Representing Underspecification in a Relational Database

Kerstin Eckart
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, Azenbergstraße 12, 70174 Stuttgart

Dealing with (syntactic) ambiguity in natural language processing

Alternative strategies

- Selecting a unique analysis
  - Reading intended by the author may be lost
  - Information about the ambiguity is lost and can therefore not be exploited (e.g. for linguistic research on ambiguities)
  - Requires great deal of effort for disambiguation, which is in some cases of no use for subsequent processing
- Spelling out all reading alternatives
  - Requires great deal of effort and disk space, where only some alternatives will be relevant for subsequent processing
- Partial analysis
  - Existing knowledge about ambiguity not represented in analysis (e.g. details about alternatives, ambiguity type)

⇒ Preferred representation: Underspecification

Underspecification

- As a concept of representation:
  - All readings of a linguistic object can be deduced
  - No additional or incorrect readings can be deduced
- Upon acquisition of new knowledge (e.g. context)
  - Transferring the underspecified representation into a new (underspecified) representation via partial specification
  - → Removing readings without need for full specification

⇒ Representation of a given level of system knowledge

Objectives

- Representing ambiguity in an efficient representation format: Underspecification
- Mapping the representation format onto data structures of a relational database mangement system

A constraint based representation for underspecified analyses

LAF/GrAF based encoding scheme by [Kountz et al. 2008]

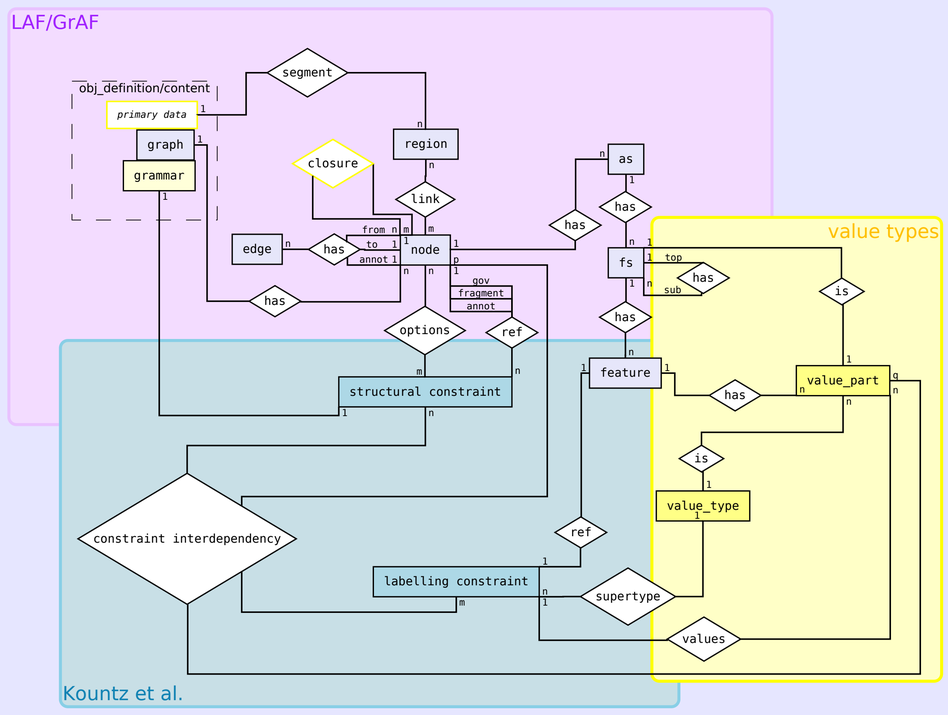
- **Structural Constraints** encode structural ambiguity:
  Ich sehe [den Mann [mit dem Fernglas]].
  Ich sehe [den Mann] [mit dem Fernglas].
  Relating partial, non-ambiguous structures
  ⇒ Constraint instanciated by edges
- **Labelling Constraints** encode labelling ambiguity:
  Peter$_{SUBJ|OBJ}$ kennt Karl$_{OBJ|SUBJ}$
  Defining annotation alternatives for structural elements
  ⇒ Constraint instanciated by atomic labels
- **Constraint Interdependencies** encode interdependent ambiguities:
  Karl fügte [einige Gedanken] [zu dem Werk$_{PP/zu|ADJUNCT}$] hinzu.
  Karl fügte [einige Gedanken [zu dem Werk$_{ADJUNCT}$]] hinzu.
  Defining interdependencies between instanciations of constraints

A database schema based on an upcoming ISO-standard

Linguistic Annotation Framework (LAF)
ISO/DIS 24612 (2009)

- Nancy Ide, Laurent Romary [Ide/Romary 2006]
- Data (meta)model to represent primary data and annotations
- References to primary data (e.g. character offsets for text)
- Stand-off annotation
- XML-Serialisation: **GrAF** [Ide/Suderman 2007]
  - Directed graph structure (nodes, edges)
  - Primary data segments are referenced by nodes
  - Annotations (feature structures) attached to nodes

⇒ Infrastructure for data exchange
⇒ Infrastructure for comparison and merging of different analyses, to increase reliablity



Relational Database Management System (RDMS)

RDMS provides off-the-shelf

- Efficient data processing
- Huge amounts of data can be handled
- Processing optimisation regarding query type
- Flexible queries

Using a data model for underspecification in a RDMS

Combining the advantages of both concepts

- Efficient queries on large amounts of data
- Extracting data on ambiguity using SQL-Queries or SQL query templates
- Disambiguation only when explicitly needed

Querying ambiguous analyses

Three types of queries

1. Searching explicitly for (non-resolved) ambiguities, e.g. with a view to
  - Data extraction regarding hypotheses on ambiguity phenomena
  - Tool development (*Which information is missing?*)

Query example: All sentences containing an ambiguous element which is in genitive case in one of its possible readings; extracted along with its lemma and the lemma of its governor (*dependency relation*).

Ausgabefeld

Datenanzeige | Zerlegung | Meldungen | Historie

	dep_lemma text	gov_lemma text	alternative text	sentence text
2	Weisheit	Positivismus/seinA/sollenI	NP:4	Es ist ja auch ein_bißchen unbefriedigend für einen Menschen unserer Gesellschaft , wenn der Weisheit letzter Schluss der P
3	Gesellschaft	wert/seinA	NP:4	Je älter das Kind , umso mehr ist es der Gesellschaft wert - und umso mehr dürfen die Menschen kosten , die sich um seine E
4	Mensch	verwehrenP/bleibenA	NP:4	Damit sei wieder ein Jahr verstrichen , in dem vielen jungen Menschen eine berufliche Perspektive verwehrt bleibe .
5	Mensch	verwehrenP/bleibenA	NP:8	Damit sei wieder ein Jahr verstrichen , in dem vielen jungen Menschen eine berufliche Perspektive verwehrt bleibe .

2. Searching explicitly for non-ambiguous data

Query example: All sentences containing a non-ambiguous element in genitive case; extracted along with its lemma and the lemma of its governor (*dependency relation*).

Ausgabefeld

Datenanzeige | Zerlegung | Meldungen | Historie

	dep_lemma text	gov_lemma text	sentence text
1	Mensch	Begleiter/seinA	Dabei waren sie über viele Generationen hinweg der Menschen treueste und arbeitswilligste Begleiter .
2	Schutz	bedürfen	Besonders Kinder und ältere Menschen bedürfen des Schutzes , vor_allem vor Schulen , Kindergärten und Altenheimen heißt es für Autofahrer :
3	Mensch	Alp#@traum/seinA	Im Zentrum aller Fassbinder-Filme steht die Annahme , dass der Mensch des Menschen Alptraum sei .
4	Mensch	bedürfen	Es bedarf dazu engagierter und risikofreudiger Menschen , die ihr Know-how anwenden und den Mut aufbringen , eigene selbständige Wege zu ge
5	Interesse	bedürfen	Freundschaft zur Welt als Sorge um die von verschiedenen Menschen bewohnte Welt bedarf eines Interesses , das die Welt zum Gegenstand de
6	Idee|Mensch	bedürfen	Denn es bedarf nicht nur einer Idee , sondern auch eines Menschen , der sie umsetzt&#x201C ; , so der Bürgermeister .
7	Mensch	würdig/seinA	Das ist eines intellektuellen Menschen nicht würdig , wie sich unser Bürgermeister da verhalten hat .
8	Beweis	bedürfen/habenP	Wenn es noch eines Beweises bedurft hätte , dass die deutschen Medienmacher mindestens so selbstbezüglich und abgehoben von den Interes

3. Queries without ambiguity restrictions

Data of both types (with and without ambiguities) → Providing larger amounts of data

Query example: All sentences containing an element which is in genitive case in at least one of its possible readings; extracted along with its lemma and the lemma of its governor (*dependency relation*).

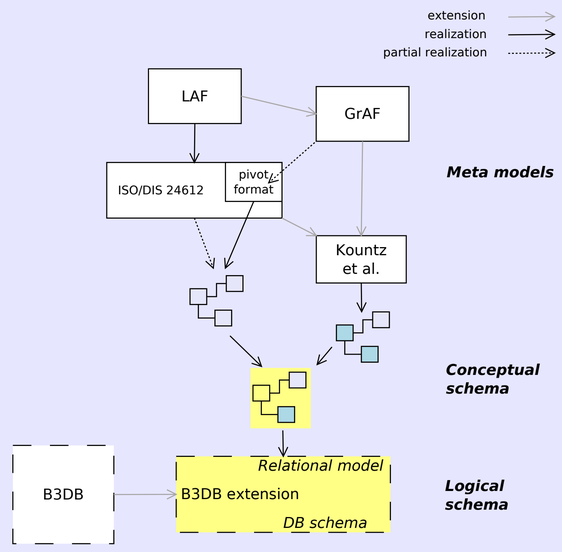
Ausgabefeld

Datenanzeige | Zerlegung | Meldungen | Historie

	dep_lemma text	gov_lemma text	sentence text
2	Mensch	Begleiter/seinA	Dabei waren sie über viele Generationen hinweg der Menschen treueste und arbeitswilligste Begleiter .
3	Gesellschaft	wert/seinA	Je älter das Kind , umso mehr ist es der Gesellschaft wert - und umso mehr dürfen die Menschen kosten , die sich um seine Bildung und Erziehung
4	Schutz	bedürfen	Besonders Kinder und ältere Menschen bedürfen des Schutzes , vor_allem vor Schulen , Kindergärten und Altenheimen heißt es für Autofahrer : *
5	Mensch	Alp#@traum/seinA	Im Zentrum aller Fassbinder-Filme steht die Annahme , dass der Mensch des Menschen Alptraum sei .
6	Mensch	bedürfen	Es bedarf dazu engagierter und risikofreudiger Menschen , die ihr Know-how anwenden und den Mut aufbringen , eigene selbständige Wege zu gel
7	Interesse	bedürfen	Freundschaft zur Welt als Sorge um die von verschiedenen Menschen bewohnte Welt bedarf eines Interesses , das die Welt zum Gegenstand des
8	Idee|Mensch	bedürfen	Denn es bedarf nicht nur einer Idee , sondern auch eines Menschen , der sie umsetzt&#x201C ; , so der Bürgermeister .
9	Mensch	würdig/seinA	Das ist eines intellektuellen Menschen nicht würdig , wie sich unser Bürgermeister da verhalten hat .

Data for the examples were processed by FSPar [Schiehlen 2003].
Database GUI displaying query results: Pgadmin3.

Technical aspects



- Entity-Relationship Models (ERM) for LAF and the LAF/GrAF-based encoding scheme of [Kountz et al. 2008]
- ERMs merged into a single conceptual schema for the database
- Structure for typing of annotation feature-values
- Linked to an existing database B3DB [Eberle et al. 2009] (collaborative research center SFB 732)

Implementation

- PostgreSQL relational database system
- Applying NestedSet model to represent type hierachies cf. [Celko 2004]
  → no recursive queries for tree-like structures needed

Framework

- Diplomarbeit *Repräsentation von Unterspezifikation in relationalen Datenbanksystemen* (08/2009)
- Institut für Parallele und Verteilte Systeme, Universität Stuttgart
- Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
- Collaborative research center (Sonderforschungsbereich) 732: Incremental Specification in Context
  - Projekt B3: Disambiguierung von Nominalisierungen bei der Extraktion linguistischer Daten aus Corpustext
  - Laufzeit: 1. Juli 2006 – 30. Juni 2010
  - `http://www.uni-stuttgart.de/linguistik/sfb732/`

References

[Celko 2004] Joe Celko. Trees and Hierarchies in SQL. Morgan Kaufmann. 2004

[Eberle et al. 2009] Kurt Eberle, Kerstin Eckart and Ulrich Heid. Eine Datenbank als multi-engine für Sammlung, Vergleich und Berechnung möglichst verlässlicher unterspezifizierter syntaktisch/semantischer Satzrepräsentationen. Poster presentation at the DGfS 2009, Osnabrück. 2009

[Eckart 2009] Kerstin Eckart. Repräsentation von Unterspezifikation in relationalen Datenbanksystemen. Diplomarbeit. Universität Stuttgart. 2009

[Ide/Romary 2006] Nancy Ide and Laurent Romary. Representing Linguistic Corpora and Their Annotations. In: Proceedings of the Fifth Language Resources and Evaluation Conference, LREC 2006. Genoa, Italy. 2006

[Ide/Suderman 2007] Nancy Ide and Keith Suderman. GrAF: A Graph-based Format for Linguistic Annotations. In: Proceedings of the Linguistic Annotation Workshop, held in conjunction with ACL 2007, Prague. 2007

[ISO/DIS 24612] Language resource management - Linguistic annotation framework (LAF). 2009.

[Kountz et al. 2008] Manuel Kountz, Ulrich Heid and Kerstin Eckart. A LAF/GrAF based Encoding Scheme for underspecified Representations of syntactic Annotations. In: Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC'08 (CD-ROM). Marrakech, Morocco. 2008.

[Schiehlen 2003] Michael Schiehlen. A Cascaded Finite-State Parser for German. In: EACL'03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, S. 163-166. Budapest, Hungary. 2003.

PostgreSQL: `http://www.postgresql.org/` Pgadmin: `http://www.pgadmin.org/`