Eberhard Karls Universität Tübingen

Technische Universität Darmstadt

Semantic Information Retrieval on the Basis of an Extended GermaNet

Iryna Gurevych (TU Darmstadt)

Erhard W. Hinrichs (University of Tübingen)

27.06.2006

SALSA Workshop Saarbrücken

The Project

Dr. Iryna Gurevych
Project Leader, Principal Investigator
gurevych@tk.informatik.tu-darmstadt.de

Dr. Piklu Gupta
Research Associate
Acquisition of Lexical-Semantic Relations
gupta@sfs.uni-tuebingen.de

Prof. Dr. Erhard Hinrichs
Project Leader, Principal Investigator
eh@sfs.uni-tuebingen.de

M.Sc. Christof Mueller
Doctoral Researcher
Semantically Enhanced Information Retrieval
mueller@tk.informatik.tu-darmstadt.de

Prof. Dr. Max Mühlhäuser
Scientific Advisor
max@informatik.tu-darmstadt.de

Dipl.-Inf. Torsten Zesch
Doctoral Researcher
Computing Semantic Relatedness
zesch@tk.informatik.tu-darmstadt.de

Deutsche Forschungsgemeinschaft
DFG

funded by the German Research Foundation

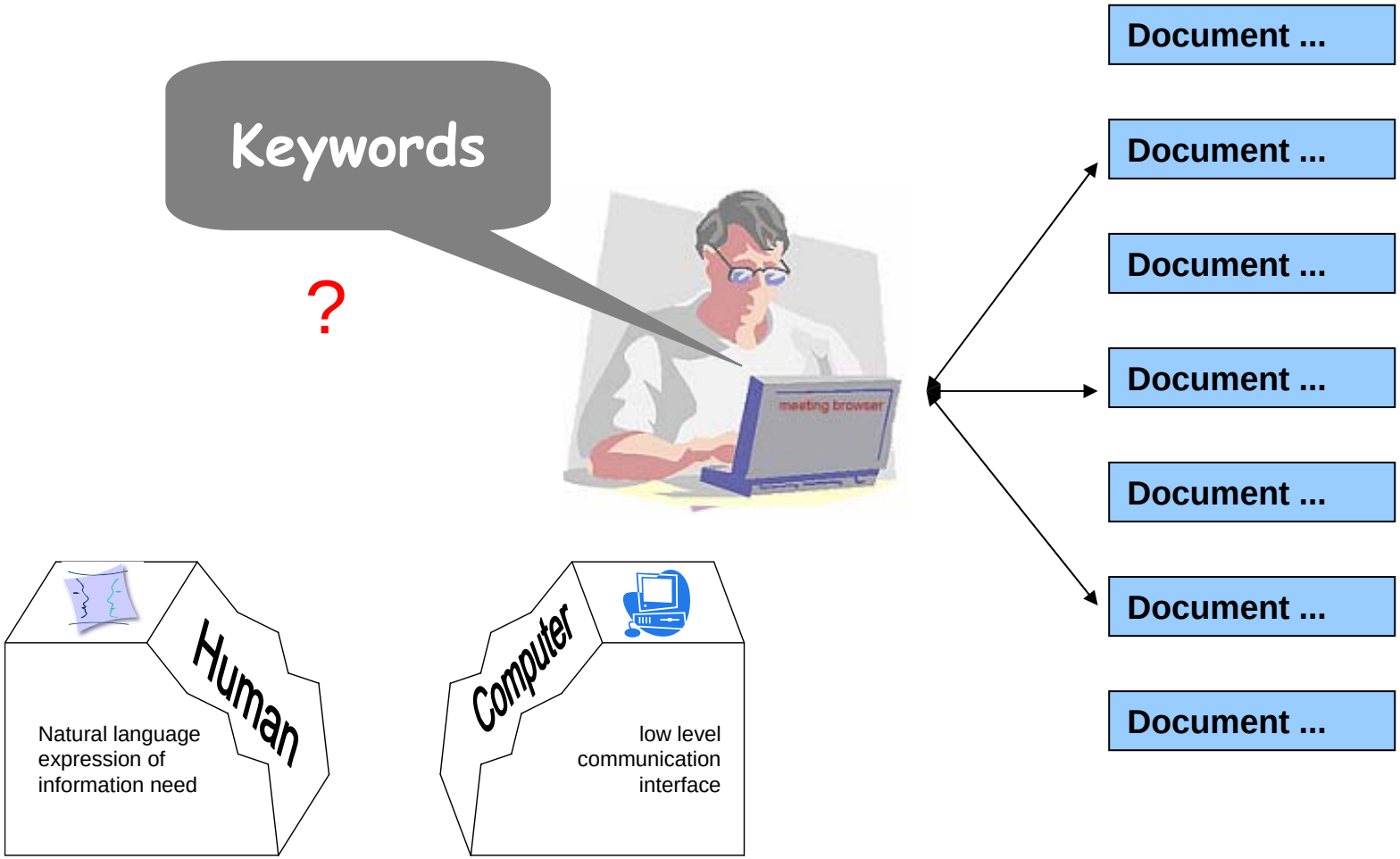
Bundesagentur für Arbeit

in cooperation with German Federal Employment Agency

Outline

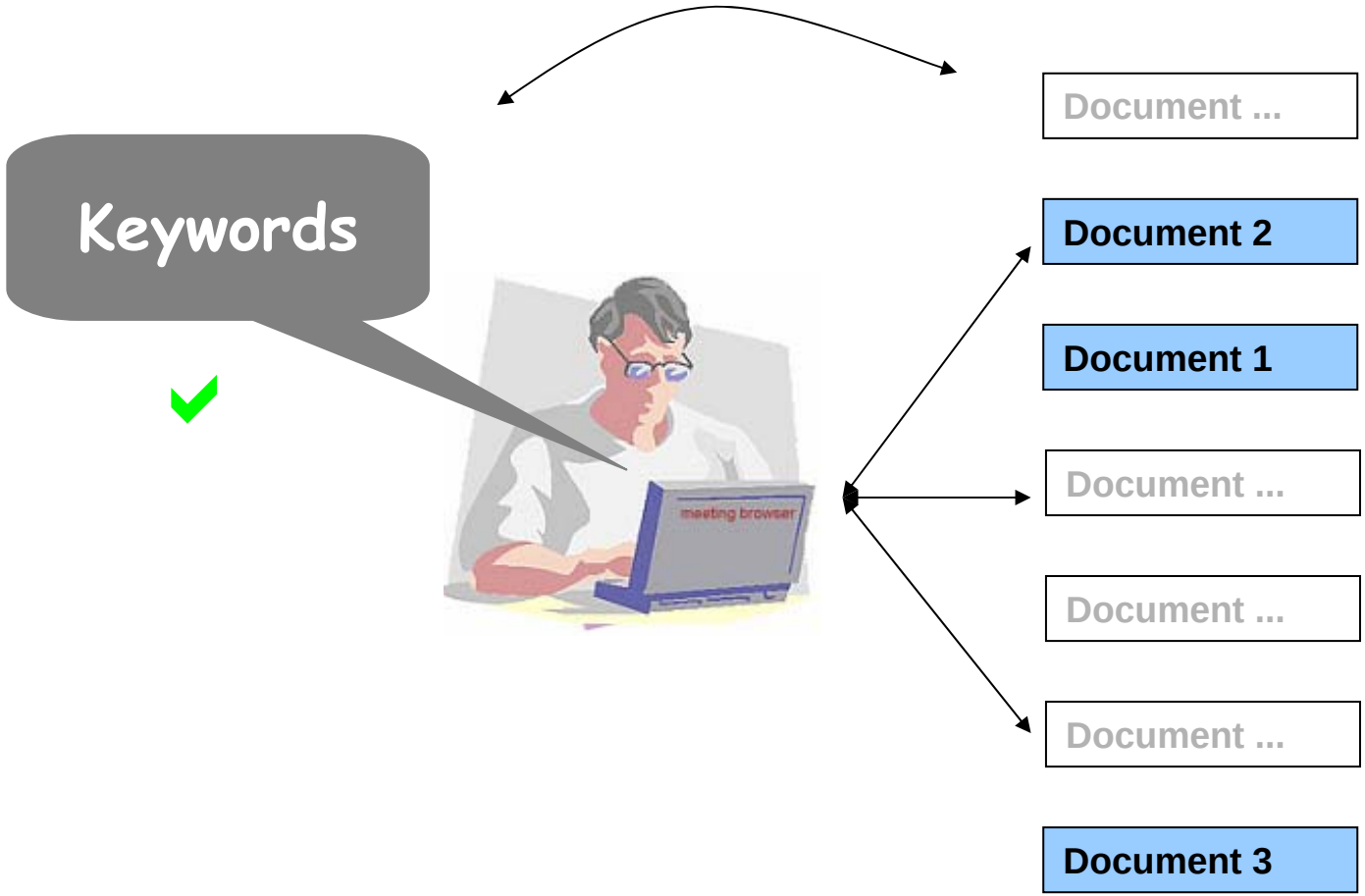
- Beyond Conventional Information Retrieval
- Semantic Relatedness Measures
- Extending GermaNet
- Summary

Information Retrieval (IR)



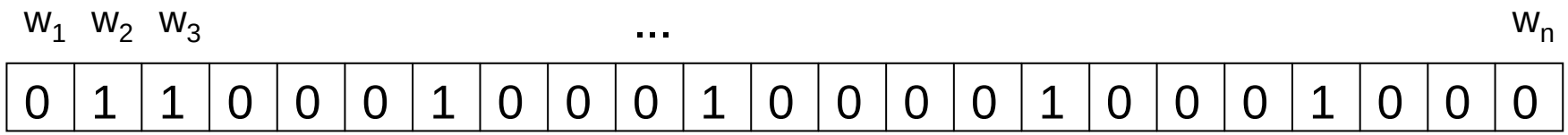
Information Retrieval (IR)

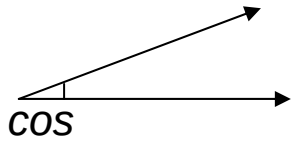
Boolean, Vector Space, ...



Beyond Conventional IR

- Go beyond conventional IR (“bag-of-words” model)

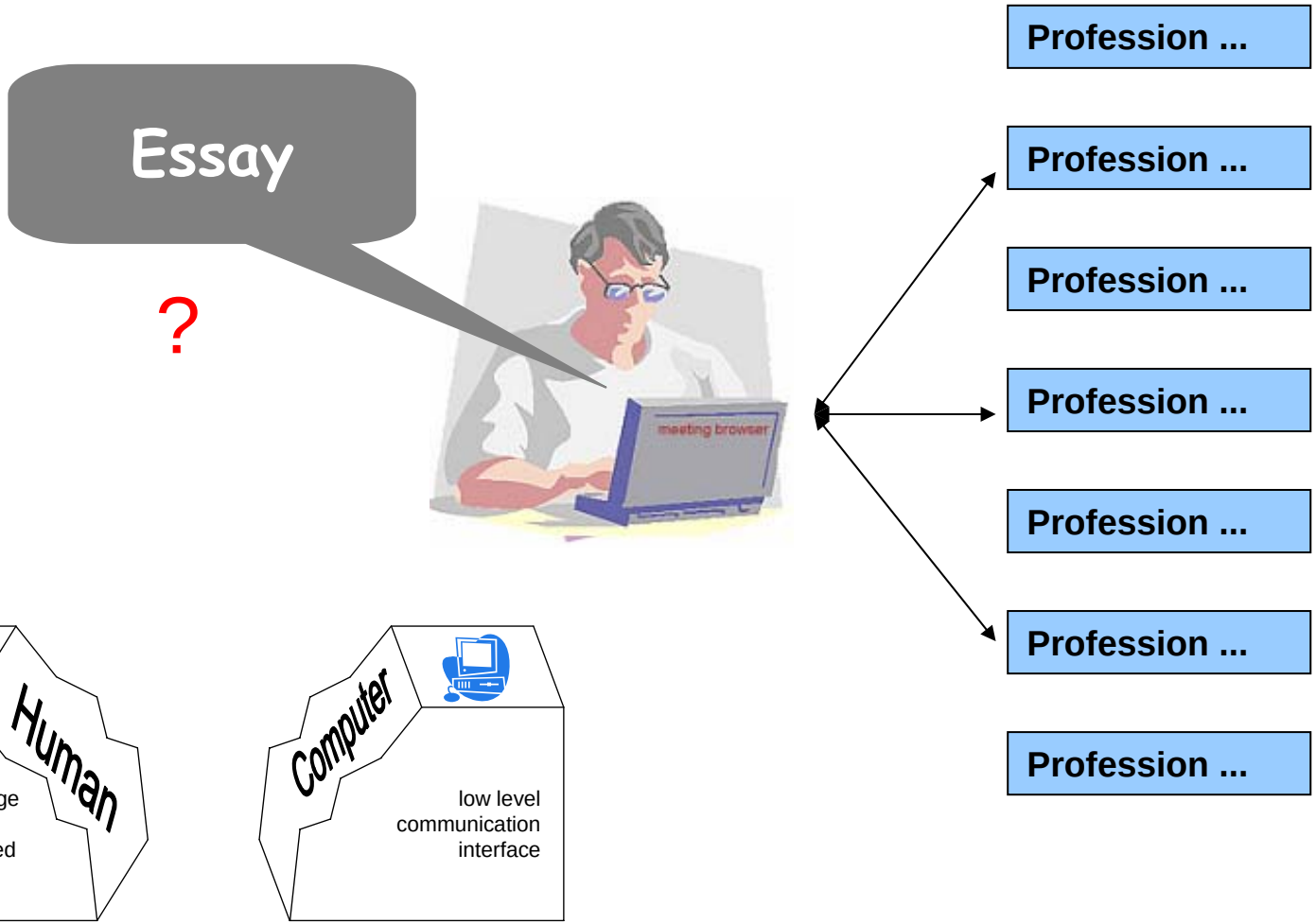




  - use relations between words
  - use knowledge encoded in ontologies

- Semantic Information Retrieval (SIR) approach:
  GermaNet + semantic relatedness measures

SIR-Project


Essay
?
meeting browser
Profession ...
Profession ...
Profession ...
Profession ...
Profession ...
Profession ...
Profession ...
Human
Natural language expression of information need
Computer
low level communication interface

SIR-Project

Semantic Relatedness

baker, to program, quality assurance

Profession ...

Profession 2

Profession 1

Profession ...

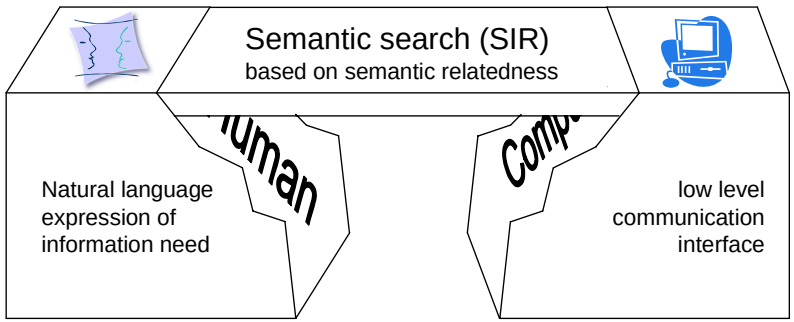
Profession ...

Profession ...

Profession 3

Essay

cake, computer, to read, ...



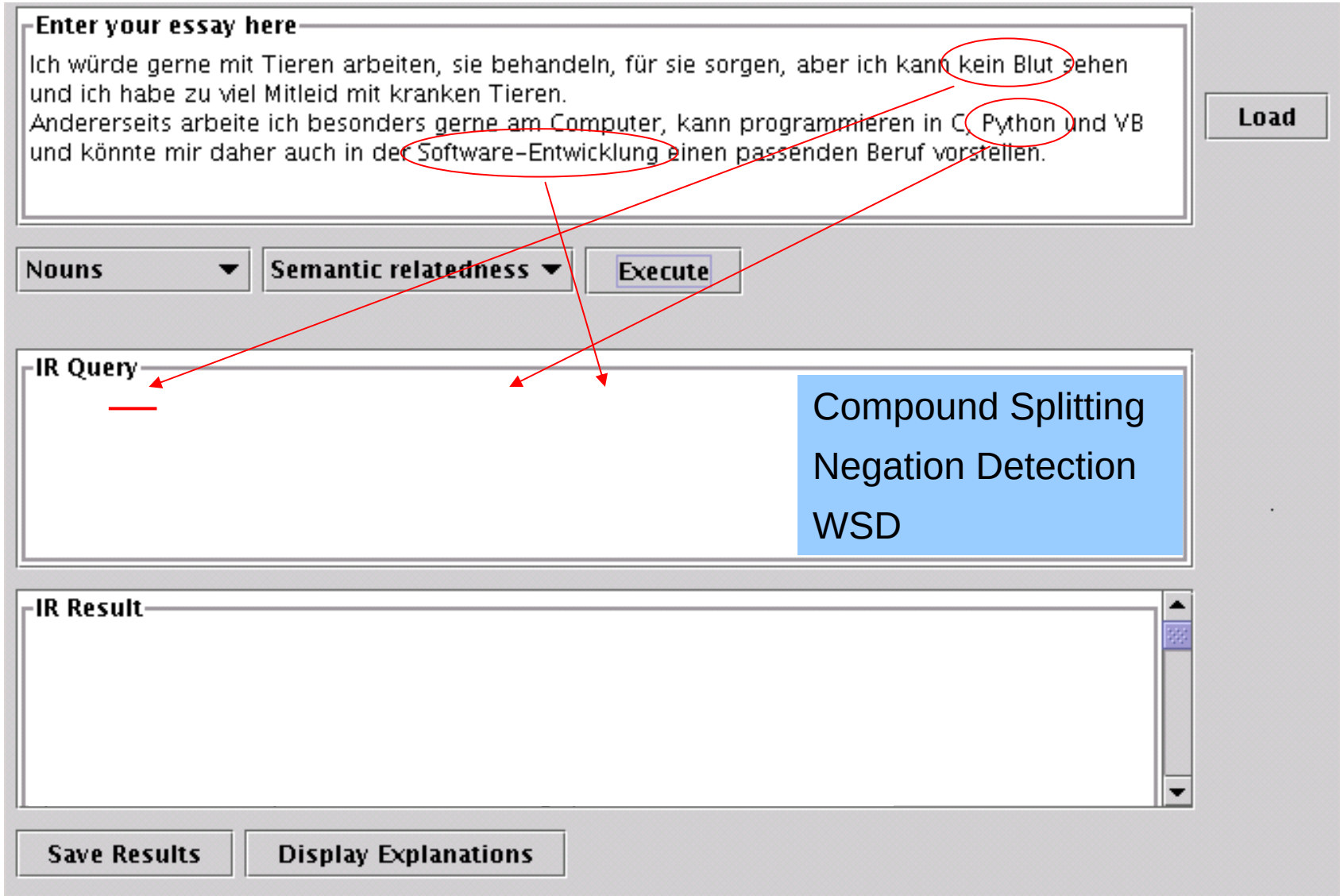
Hypothesis

- Semantic relatedness is a better model of relevance between query and document
- Why?
  - Essays and descriptions of professions often use different vocabularies to describe the same thing
  - Documents are fairly short → not enough context
  - Bag-of-words approaches are likely to fail
  - Deeper semantic analysis is required
- Pre-processing of queries error-prone → it IS a challenge, see example.

SIR Example

find good index terms

compute semantic relatedness

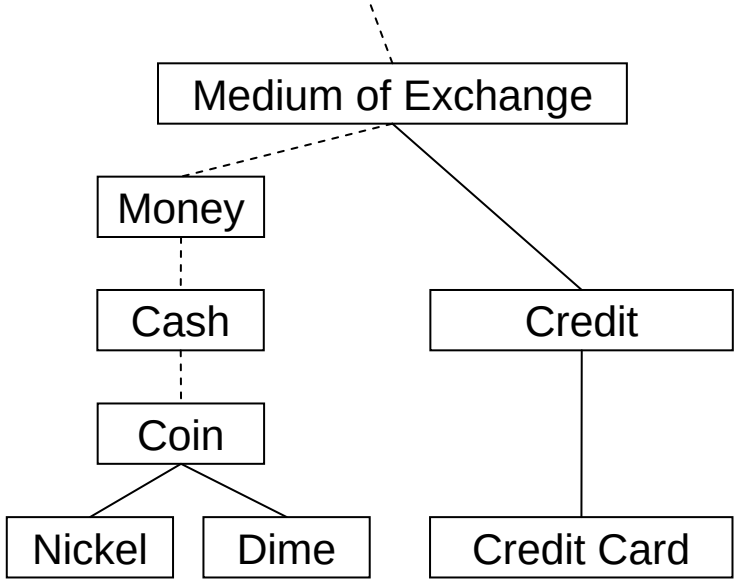


Semantic Relatedness

- Intuitive human judgments of semantic relatedness
- How to compute (quantify)?
- In this project → use GermaNet, a lexical-semantic wordnet for German

Semantic Similarity Measures

- Basic methods
  - using the structure of the ontology (Resnik95, Lin98, Hirst&St-Onge98, Leacock&Chodorow98)
  - using the definitions (Lesk86, Banerjee&Pedersen03, Gurevych05)

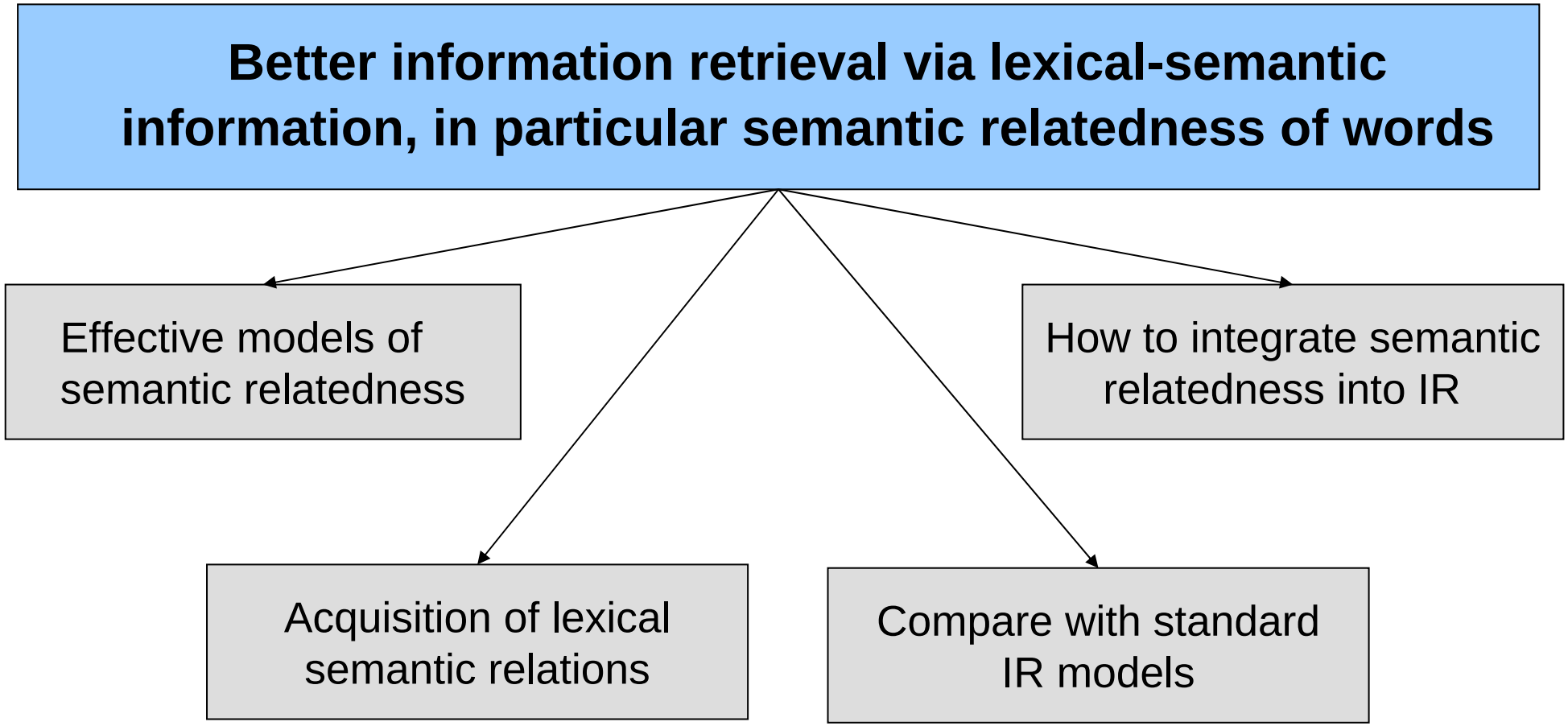


Credit card: a card (usually plastic) that assures a seller that the person using it has a satisfactory credit rating and that the issuer will see to it that the seller receives payment for the merchandise delivered

Dime: a United States coin worth one tenth of a dollar

Example: Resnik95

Research Goals



Particular motivations

Information retrieval needs:

- robust processing → broad coverage, reflect corpus reality
- rich representation of queries and documents → entities, events, attributes

This entails:

- extending / designing better metrics of semantic relatedness
  - work across word classes
  - broad coverage – use of distributional information or other knowledge sources/ontologies
- construction of datasets for testing and evaluation

bank – money (noun – noun)
bank – financial (noun – adj.)

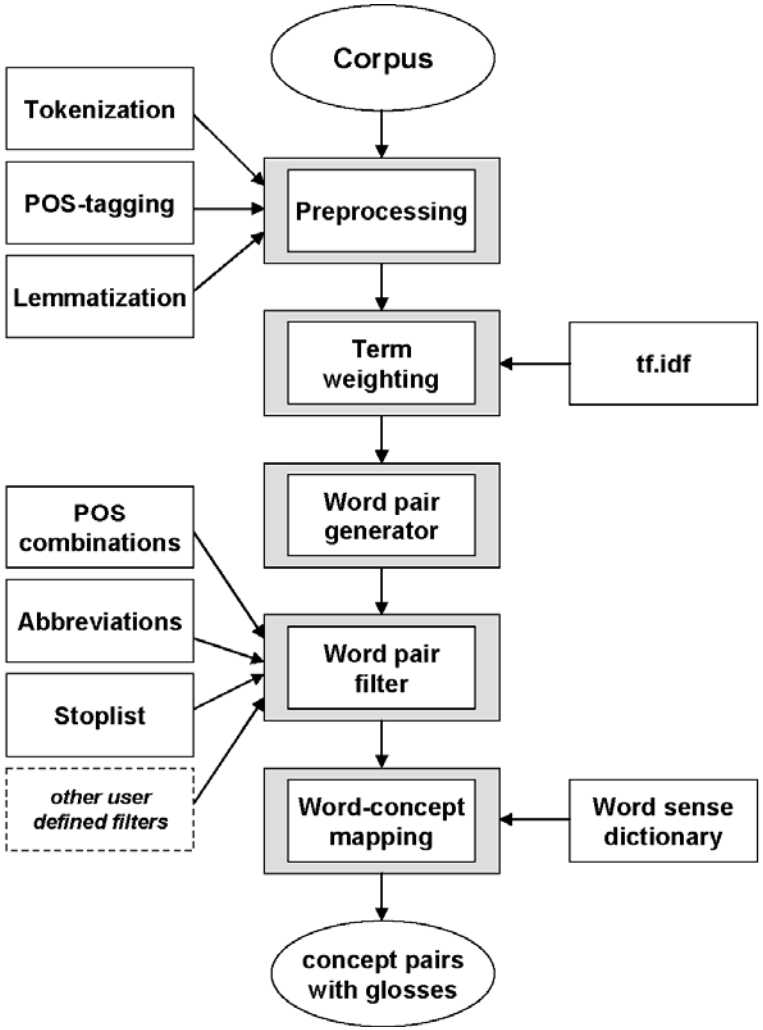
Experiments

Authors	Language	# Pairs	POS	Rel-Type	Scores	Subjects	Correlation	
							Inter	Intra
R/G	English	65	N	sim	continuous 0-4	51		.850
M/C	English	30	N	sim	continuous 0-4	38		
Res	English	30	N	sim	continuous 0-4	10	.903	
Fin	English	353	N, V, A	relat	continuous 0-10	16		
Gur	German	65	N	sim	discrete {0,1,2,3,4}	24	.810	
Gur	German	350	N, V, A	relat	discrete {0,1,2,3,4}	8	.690	
Z/G	German	328	N, V, A	relat	discrete {0,1,2,3,4}	21	.478	.647

Shortcomings of previous experiments:

- semantic similarity instead of semantic relatedness
- small number of word pairs
- manual selection of word pairs
- only noun-noun combinations
- only general terms
- annotated words instead of concepts

From Corpus to Dataset



Corpus	# Docs	# Tokens
BERUFEnet	9,022	7,728,501
GIRT	151,319	19,645,417
Scientific Presentations	106	144,074

Annotation GUI

Wortpaarbewertung
3 / 328

Bewerten

(nicht verwandt) 0 1 2 3 4 (stark verwandt)

	verarbeiten	**dichten**	
Synonyme	aufbereiten verarbeiten	dichten abdichten stopfen verstopfen	**Synonyme**
verwandte Wörter	wandeln bearbeiten verarbeiten aufbereiten verändern ändern	schließen stopfen abdichten wandeln verstopfen dichten verändern ändern	**verwandte Wörter**

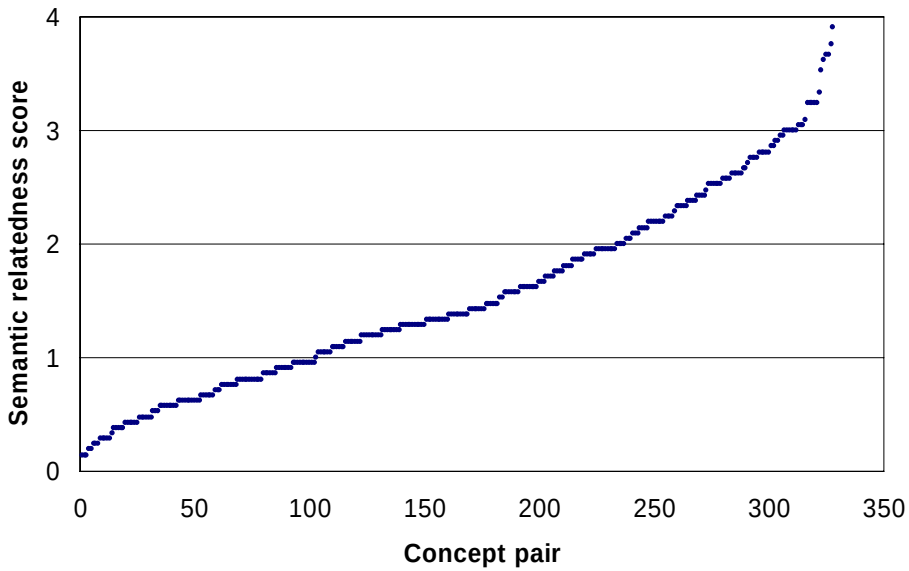
Wikipedia (Schlagen Sie hier nach, wenn Sie ein Wort nicht kennen.)

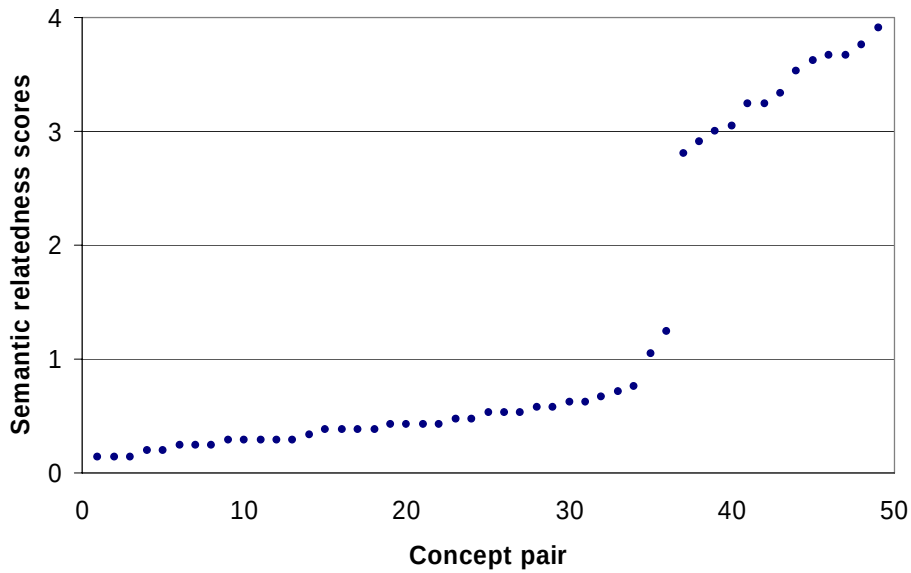
Hilfe (Den Einführungstext nochmals lesen.)

Logout

Results

- Medium-sized datasets can be created with reasonable effort
  - almost balanced (slight under-representation of highly related concepts)





only pairs with std-dev < 0.8

- Human judgments are most reliable for pairs that are closely related or not related at all

Results (2)

- Although distribution of judgments is balanced → few classical relations
- Most relations are non-classical (Morris & Hirst, 2004)
- Domain-specific vocabulary could not be found in GermaNet

→ increase coverage of GermaNet

GermaNet Example

Berufstätiger

Has lexical unit

Has hyponym

Hyp1 *Industriearbeiter*

Hyp2 *Metallarbeiter*

Hyp3 *Metallverbinder*

Schweißer

Current status of GermaNet

- At present GermaNet contains a total of 53,241 synsets and 76,415 lexical units
- Mainly general language vocabulary, only limited terminological coverage
- Relations mainly hold within one word class
- Standard relations include hyperonymy/hyponymy, meronymy/holonymy
- No details on selectional restrictions, semantic roles

Planned GermaNet enhancements

- Coverage of careers related terminology (e.g. job titles, work activities) to be extended
- Existing taxonomy of professions and qualifications to be integrated into GermaNet
- New relations spanning word classes to be introduced (e.g. semantic roles, selectional restrictions)

Modeling issues for GermaNet

- Specialist terminology to be integrated into GermaNet, but should also be possible to decouple it without adverse side-effects

- Trade-off between fine-grained modeling and keeping path length from hyperonym to hyponym consistent for purposes of similarity measures

- Currently investigating modeling strategies to ensure that new synsets are inserted at appropriate levels in hierarchy

Knowledge sources

- Central resource for extension of GermaNet is *Berufenet* specialist corpus of detailed career descriptions (c. 14.5 million tokens)
- Texts provide details of activities, qualifications required

Example text (translated)

Demographer

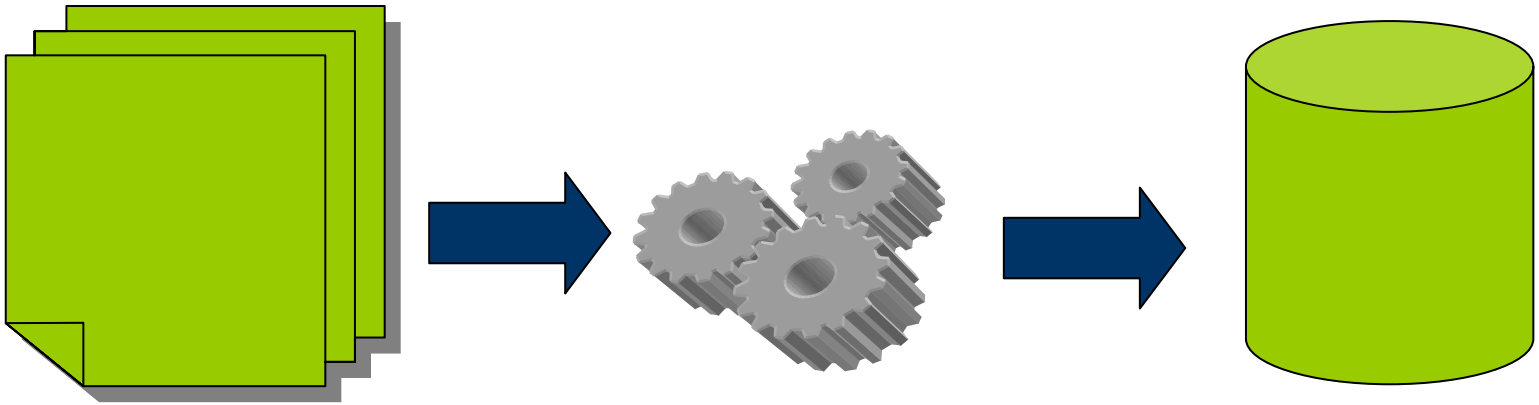
Overview of activities – Demographers work in the field of describing economic and socio-political movements within the population. They collect relevant statistical data, prepare and analyse them. Their work helps solve demographic problems and in preparing relevant decisions.

Access to profession – In order to work as a demographer, a university degree in a mathematical or statistical subject is required. Additional knowledge of economics or social sciences is advantageous. Along with analytical and combinatorial skills, the ability to communicate and to coordinate activities is required. Furthermore, experience of making presentations and knowledge of English are desirable.

Exploitation of knowledge sources

- Corpus was partially parsed using tools developed in Tübingen, annotation is in XML
- XML format of corpus allows the use of XML technologies (e.g. XSLT) for extraction of salient features
- Verbal and nominal heads extracted, verb/object pairs identified
- New GermaNet relations arrived at using unsupervised learning methods – soft clustering (Latent Semantic Clustering)

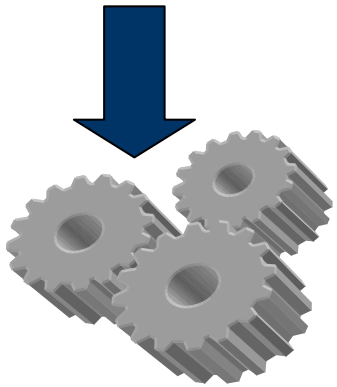
Processing pipeline



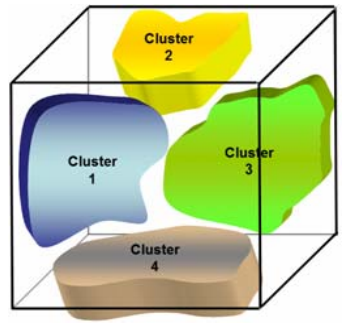
Unstructured texts

Morphological analysis, partial parsing of texts

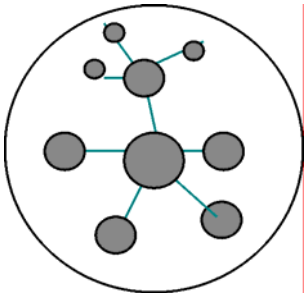
Syntactically annotated corpus



Extraction of verb/object, subject/verb pairs



Soft clustering (LSC)



Candidate new GermaNet relations spanning word classes

Semantic Relatedness <-> IR

Completely different (alternative) model of IR

Via query expansion → add semantically related vocabulary to query terms

Via term weights by lexical chaining → not all query terms are equally important

Prototype implemented, but experimental results depend on the extension of GermaNet

Evaluation of SIR

BerufeNet-based: compare professions ranked relevant by SIR to expert decisions

GIRT-based: standard benchmark, monolingual domain-specific task for German (CLIR)

E-Learning: semantic search in course materials (PowerPoint-based scientific presentations) → collaboration with Prof. Max Mühlhäuser, Digital Lecture Hall

Summary

- SIR – recently started DFG-Project:
  - Enhance information retrieval by lexical-semantic information
  - Measures of semantic relatedness at its core
  - Based on *extended* GermaNet, semi-automatic, corpus-based
  - Domain of career guidance, in collaboration with the German Federal Employment Agency

- Expected results:
  - Better models of semantic relatedness (coverage, accuracy, efficiency)
  - Breakthrough in Information Retrieval (beyond bag-of-words)
  - Corpus-based extension of GermaNet by domain-specific vocabularies

Thank you for your attention!