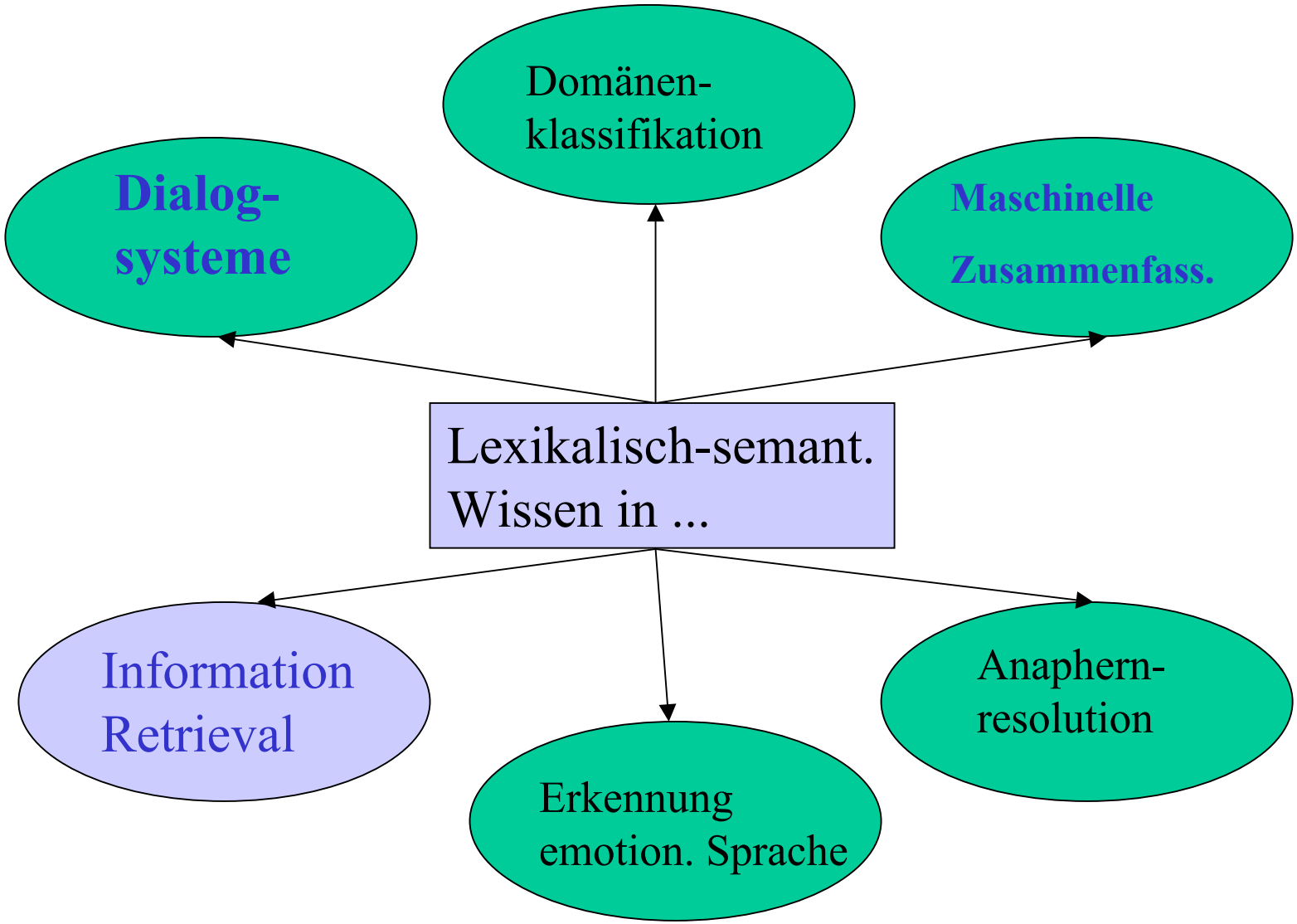
Erschließung großer Korpora mit Verfahren der lexikalischen Semantik

Dr. Iryna Gurevych
EML Research gGmbH
iryna.gurevych@eml-research.de
http://www.eml-research.de/~gurevych

Zu meiner Person ...

- Background
- Erfahrungen mit Heidelberger und Nicht-Heidelberger Studenten
- Wissenschaftliche Interessen und laufende Projekte

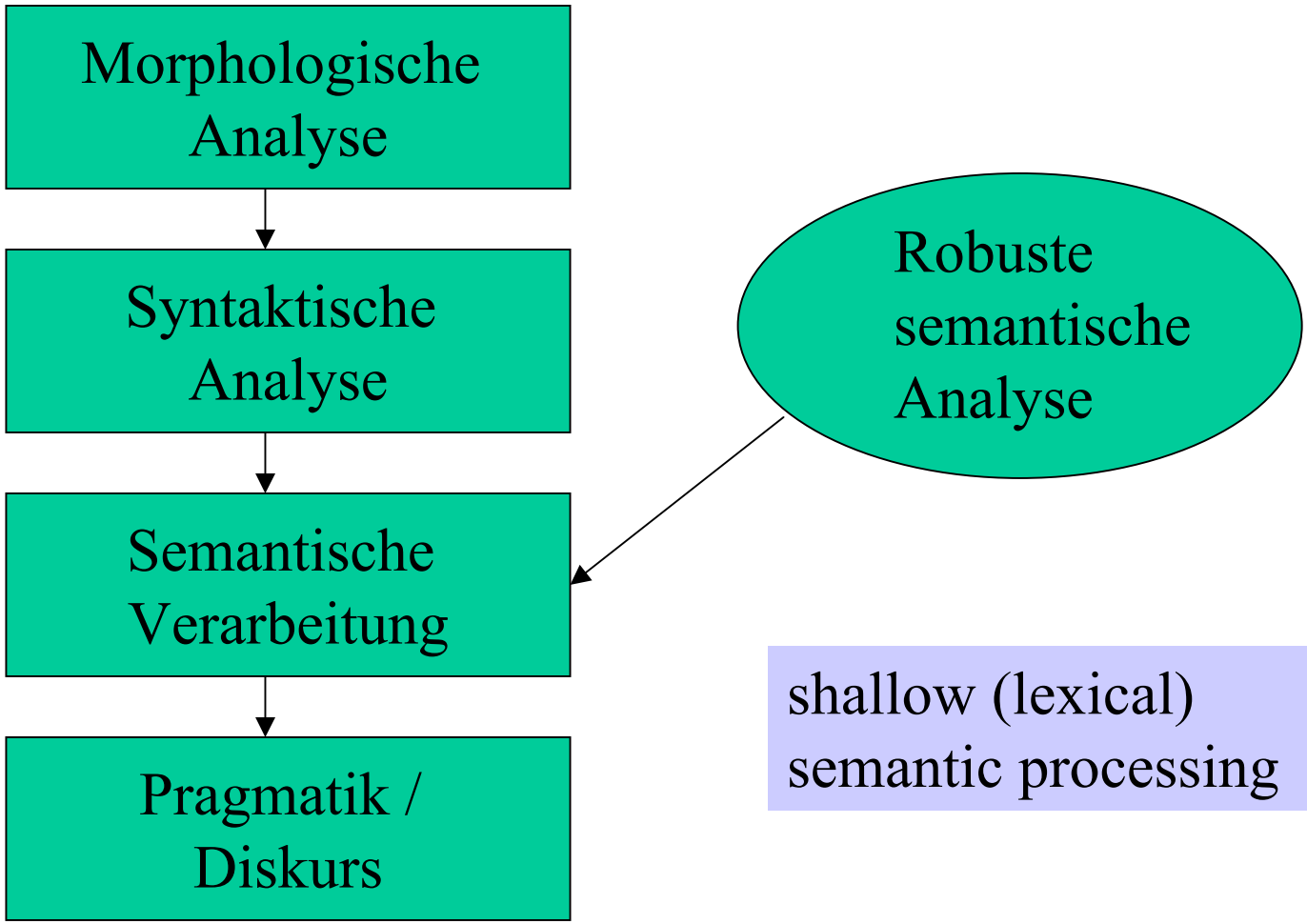
Eigene Forschung



Heute ...

- Vorstellung
- Lexikalisch-semantische Sprachverarbeitung
- Organisatorisches zum Seminar
- Einführung in mögliche Themen
- Ressourcen, Schwerpunke
- Fragebogen

Ebenen der automatischen Sprachverarbeitung



Verfahren der lexikalischenSemantik

1. Linguistische Grundlagen:
   - lexikalische Semantik
   - Wissensquellen (Thesauri, Ontologien, ...)
2. Informatik:
   - automatischer Zugriff auf die Wissensquellen
   - Anwendung des Wissens in Applikationen
3. Computerlinguistik:
   - korpus-basierte Arbeiten
   - Evaluierung der Verfahren

Lexical Semantic Processing

... applications that make use of word meanings, but which are to varying degrees decoupled from the more complex tasks of compositional sentence analysis and discourse understanding

Jurafsky & Martin, S. 631

Nächstes Mal ...

- Vorbesprechung der Themen
- Genaue Seminarplanung
- Lexikalische Semantik (ling. Grundlagen)
- Experimente über semantische Verwandtschaft
- SIR (GermaNet API, Maße semantischer Verwandtschaft, Demo)
- ...

Heute ...

- Vorstellung
- Lexikalisch-semantische Sprachverarbeitung
- Organisatorisches zum Seminar

Ziele des Seminars

- theoretische Grundlagen der lexikalischen Semantik kennenlernen;
- Anwendung der Verfahren lexikalischer Semantik auf große Korpora;
- Arbeit mit / Implementierung der Software im Bereich computationelle lexikalische Semantik;
- einzelne Themenstellungen strukturiert aufbereiten;
- Vorträge vorbereiten und Präsentationstechniken einüben;
- Kennenlernen verschiedener Aspekte der Verarbeitung und formalen und semantischen Modellierung natürlichsprachlicher Texte

Zentrale Frage

- Wie kann man NLP Anwendungen mit dem Wissen der lexikalischen Semantik verbessern?
  - Wortlesartendisambiguierung
  - Information Retrieval
  - maschinelle Textzusammenfassung
  - ...

Struktur der Lehrveranstaltung

1. Hälfte
   - Klärung der Begriffe
   - Was ist lexikalische Semantik?
   - Welche Wissensquellen gibt es?
   - Wie kann man auf sie in Programmen zugreifen?
   - Maße semantischer Ähnlichkeit / Verwandtschaft

Struktur der Lehrveranstaltung

EML Research gGmbH

2. Hälfte

- Information Retrieval (!)
- Wortlesartendisambiguierung
- Informationserschließung und Extraktion
- Rechtschreibkorrektur
- maschinelle Text-/Dialogzusammenfassung
- Dialogsysteme
- ...

Darüber hinaus...

- Korpus-basierte Arbeiten
- Annotationstools
- Wie kann ich eine NLP-Aufgabe definieren, wie kann ich die Daten annotieren, wie kann ich mein Programm evaluieren?
- Praxisbezogene Arbeit !! Nicht nur was, sondern WIE?

Empirisches Experiment

- Semantische Verwandtschaft der Wörter im Deutschen
- Ziel: einen wortartenübergreifenden Test-Datensatz für das Deutsche entwickeln
- Schritte:
  1. Wortpaare aus Texten auswählen;
  2. Nach bestimmten Merkmalen einen Datensatz zusammenstellen;
  3. Wortpaare bewerten (auf der Skala von 0 bis 4).
- Teilnahme am Experiment **Pflicht** für Seminar-teilnehmer

Scheinanforderungen

Proseminar: ein Thema nach Wahl bearbeiten

- Referat darüber halten
- Implementierung willkommen
- Eine Hausarbeit schreiben → Abgabe der Hausarbeit + Vortragsfolien

Hauptseminar: ein Thema nach Wahl bearbeiten

- Referat darüber halten
- Implementierung (nach Absprache)
- Eine Hausarbeit schreiben

  (Unterschrift durch Prof. Hellwig)

Seminar-Info

- **Kurs-Webseite:**

  http://www.eml-research.de/english/homes/gurevych/SS2005.html

- **Seminar-Materialien:**

  http://www.eml-research.de/english/homes/gurevych/Materialien.html

- **Die Web-Seite im Aufbau, wird laufend aktualisiert, Folien dort verfügbar**

Heute ...

- Vorstellung
- Lexikalisch-semantische Sprachverarbeitung
- Organisatorisches zum Seminar
- Einführung in mögliche Themen

Themen

Schwerpunkte:

- Maße semantischer Ähnlichkeit / Verwandtschaft
- Semantisches Wissen in Information Retrieval

Eine nicht deterministische Liste, weil:

- vielfältige Anwendungen lexikalischer Semantik in NLP;
- hängt von der Beteiligung und Interessen der Studierenden ab;
- Gruppenbildung, Projektdefinition → was wird im Projekt benötigt?

Mögliche Themen I

- WordNet, WordNet Domains, SUMO/WordNet linking, MultiWordNet ...
- GermaNet (Deutsch)
- OpenThesaurus (Deutsch)
- EuroWordNet
- Roget's Thesaurus
- DMOZ
- Eigene Themenvorschläge möglich und willkommen!

Mögliche Themen II

- Leacock C. and Chodorow M. 1998. **Combining local context and WordNet similarity for word sense identification.** In Fellbaum 1998, pp. 265-283.
- Jiang J. and Conrath D. 1997. **Semantic similarity based on corpus statistics and lexical taxonomy.** In *Proceedings of International Conference on Research in Computational Linguistics,* Taiwan.
- Resnik P. 1995. **Using information content to evaluate semantic similarity.** In *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* pages 448-453, Montreal.
- Resnik P. 1999. **Semantic Similarity in a Taxonomy: An Information- Based Measure and its Applications to Problems of Ambiguity in Natural Language.** *Journal of Artificial Intelligence Research*, 11, 95-130.

Mögliche Themen III

- Lin D. 1998. **An information-theoretic definition of similarity.** In *Proceedings of the 15th International Conference on Machine Learning,* Madison, WI.
- Hirst G. and St-Onge D. 1998. **Lexical Chains as representations of context for the detection and correction of malapropisms.** In Fellbaum 1998, pp. 305-332.
- Wu Z. and Palmer M. 1994. **Verb Semantics and Lexical Selection.** In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics.* Las Cruces, New Mexico.
- Banerjee S. and Pedersen T. 2002. **An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet.** In *Proceeding of the Fourth International Conference on Computational Linguistics and Intelligent Text Processing (CICLING-02).* Mexico City.

Semantische Beziehungen in GermaNet

- Iryna Gurevych and Hendrik Niederlich. 2005. **Computing semantic relatedness of GermaNet concepts.** In Fisseni et al. (Hrsg.) "Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen, Beträge zum Workshop *Anwendungen des GermaNet II*", 31 March 2005, Bonn, S. 462-474.

- Java API für GermaNet, und Semantic Relatedness Software (Java) auf Anfrage bei mir erhältlich

- Fünf unterschiedliche Maße semantischer Verwandtschaft implementiert, GUI

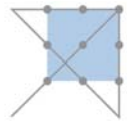
Wortlesartendisambiguierung

- Schütze H. 1998. **Automatic Word Sense Discrimination.** *Computational Linguistics*, 24(1):97-123.
- Patwardhan S., Banerjee S. and Pedersen T. 2002. **Using Semantic Relatedness for Word Sense Disambiguation.** In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.
- Jäger, Jutta. 2005. **Wortlesartendisambiguierung in einem System zur automatischen Dialogzusammenfassung.** Magisterarbeit an der Uni Heidelberg und EML Research gGmbH (Betr. Iryna Gurevych)

Rechtschreibkorrektur

- Hirst, Graeme and Budanitsky, Alexander. **Correcting real-word spelling errors by restoring lexical cohesion.** Submitted for publication. (http://www.cs.toronto.edu/compling/Publications/h-i.html#Hirst)

Informationserschließung und Extraktion



- Informationserschließung und Extraktion. Gonzalo, J., F. Verdejo, I. Chugur and J. Cigarrán (1998) **Indexing with WordNet synsets can improve text retrieval,** in *ACL/COLING Workshop on Usage of WordNet for Natural Language Processing*.
- Sanderson, M. (1994) **Word Sense Disambiguation and Information Retrieval**, *SIGIR'94.*
- Richardson, R. and Smeaton, A. (1995) **Using WordNet in a Knowledge-Based Approach to Information Retrieval, Dublin City Univ**. *TR CA-0395*.
- Smeaton, A. and Quigley, I. (1996) **Experiments on Using Semantic Distances between words in Image Caption Retrieval**, *SIGIR'96*.

Text-/Dialogzusammenfassung

- Iryna Gurevych and Michael Strube. 2004. **Semantic Similarity Applied to Spoken Dialogue Summarization.** In *Proceedings of COLING*, Geneva, Switzerland, August.

Dialogsysteme

- Iryna Gurevych, Rainer Malaka, Robert Porzel, Hans-Peter Zorn. 2003. **Semantic coherence scoring using an ontology.** In *Proceedings of the Joint Human Language Technology and Northern Chapter of the Association for Computational Linguistics Conference (HLT-NAACL)*, Edmonton, Canada, p.p. 88 - 95.

Anaphernresolution

- Iryna Zhmaka. 2005. **Auflösung der Pronomen mit Nicht-NP-Antezedenten in spontansprachlichen Dialogen.** Magisterarbeit an der Uni Heidelberg und EML Research gGmbH (Betr. Iryna Gurevych)
- Iryna Zhmaka. 2005. **Auflösung der Pronomen mit Nicht-NP-Antezedenten in spontansprachlichen Dialogen.** In Beiträgen zur GLDV-Tagung 2005 "Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen". Bonn, 30. März – 1. April, S. 619-632.

Heute ...

- Vorstellung
- Lexikalisch-semantische Sprachverarbeitung
- Organisatorisches zum Seminar
- Einführung in mögliche Themen
- Ressourcen, Schwerpunke

Literatur

- Grundlegende Bücher
- Papiere, meistens online verfügbar
- In einer Vorbesprechung kann ich weitere Hinweise geben
- Von mir betreute Magisterarbeiten auf Anfrage bei mir

Grundlegende Literatur

- Cruse, D.A. **Lexical Semantics.** (Cambridge Textbooks in Linguistics). Cambridge University Press, 1997.
- Fellbaum C. (editor) **WordNet: An electronic lexical database.** MIT Press, 1998.
- Green, R., Bean, C.A., Myaeng S.H. (editors) **The Semantics of Relationships: An Interdisciplinary Perspective.** Kluwer Academic Publishers, 2002.
- Kunze, C., Lemnitzer, L., Wagner, A. (editors) **Proceedings of the Workshop "GermaNet: Application of the German wordnet in Theory and Practice".** 9.-10. Oktober, Tübingen, 2003. Hier eine online Version.
- Fisseni, B., Schmitz, H.-C., Schröder, B., Wagner, P. (editors) **Proceedings of the Workshop "Applications of GermaNet II"** in Sprachtechnologie, mobile Kommunikation und linguistische Resourcen, volume 8 of *Computer Studies in Language and Speech*. Peter Lang, Frankfurt am Main, 2005.

Software

- WordNet: s. WordNet Web-Seite, API in mehreren Programmiersprachen, viele zusätzliche Anwendungen
- WordNet:Similarity Package von Ted Pedersen et al., s. Web

  http://search.cpan.org/dist/WordNet-Similarity (Perl)
- GermaNet API & semantic relatedness package, auf Anfrage bei mir (Java/XML)
- ...

Schwerpunktthema: semantisches Wissen in IR

- Zentrales Forschungsziel: Semantic Information Retrieval
- Anwendungsdomäne: elektronische Berufsberatung
- Sprache Deutsch → GermaNet
- Programmiersprache → Java
- Daten: 30 natürlichsprachliche Profile, ca. 600 textuelle Berufsbeschreibungen (Bundesagentur für Arbeit)

Stand des Projekts I

- Daten morphologisch und syntaktisch analysiert:
- externe Software, Integration in XML-Format (Praktikum Steffen Eger)
- Eine Evaluationsumgebung incl. Gold-Standard (Praktikum Sinian Zhang) → Erfolg messen!
- Semantische Annotationen, positive und negative emotionale Einstellungen via GermaNet (Magisterarbeit Vesna Cvoro)

Stand des Projekts II

- NLP in Information Retrieval (IR) und statistische Baselines → Diplomarbeit Hendrik Niederlich
- Semantische Query Expansion via GermaNet:

  eigene Experimente fortlaufend
- Offene Punkte (semantisches Wissen in IR):
  - IR via Maße semantischer Verwandtschaft
  - Integration lexikalischer Ketten in die Indexierung
  - Wortlesartendisambiguierung → großes Thema !!
  - ...

Vorgehensweise

- Wer Interesse hat, im Rahmen dieses Projekts eine Seminararbeit zu gestalten, bei mir melden
- Anmeldung für weitere vorgeschlagene Themen an mich bis 26.04 mailen
- Eigene Vorschläge, etc. ebenso bis 26.04 bitte mailen
- Eine Vorbesprechung am 27.04 von 13 bis 14 Uhr möglich, bitte anmelden!
- Nun Fragebogen ...