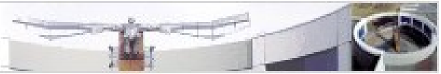


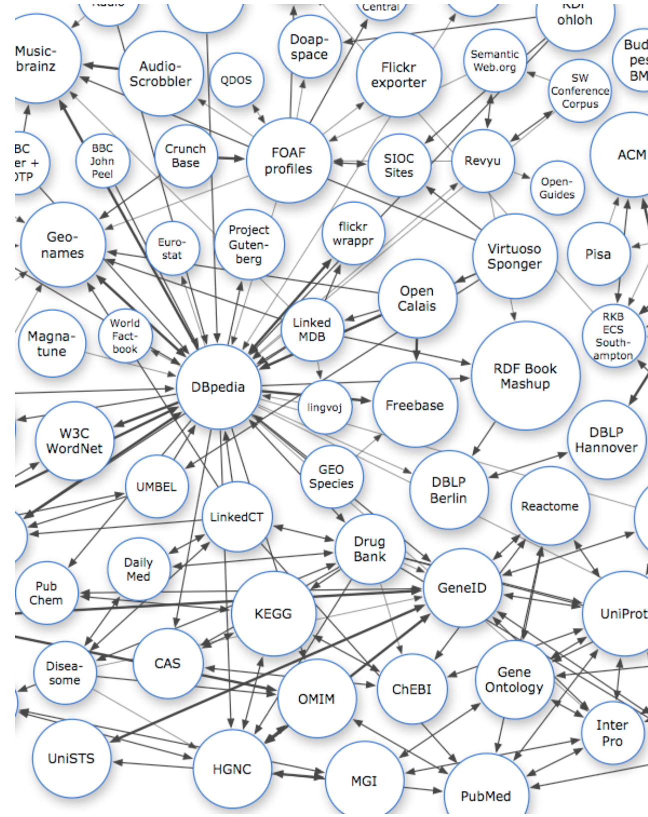
Question Answering Biographic Information and Social Networks Powered by the Semantic Web

Peter Adolphs, Xiwen Cheng, Tina Klüwer, Hans Uszkoreit & Feiyu Xu
German Research Center for Artifical Intelligence (DFKI)
Language Technology Lab

Presenter: Peter Adolphs
peter.adolphs@dfki.de

Motivation

- Semantic Web:
  - "The Semantic Web will bring structure to the meaningful content of Web pages" (Berners-Lee et al, 2001)
  - Today: genuine Semantic Web resources + Semantic Web versions of large, sometimes community-driven databases and websites
- Our questions:
  - How can we use these data in an knowledge-intensive AI applications?
  - How can we acquire such data from the Web?
  - How can we interface Semantic Web data with the human?

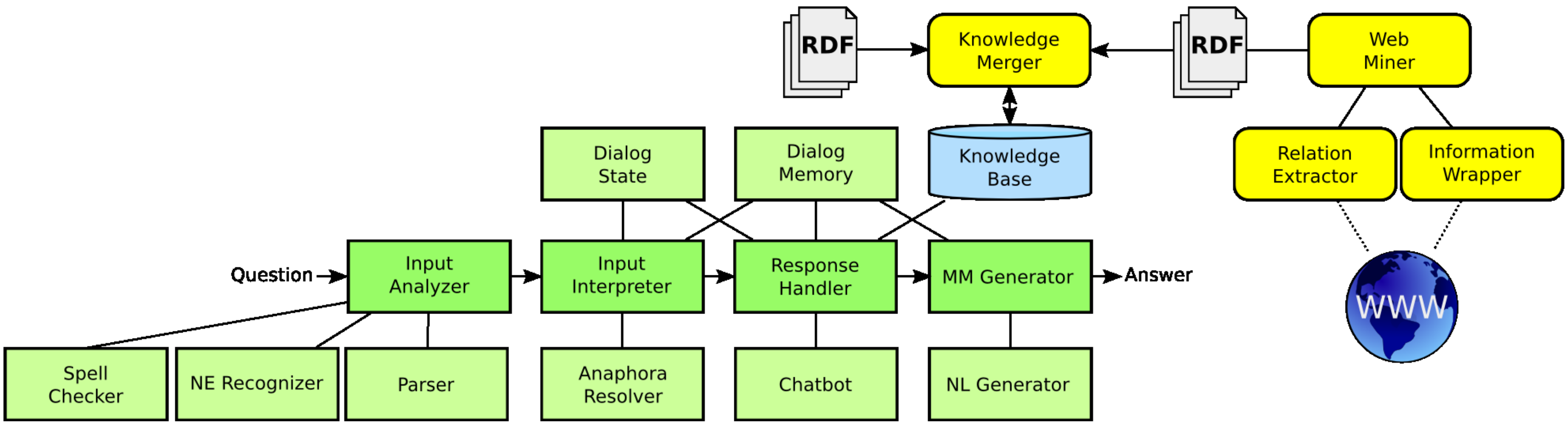


Linked Data Visualization from http://linkeddata.org/

Gossip Galore

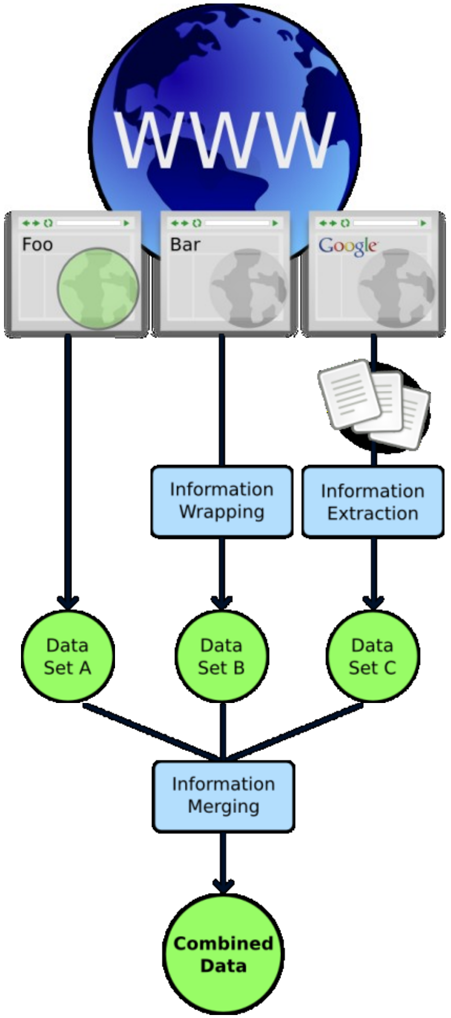
- A user-friendly natural language interface to biographical information
- Embodied Conversational Agent *Gossip Galore*
- Q/A methods employed:
  - Semantic Knowledge Encoding and Retrieval
  - Natural Language Query Analysis
  - Multimodal Answer Generation
  - Finite-State Dialogue Models

- Two major parts:
  - Knowledge Management Components (yellow)
  - Dialogue-Enabled Question Answering Components (green)
- Interface between the components: Knowledge Base



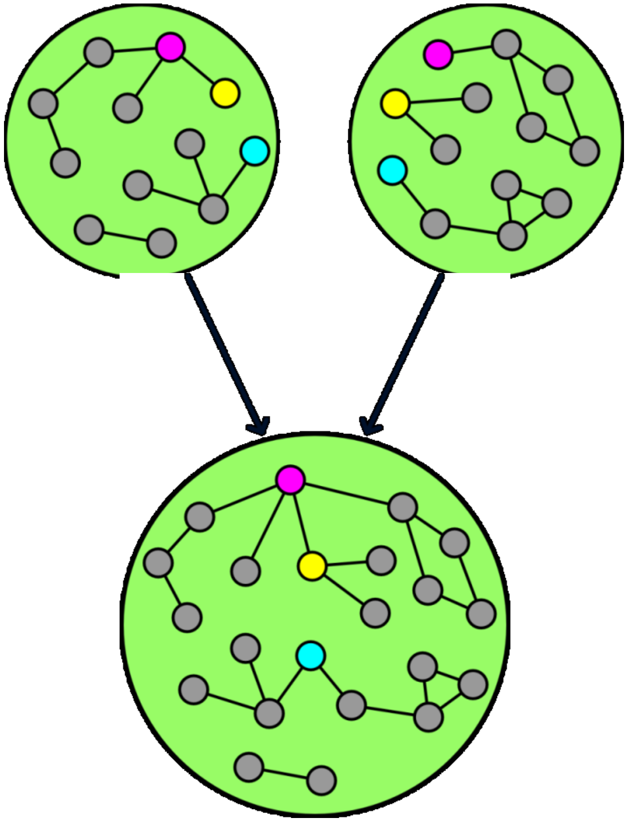
Part 1
Knowledge Acquisition

- Different kinds of knowledge sources
  - Information is offered in structured form (e.g. as SQL or RDF exports)
  - Information provided in semi-structured form on web pages (e.g. price tables for products, info boxes in Wikipedia, etc.)
  - Free natural-language text
- Different approaches for these sources
  - Structured data can be used more or less directly
  - Information Wrapping for accessing semi-structured web pages
  - Information Extraction



Information Merging

- Procedure:
  - Instances with the same referent have to be identified
  - Knowledge bases are then merged by graph union
- Semantic Web:
  - RDF provides a simple framework for such a scenario
  - Ideal for fragmentary data as delivered by Information Extraction
  - Missing data can sometimes be inferred from fragmentary data using domain models

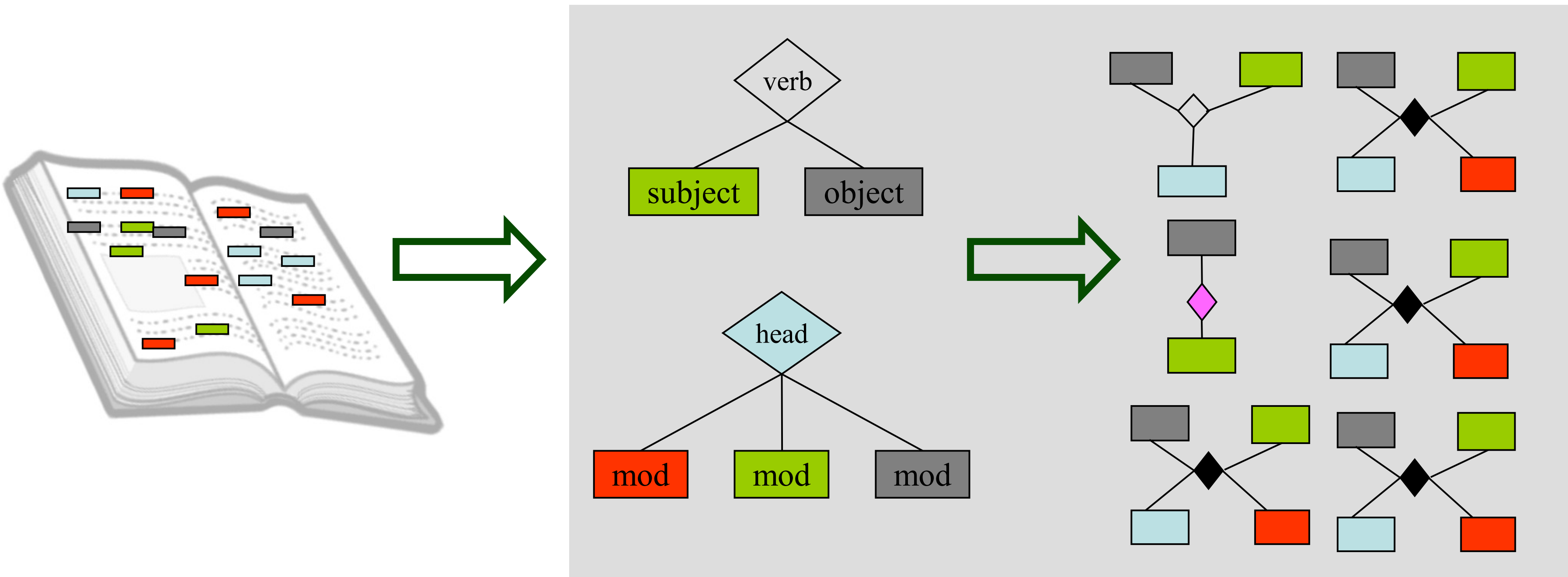


RASCALLI Gossip Knowledge Base

- Knowledge Base (KB) about people in the pop music domain
- Populated using
  - Information Wrapping from semi-structured web sites such as Wikipedia and NNDB
  - Minimally supervised relation extraction with DARE from raw text
- Entities:
  - 38,758 people including 16,532 artists
  - 1,407 music groups
- Relations:
  - 14,909 parent-child
  - 16,886 partner
  - 4,214 sibling
  - 308 influence/influenced
  - 9,657 group membership

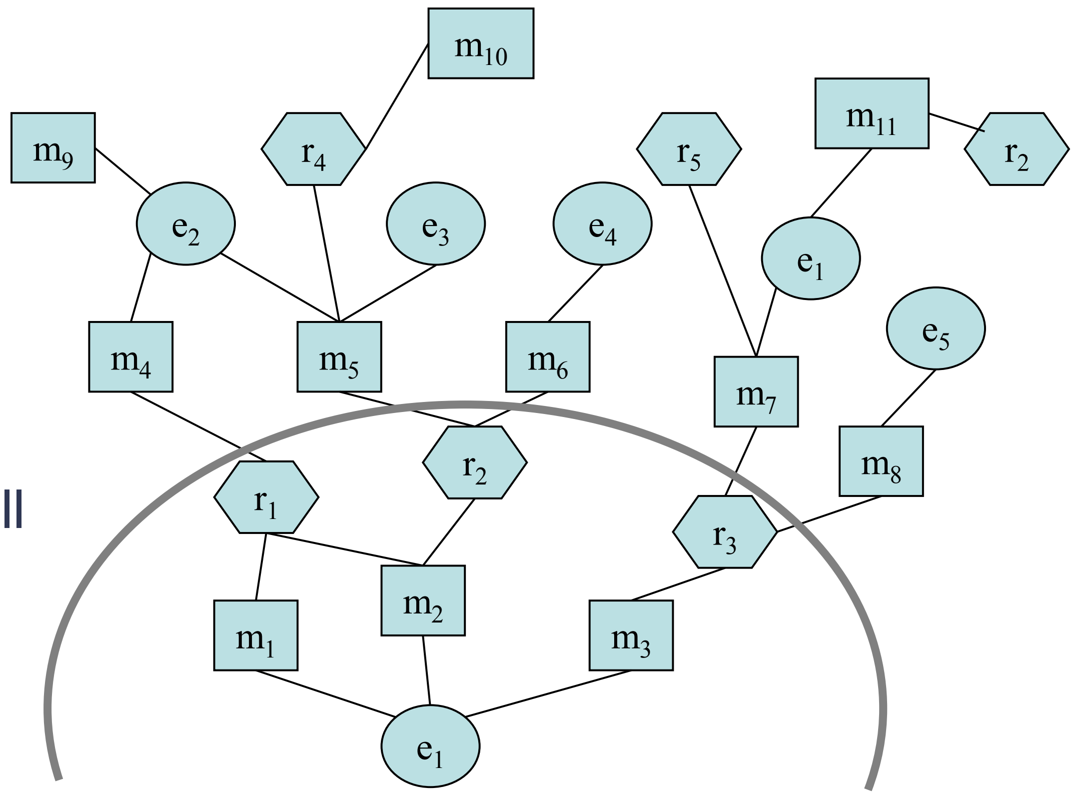
Relation Extraction with DARE

- **D**omain **A**daptive **R**elation **E**xtraction Based on Seeds
- General framework for automatically learning mappings between linguistic analyses and target semantic relations with minimal human intervention (Xu et al, 2008; Xu, 2007)

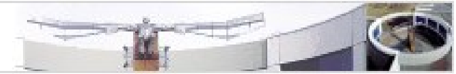


Relation Extraction with DARE

- Relation instances, mentionings, rules
- Rule learning with bootstrapping (sketch):
  - Use confirmed relation instances as seed data
  - Find mentionings of the seed in the text
  - Bottom-up extraction of all patterns for the *i*-ary projections of the target relation ($1 \leq i < n$)
  - Extract further relation instances with the new rules and use these as seeds in the next iteration

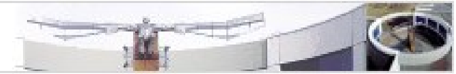


Merging with YAGO



- YAGO is a huge semantic knowledge base, being developed by the group of Gerhard Weikum at Max-Planck-Institute Saarbrücken
- Automatically constructed from the semi-structured parts of Wikipedia (infoboxes) and the taxonomic structure of WordNet
- Made available in RDF format (among others)
- Currently YAGO knows
  - more than 2 million entities (like persons, organizations, cities, etc.).
  - 20 million relations
- We mainly use facts about persons, such as
  - full name, given name,
  - bornIn, bornOnDate, diedIn, diedOnDate
  - actedIn, created, directed, discovered, graduatedFrom, interestedIn, isCitizenOf, participatedIn, produced, worksAt, wrote

Merging with YAGO: Identity Resolution



- Merging rules operating on name and full name from Rascalli, full name and given name from YAGO (<*Rascalli Name, Rascalli Full Name, Yago Full Name, Yago Given Name*>)
  - Rascalli Name == Yago Full Name e.g. <"Clarence Brown"; "Clarence Leon Brown"; "Clarence Brown"; "Clarence">
  - Rascalli Full Name == Yago Full Name e.g. <"Lord Haw-Haw"; "William Joyce"; "William Joyce"; "William">
- \+ additional info if necessary, e.g.: Rascalli Name == Yago Given Name && Rascalli Birthday == Yago bornOnDate

- Dealing with fragmentary name information (culture-dependent heuristics)
  - Siblings sharing same surname could have the same parents, e.g.
    - *Julia Roberts hasParent Walter Roberts;*
    - *Eric Roberts hasParent Walter;*
    - *Julia Roberts hasSibling Eric Roberts;*

    ➔ *Walter == Walter Roberts*
  - A couple could have the same children, e.g.
    - *Madonna hasChild Rocco;*
    - *Guy Richie hasChild Rocco Richie;*
    - *Madonna hasHusband Guy Richie;*

    ➔ *Rocco == Rocco Richie*

Merged Knowledge Base

People: 618,445
Published: 50,601
Movies: 34,458
Locations: 20,733

bornIn = 44339
bornOnDate = 442319
diedIn = 15886
diedOnDate = 205808
originatedFrom = 11693
livesIn = 14707
hasGender = 30815
actedIn = 14088
created = 22473
directed = 5859
discovered = 75
graduatedFrom = 4968
hasNationality = 8256

hasWebsite = 118211
interestedIn = 1806
isCitizenOf = 4865
madeCoverFor = 257
participatedIn = 1158
produced = 9706
worksAt = 1401
wrote = 4152
causeOfDeath = 1888
hasPartyAffliation = 268
hasProfession = 8596
hasReligion = 1533
hasSexualOrientation = 8560
hasRemain = 803

hasMember = 1407
isMemberOf = 8924

hasWonPrize = 16967
hasAlbum = 2663

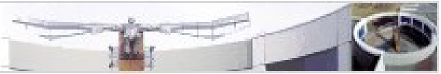
influences = 3043
academicAdvisor = 1307

hasChild = 6868
- hasSon = 4067
- hasDaughter = 2775

hasParent = 12594
- hasMother = 3383
- hasFather = 4219

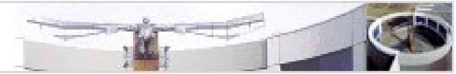
hasSibling = 2076
- hasBrother = 2076
- hasSister = 1100

hasPartner = 18793
- hasSpouse = 16323
  - hasHusband = 7034
  - hasWife = 6458
- hasBoyFriend = 1962
- hasGirlFriend = 2076



Part 2
Dialog Processing

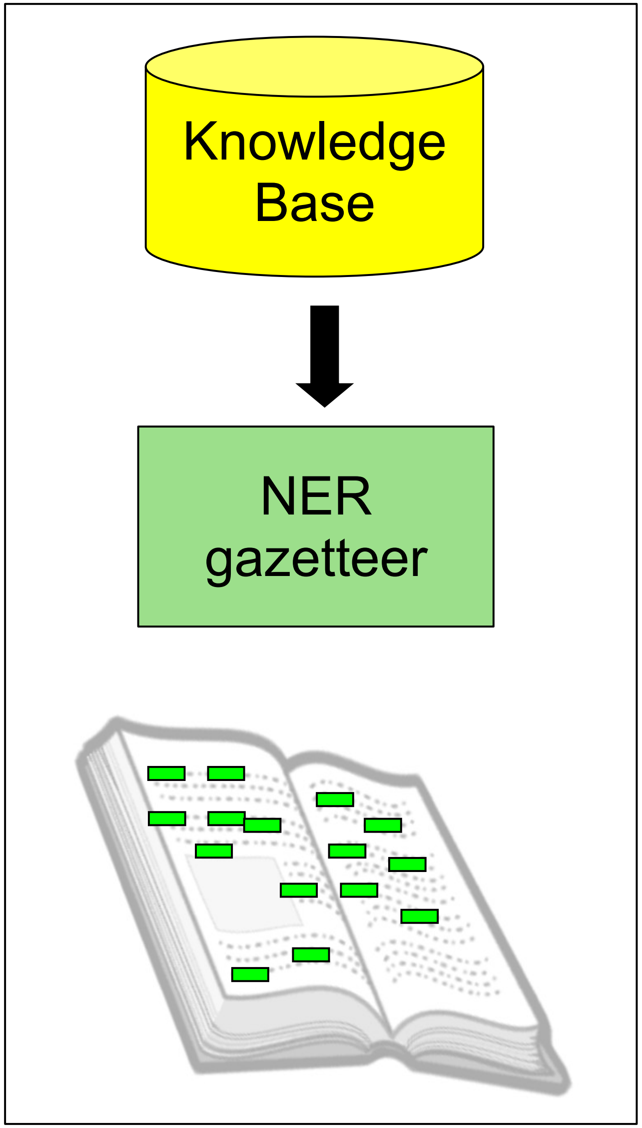
Input Analysis



- Q/A on RDF data is the task of mapping linguistic predicates and arguments to underspecified query graphs
- We support *wh-*, *yes/no*, *how many*-questions involving exactly one query triple
- Approach: linguistic input analysis component, which...
  - Gets the user input
  - Processes the dependency structure belonging to the input
  - Delivers a semantic representation belonging to the dependency structure
  - Assures robustness via an additional string pattern based component

Concept Identification

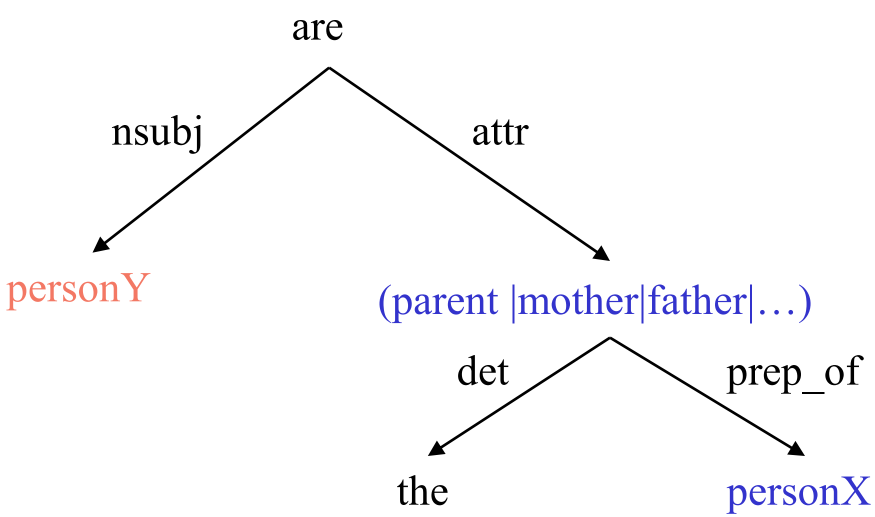
- NER as a bridge from surface strings to semantic concepts
- Gazetteers are derived from the Knowledge Base, associating names and words with ontology instance identifiers
  Examples:
  - “Richard Gere” → g:Person.8134
  - “Deep Purple” → g:Group.1358
  - “buddhist” → g:Religion.3367



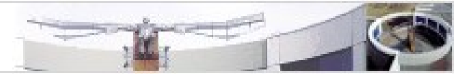
Robust Input Processing

- Hybrid approach to robust input processing
- Cascaded input processors, currently:
  - Dependency parsing
  - Fuzzy string matching baseline
- Using dependency patterns for input analysis, the 1067 paraphrases for the string matching baseline could be reduced to *212* dependency tree patterns

- E.g. „Who are the parents of Mick Jagger?“



Question Semantics



- Dependency parsing and fuzzy string matching deliver semantic representation in triple structure + question type:

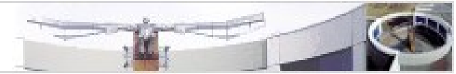
  [[RELATION] [ARG1] [ARG2]] [QTYPE]

- Possible question types, e.g.,
  - [RELATION [ARG1] [null]] [wh]
    Who is the boyfriend of Madonna?
  - [RELATION [ARG1] [null]] [yesno]
    Does Madonna have any boyfriends?
  - [RELATION [ARG1] [null]] [howmany]
    How many boyfriends does Madonna have?
  - [RELATION [ARG1] [ARG2]] [yesno]
    Is Madonna the girlfriend of Mick Jagger?
- Semantics offer more flexibility and abstraction from input and output

Answer Retrieval

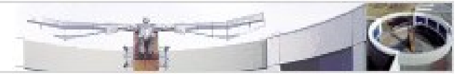
- Question semantics is mapped to query language
- We store all data in an OWLIM knowledge base, using SPARQL queries for access.
- Mapping from semantics to SPARQL is straight-forward: only 8 patterns are needed for simple factoid questions.
- Can be extended to questions with modified NPs, double questions, etc.
- Example: "Who is the boyfriend of Madonna?"
  - Semantics:
    [g:hasBoyfriend [g:Person.14193] [null]] [wh]
  - SPARQL:
    SELECT $x { g:Person.14193 g:hasBoyfriend $x}
  - Returned Answer Set:
    { g:Person.119944, g:Person.494993, …}
- A question as "Does Madonna have any boyfriends?" only differs in answer realization due to the different question type (different expected answer)

Multimodal Generation



- Set of answer triples is realized in natural language, depending on aspects of the question interpretation, answer size and general principles of cooperation
- Dimensions:
  - Question semantic type:
  - Answer size
  - Principles of cooperation:
    - overanswering questions
    - providing alternative solutions to answer the query
  - Expected answer type:
    - Person (“Who”)
    - Place (“Where”)
    - Time (“When”)
    - Quantity (“How many”)
    - Truth value (yes/no)

Natural-language Generation

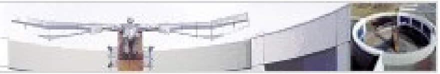


Predicate	EAT	Size	Response
g:hasBoyfriend	Person	≥ 1	Output KB answer (list people)
g:hasBoyfriend	Quantity	≥ 1	„$X has $ANSWER-SIZE boyfriends.“
g:hasBoyfriend	Truth Value	≥ 1	„Yes“ + support answer with some examples
g:hasBoyfriend	*	= 0	„I don‘t know of any boyfriends of $X.“
g:hasDeathday	Time	= 1	„$X died on $ANSWER.“
g:hasDeathday	Time	= 0	„According to my source, $X is still alive.“ + open Google search page
g:hasDeathday	Time	> 1	„My sources are not clear. $X is reported to have died on $ANSWER-CONJUNCTION.
*	*	= 0	„Sorry, I don‘t have that information.“

- Present supportive visual answers for specific answer types
  - Geographical maps for answers of type location
  - IMDB page for some movies
- Provide answer mainly visually where a verbal answer would be too long or too tiring
  - Example: “How are Richard Gere and Michael Jackson connected?”

Conclusions

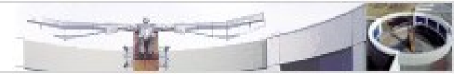
- We presented a system that
  - Enriches Semantic Web data with information extracted from natural language text, and
  - Allows to access that data in natural language (both for user questions and system answers)
  - Demonstrates how existing and freshly acquired Semantic Web data can be exploited to widen the notorious bottleneck of knowledge-driven AI applications.
- Further plans:
  - Integrate other available Semantic Web resources to extend the covered knowledge of our agent.
  - Especially focus on information available from Social Media.



THANK YOU FOR YOUR ATTENTION

Questions?

Acknowledgements



- RASCALLI project funded by the Sixth Framework Programme of the European Commission (IST-27596-2004)
- KomParse, ProFIT programme of the Federal State of Berlin and the EFRE programme of the European Union
- TAKE project, funded by the German Ministry for Education and Research (01IW08003)