Example sentence

Subsumption hierarchy

Synset organization

...

Wiktionary
[ˈwɪkʃənri] *n.*,
a wiki-based Open
Content dictionary

Pronunciation

Etymology

Syntactic knowledge

Quotations

Related terms

Translations

...

The increased coverage and the enriched sense representation yields synergies

Previously shown:

- Aligning FrameNet, VerbNet, and WordNet for semantic parsing (Shi&Mihalcea, 2005)
- Aligning VerbNet and PropBank for semantic role labeling (Loper et al., 2007)
- Aligning WordNet and Wikipedia for word sense disambiguation (Ponzetto&Navigli, 2010)

Future work:

- Semantic relatedness, information retrieval, information extraction,...
- Your application?

Conclusions

- **Aligned Wiktionary and WordNet** for the first time
- Our aligned resource is characterized by:
 - (1) **Increased coverage**
 - Different parts of speech, not only nouns
 - Humanities and social sciences from WordNet
 - Technical domains and leisure from Wiktionary
 - (2) **Enriched sense representation**
 - Pronunciation, etymology, related terms, translations, etc.
- Novel **evaluation dataset** annotated by 10 human raters
- All resources are **publicly available**

Wiktionary
[ˈwɪkʃənri] *n.*,
a wiki-based Open
Content dictionary

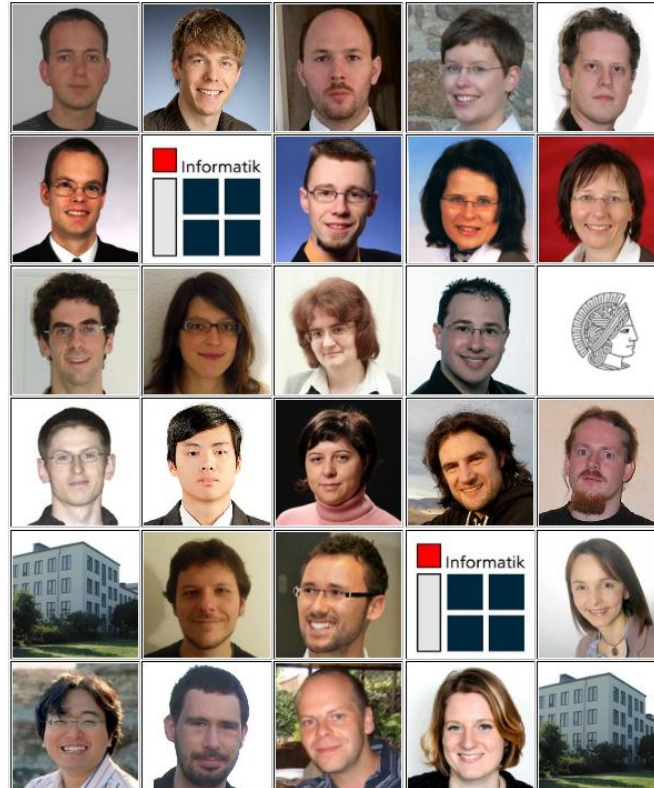


Thank you for your attention!

Ubiquitous Knowledge Processing



KLAUS TSCHIRA STIFTUNG
GEMEINNÜTZIGE GMBH



Additional Online Material:

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

Thank you for your attention!




TECHNISCHE
UNIVERSITÄT
DARMSTADT

TU | Informatik | UKP Home | People | Research | Teaching | Publications | Data | Software | Services | Contact »

UBIQUITOUS KNOWLEDGE PROCESSING

TU Darmstadt » Informatik » Data » [Sense Alignment](#)

Data	Sense Alignment of Wiktionary and WordNet
Uby	Christian M. Meyer and Iryna Gurevych :
Word Choice Problems	What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage , in: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), (to appear), November 2011. Chiang Mai, Thailand.
Lexical Resources	PDF BibTeX Proceedings
Quality Assessment	Resource Download
Question Paraphrases	<ul style="list-style-type: none">▪ Full alignment of Wiktionary and WordNet (2011-08-31, .zip, 1.5MB)▪ Classification results (2011-08-31, .zip, 4.4MB)▪ Corresponding Wiktionary data (2011-08-31, .zip, 9.1MB)▪ Evaluation dataset (2011-08-31, .txt, 97KB)
Relation Classification	
Semantic Relatedness	
Sense Alignment »	

 **UBIQUITOUS KNOWLEDGE PROCESSING**

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

Rechtliche Hinweise

Die Folien sind für den persönlichen Gebrauch der Vortragsteilnehmer gedacht. Im Vortrag verwendete Photographien, Illustrationen, Wort- und Bildmarken sind Eigentum der jeweiligen Rechteinhaber oder Lizenzgeber. Um Missverständnisse zu vermeiden, wäre eine kurze Kontaktaufnahme vor Weitergabe oder -nutzung der Vortragsmaterialien empfehlenswert. Sofern Sie Ihre Rechte verletzt sehen, bitte ich ebenfalls um Kontaktaufnahme zur Klärung der Sachlage.

Legal Issues

The slides are intended for personal use by the audience of the talk. Photographies, illustrations, trademarks, or logos are property of the holder of rights. To avoid any misconceptions, I would strongly recommend to get in touch before reusing or redistributing the slides or any additional material of the talk. The same applies if you consider your rights infringed – please let me know to initiate further clarification.