Advanced Topics and Applications of IE

Günter Neumann & Feiyu Xu

{neumann, feiyu}@dfki.de

Language Technology-Lab
DFKI, Saarbrücken

Outline

- An Information Extraction-based Tourism Information System
- Semantics and Information Extraction

Facts Sheet - MIETTA

- Title: MIETTA -Multilingual Information Extraction for Tourism and Travel Assistance
- Funding: EU Language Engineering Sector of TAP (HLT-IST)
- Technical Partners: DFKI, Celi, University of Helsinki, Polito, Unidata
- User Partners: Commune DI Rome, City of Turku, Staatskanzlei of the Saarland

Objectives

- Multilingual internet portal and specialised information system for tourist information

  Five languages: English, Finnish, French, German, Italian

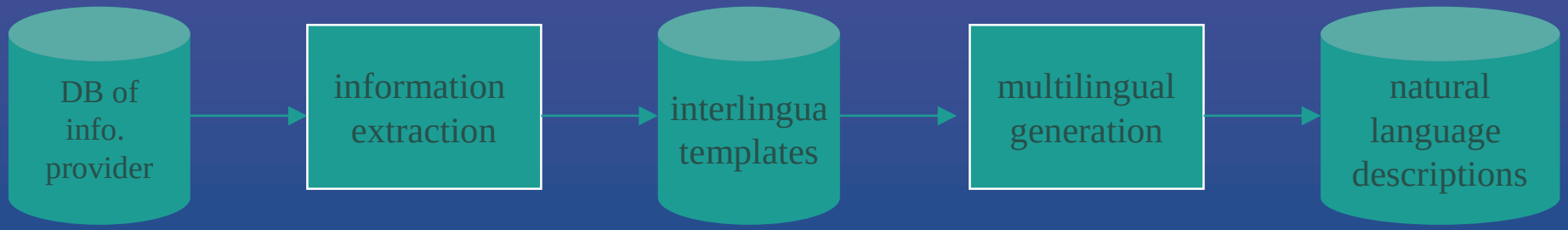
  Three regions: Rome, Saarland and Turku

- Integrated access to heterogeneous data sources and make it fully transparent to end users whether they are searching in
  - ✶ WWW documents or
  - ✶ Databases

Information Extraction and Multilingual Generation

- Motivation

  - ✶ Make the database content more structured and multilingual accessible.
  - ✶ Apply the same free text retrieval method to the generated descriptions as to the web documents



Information Extraction in MIETTA

- The objective of information extraction is twofold:
  - To extract the domain relevant information (templates) from the unstructured data so that the user can access more facts and more accurately
  - To normalise the extracted data in a language independent format to facilitate multilingual generation

- Three steps for template extraction in MIETTA
  - Natural language shallow processing: named entities, np, vp
  - Normalisation: converting information into a language independent format
  - Template filling: mapping the extracted information into template slots by employing specific template filler rules

Example of IE

German text from an event calendar in Saarland

St. Ingbert: -Sanfte Gymnastik für Seniorinnen und Senioren montags von 10 bis 11 Uhr im Clubraum, Kirchengasse 11.

English: St. Ingbert: -Gentle Gymnastic for seniors, every Monday from 10:00 to 11:00 am, in Club room, Kirchengasse 11

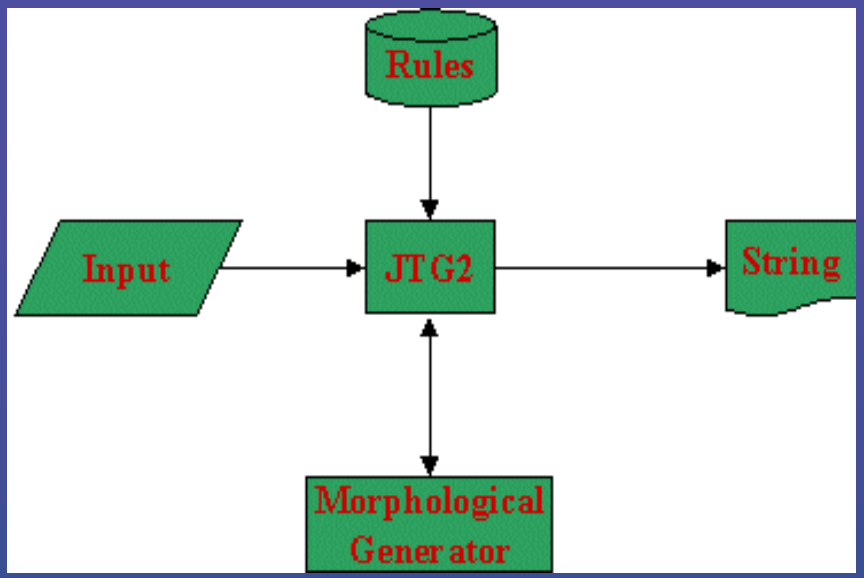
Event:

Name: gymnastic
Addressee: seniors
time:
start time:10
end time: 11
weekly: yes
weekday: 1

location:
city name: St. Ingbert
address: Club room
Kirchengasse 11

Multilingual Generation

- Template Generation system (JTG/2)



- Language independent input allows for easy extension of the generation component to other languages

Example

Level1: Event
Level2: Theater
Level3:
Event-Name: Faust
StartDate: 21.10.99
PlaceName: Staatstheater
Address: Schillerplatz, 66111 Saarbrücken
Phone: 0681-32204

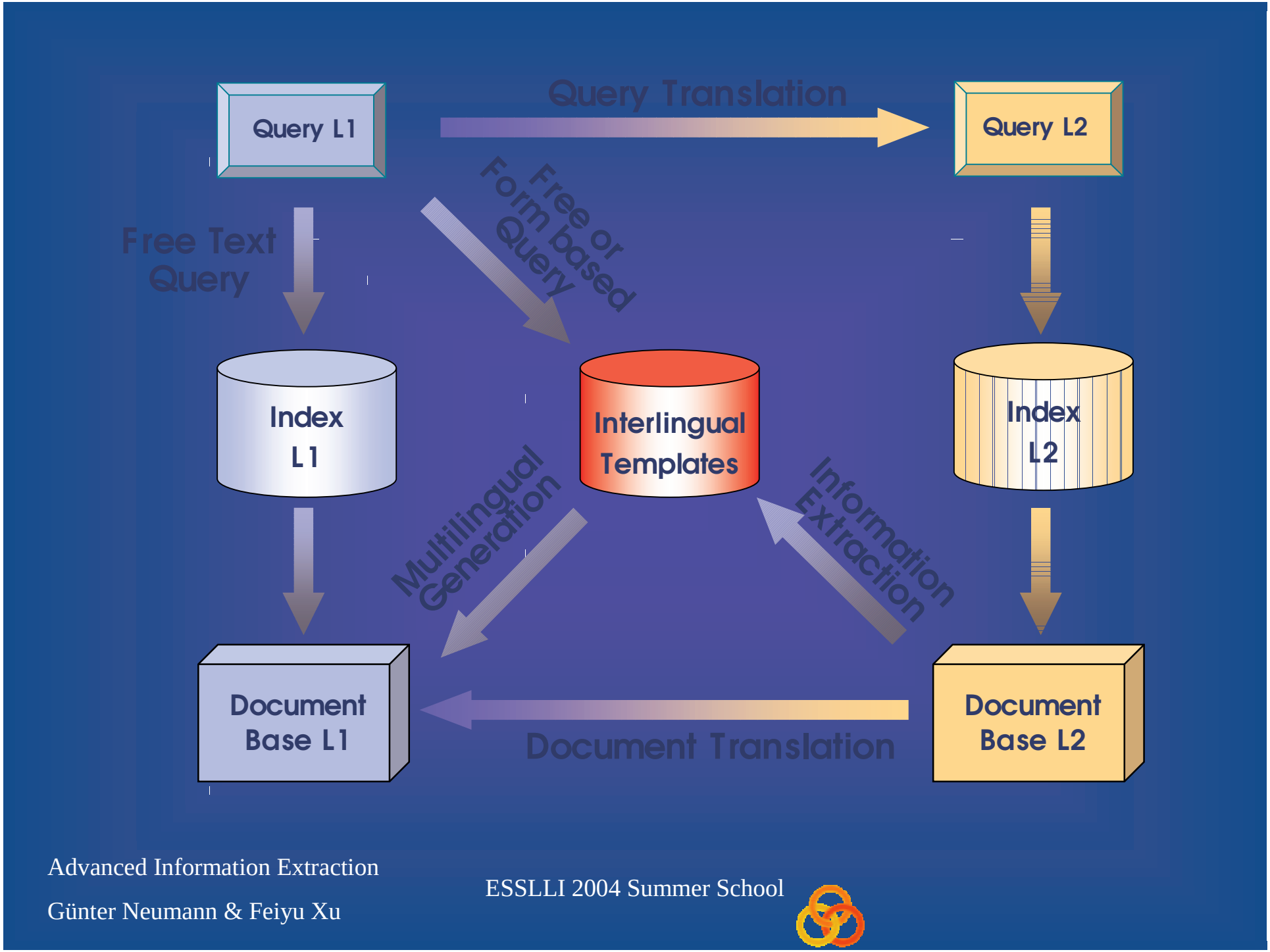
English:

The theater show Faust will take place at the Staatstheater in Schillerplatz 1, 66111 Saarbrücken (in the downtown area).

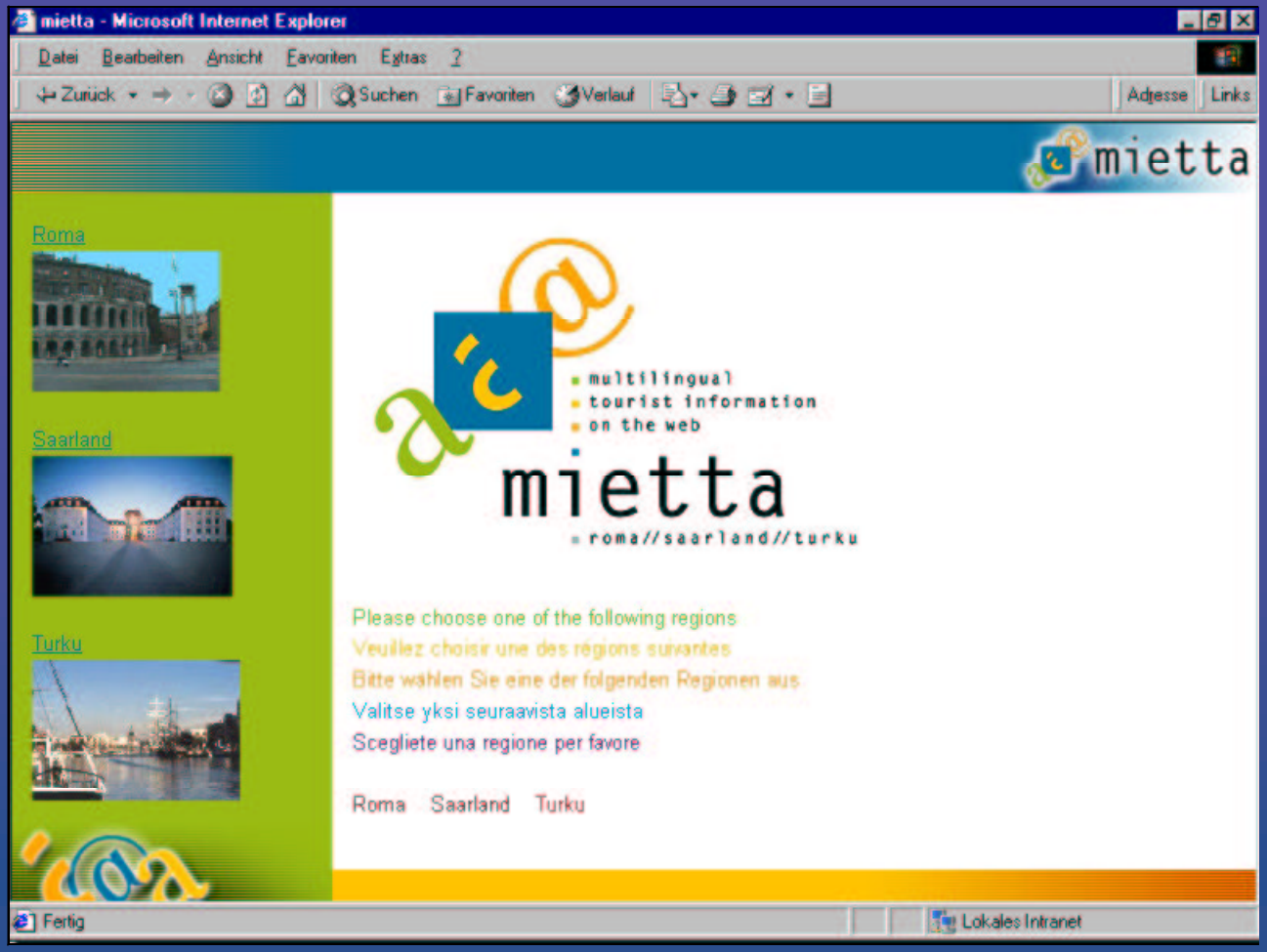
The scheduled date is Thursday, October 21, 1999. Phone: 06 81-32204

Finnish:

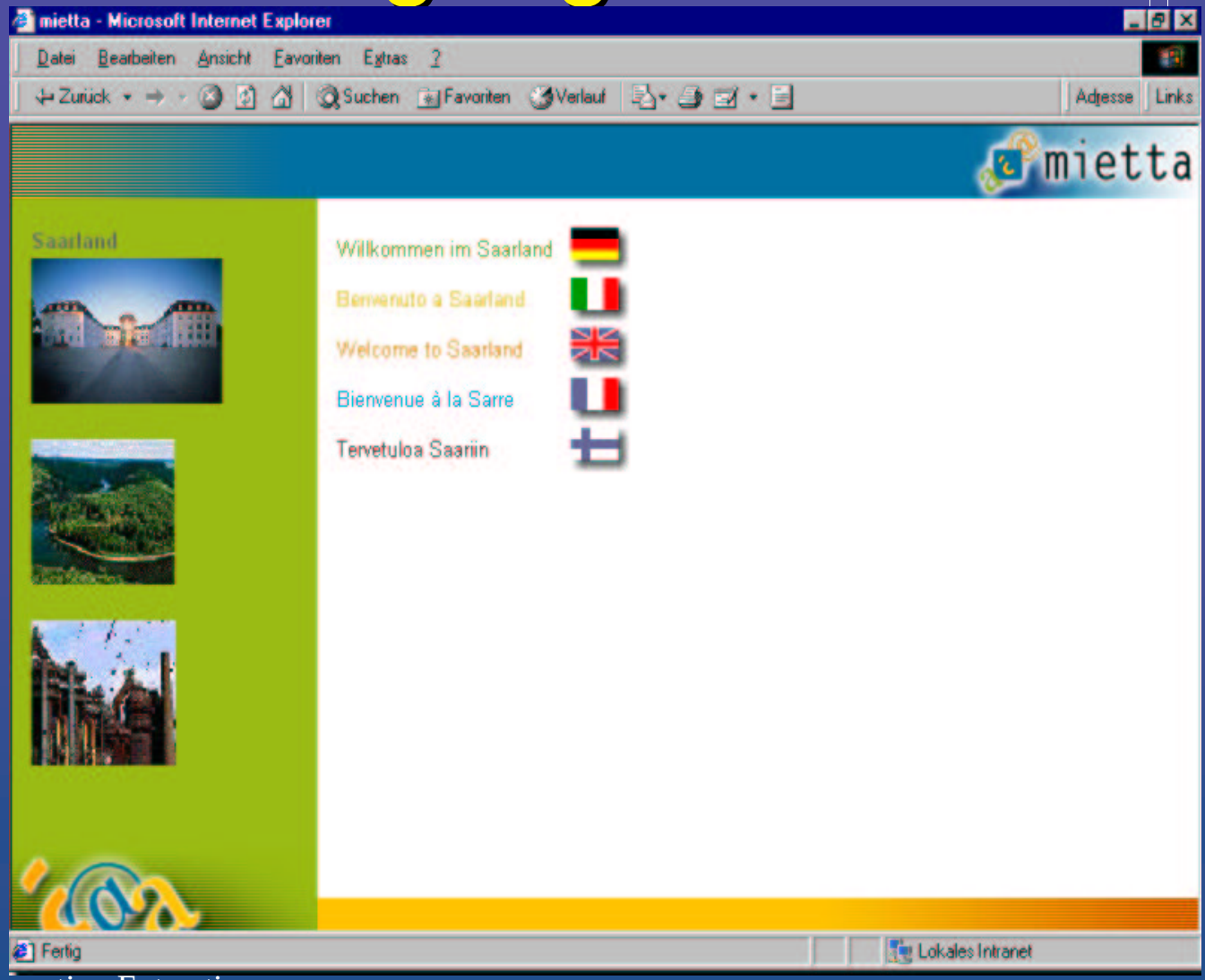
Teatteriesitys Faust järjestetään Staatstheaterissa, osoitteessa Schillerplatz 1, 66111 Saarbrücken (keskustan alueella). Tapahtuman päivämäärä on 21. lokakuuta 1999. Puhelin: 06 81-32204.


Query L1
Query Translation
Query L2
Free Text Query
Free or Form based Query
Index L1
Interlingual Templates
Index L2
Multilingual Generation
Information Extraction
Document Base L1
Document Translation
Document Base L2

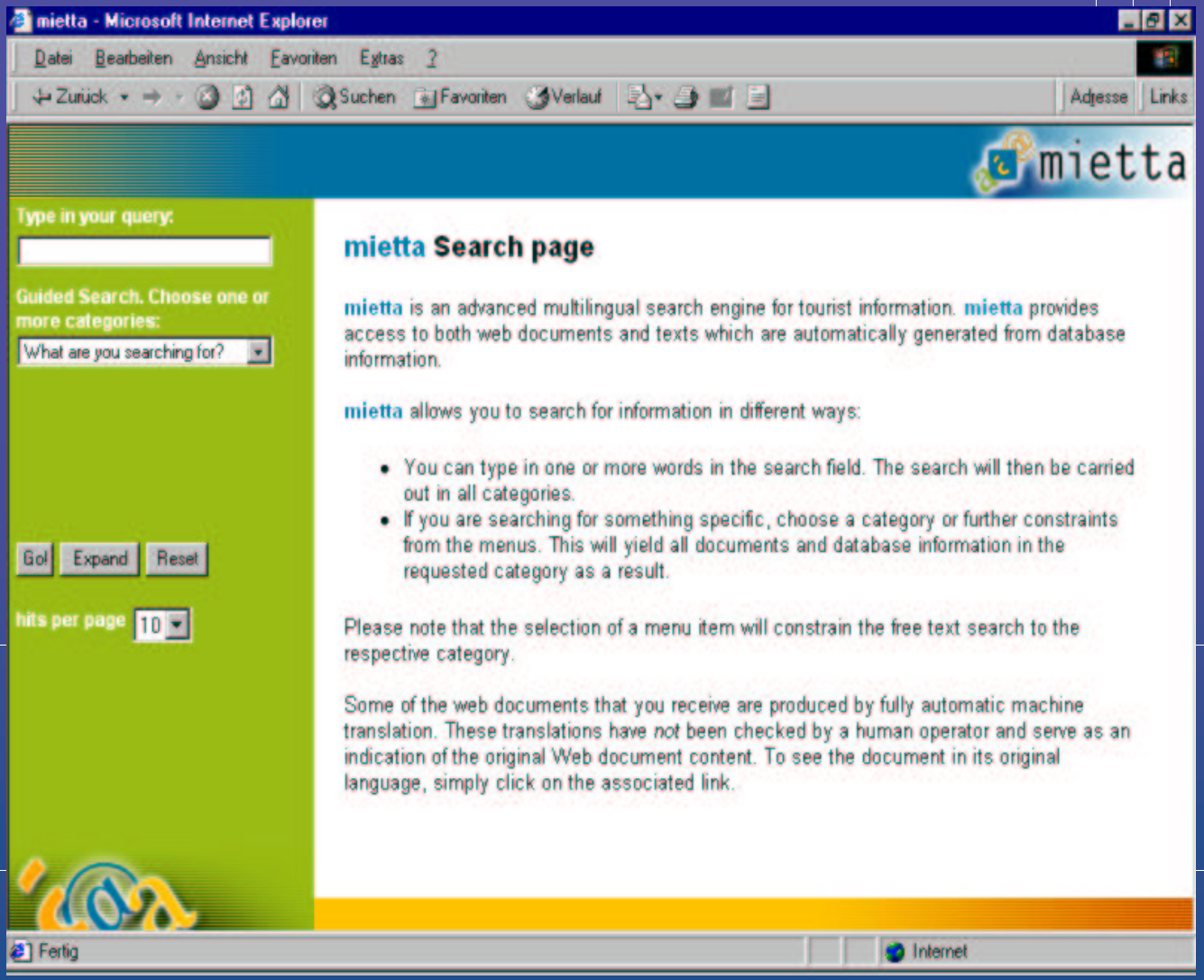
MIETTA Start Page: Choose Region



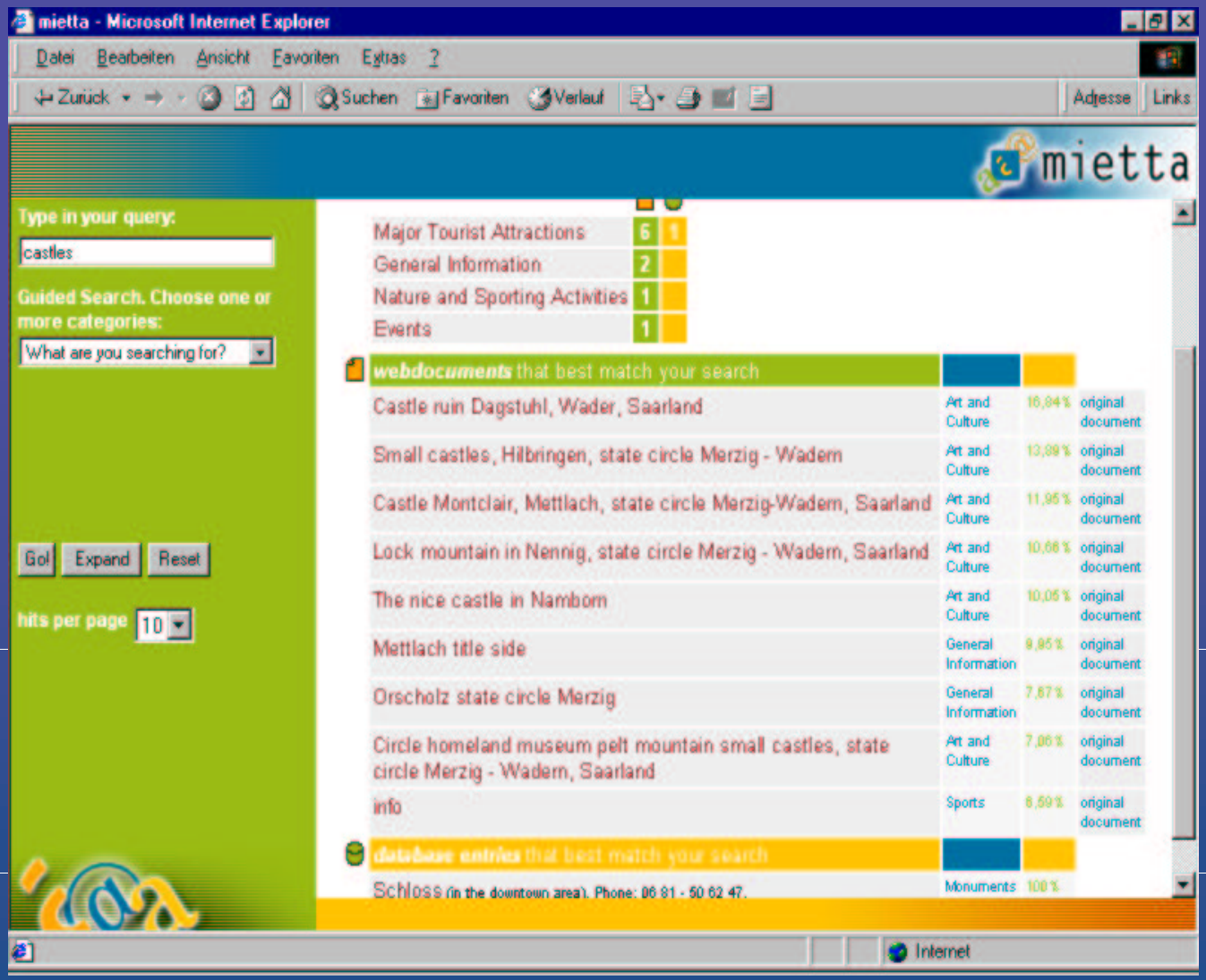
Choose Language



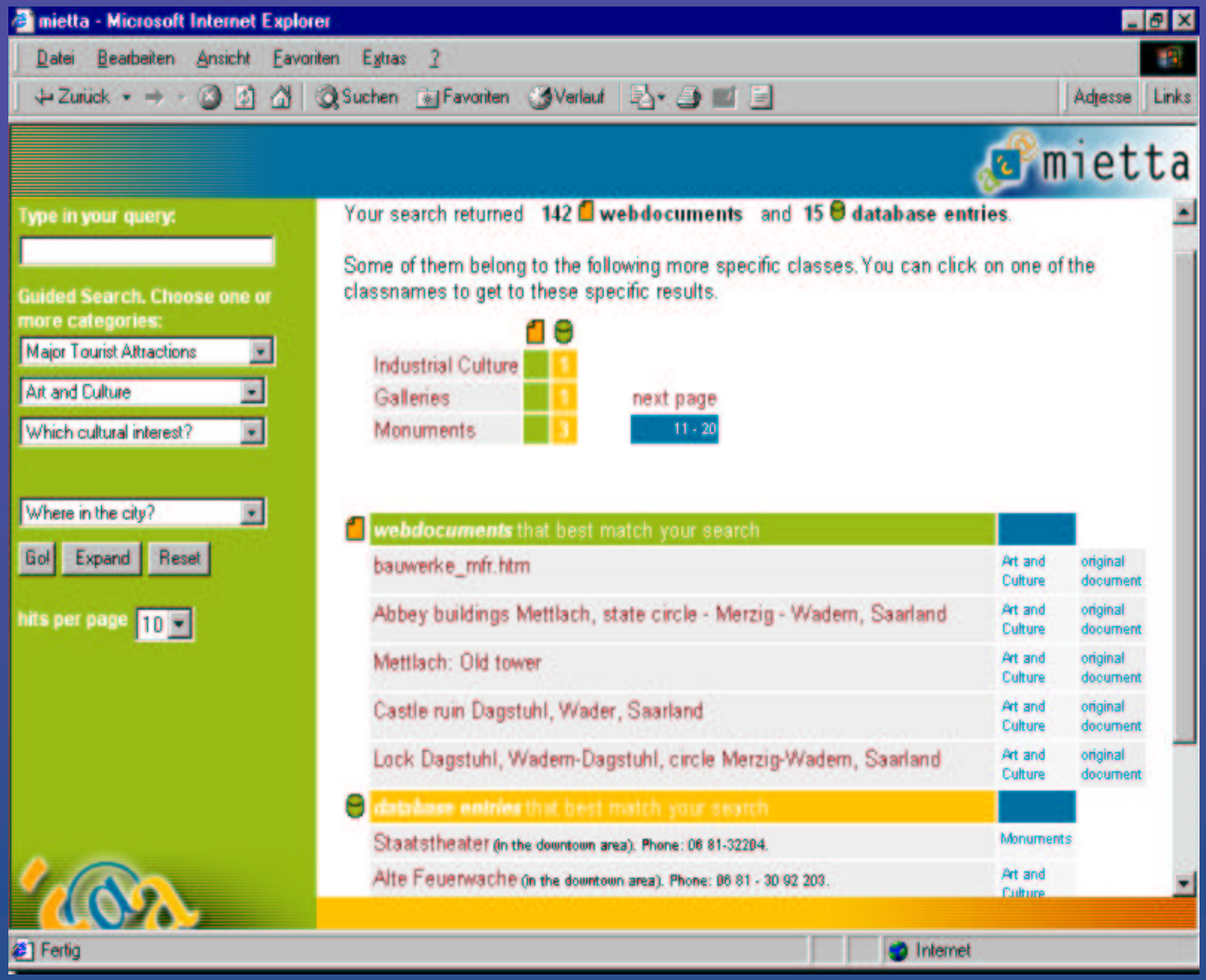
MIETTA Search Menu



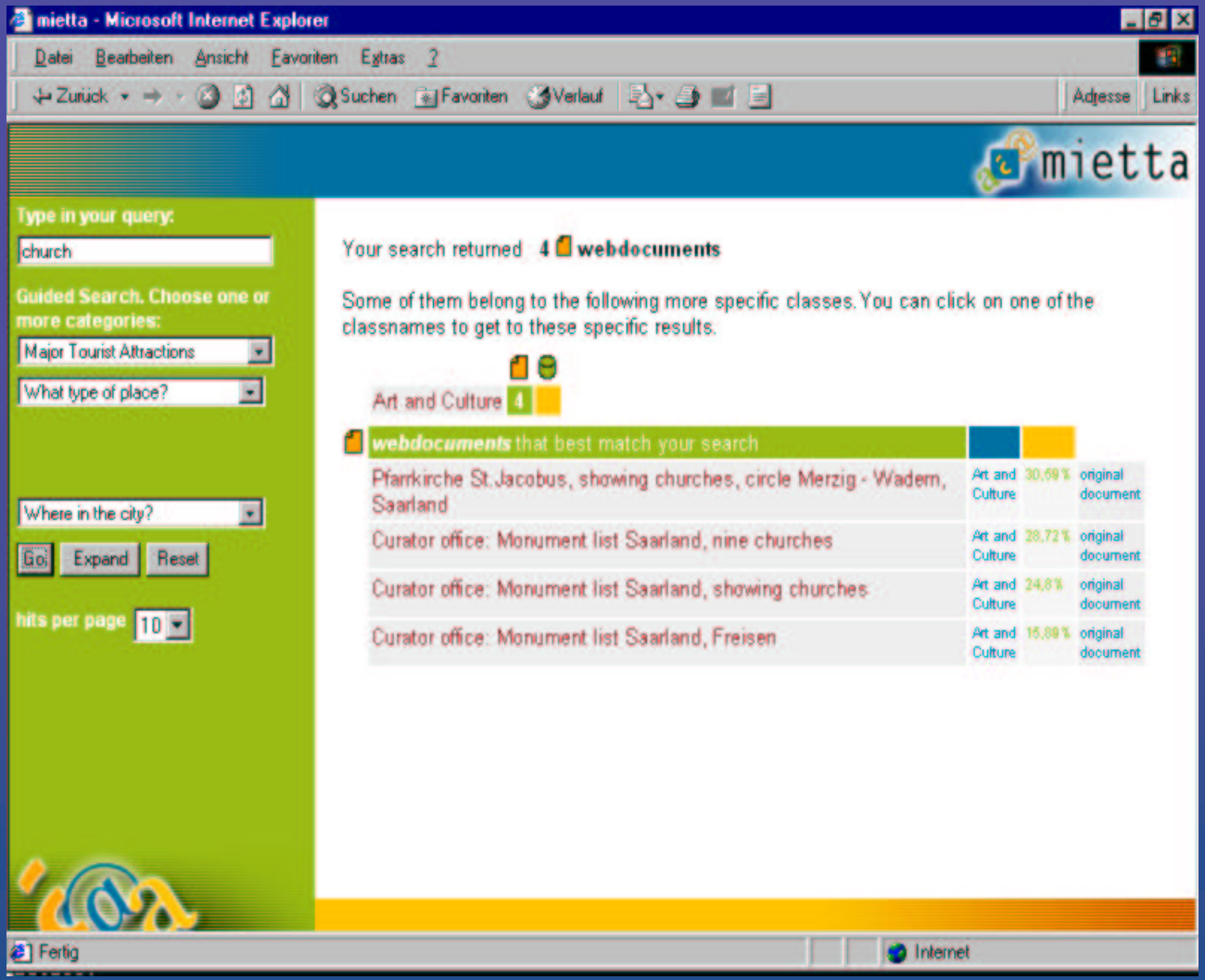
MIETTA Free Text Retrieval



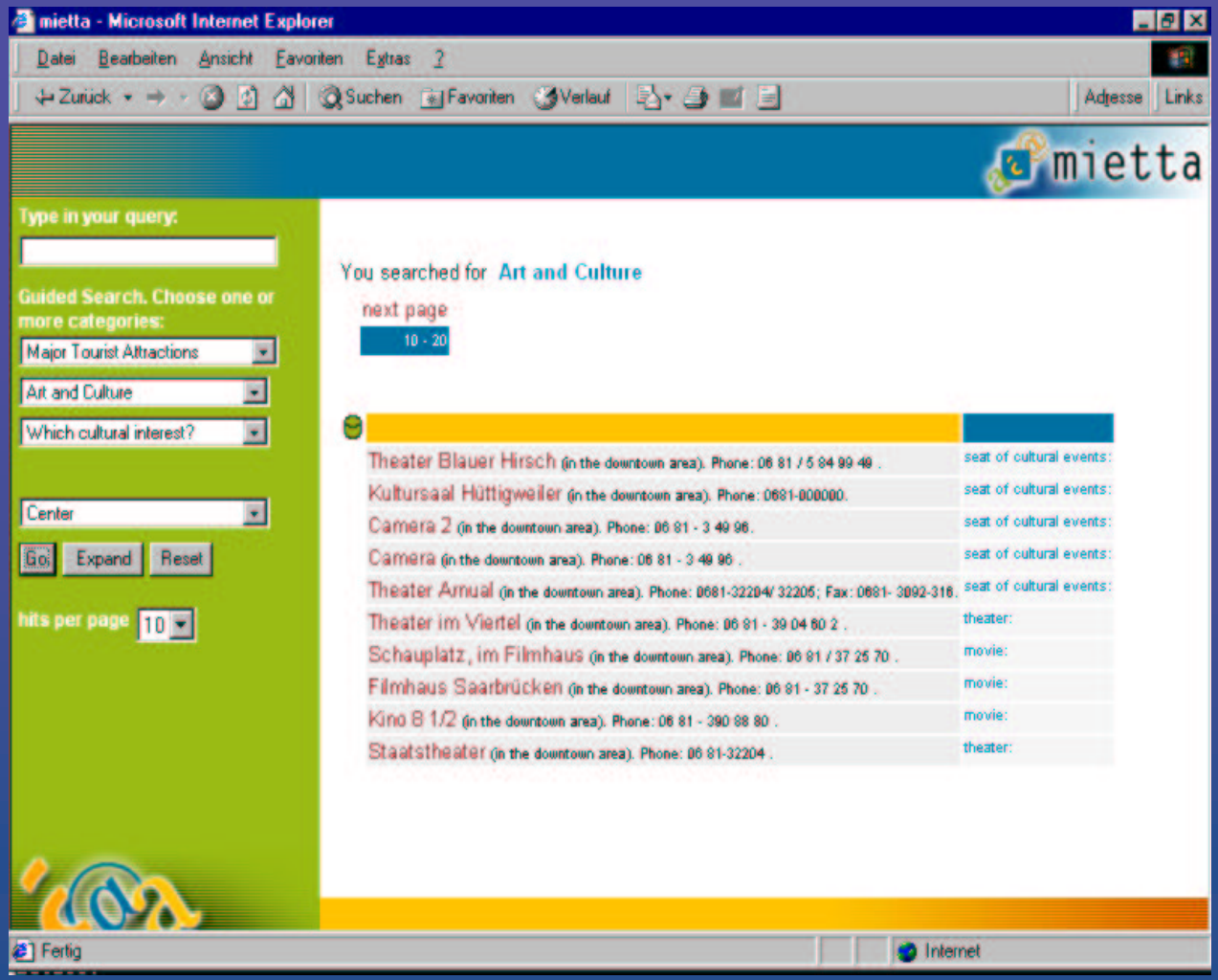
MIETTA Class-based Navigation



MIETTA Class-based Navigation with Free Text



MIETTA Form based Query

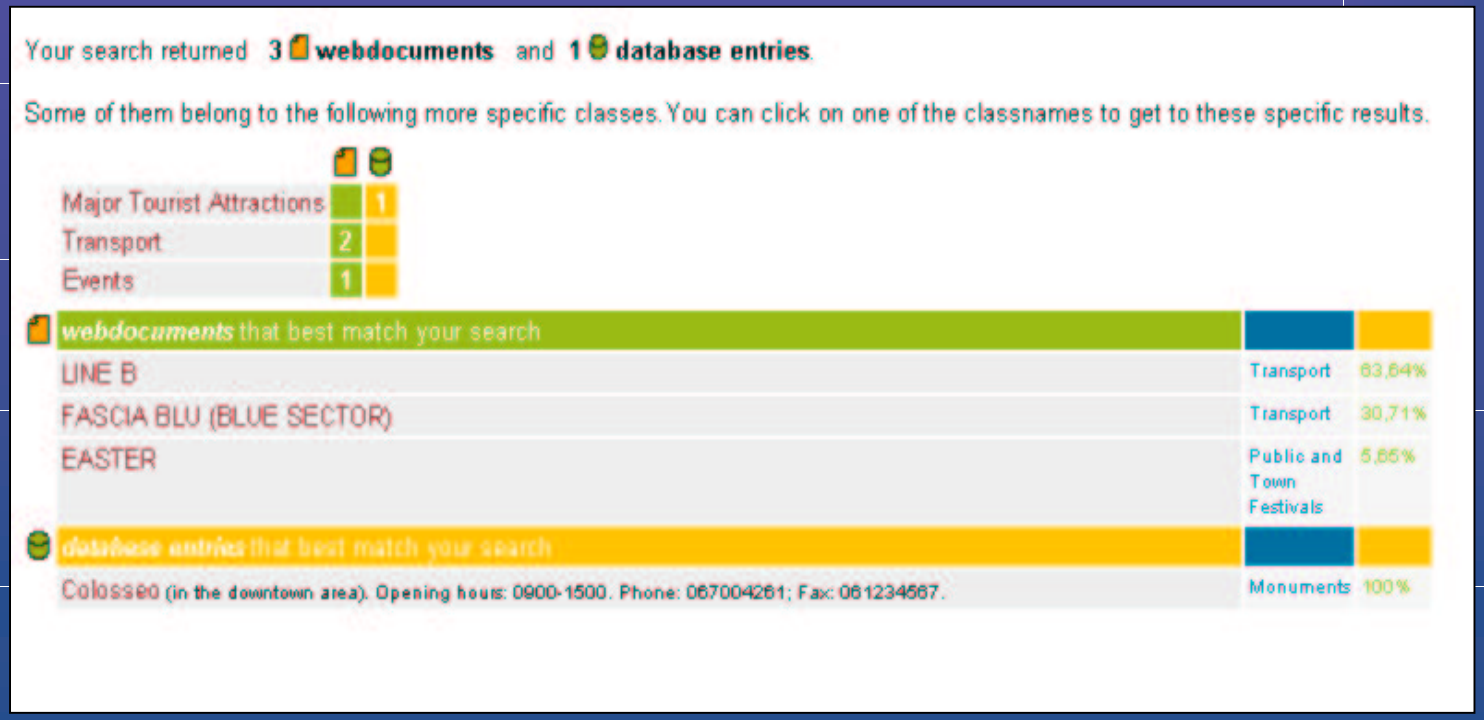


Online Text Generation

English	The theater **Staatstheater** is located in Schillerplatz 1, 66111 Saarbrücken (in the downtown area). Phone: 06 81-32204 .
Finnish	Teatteri **Staatstheater** sijaitsee osoitteessa Schillerplatz 1, 66111 Saarbrücken (keskustan alueella). Puhelin: 06 81-32204.
French	Le théâtre **Staatstheater** se trouve Schillerplatz 1, 66111 Saarbrücken (dans la zone du centre). Téléphone: 06 81-32204 .
German	Das Theater **Staatstheater** befindet sich in der Schillerplatz 1, 66111 Saarbrücken (im Stadtzentrum). Phone: 06 81-32204 .
Italian	Il teatro **Staatstheater** si trova in Schillerplatz 1, 66111 Saarbrücken (nella zona del centro). Telefono: 06 81-32204.

Result Presentation

- Result contains both database entries and documents
- All information is presented in uniform format
  - ✶ Classified
  - ✶ Ordered according to the relevance



What is Semantics?

- the philosophical and scientific study of meaning [Encyclopedia Britannica]
- Semantics is, generally defined, the study of meaning of linguistic expressions.
  [SIL Glossary of Linguistics]
- Semantics is the study that relates signs to things in the world and patterns of signs to corresponding patterns that occur among the things the signs refer to. [Charles Sanders Peirce]
- Theory of the relationship between formal aspects of language and objects and facts in the world. [Appelt, 2003]

IE: Concepts and Relations

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates
Microsoft

* Bill Veghte
Microsoft
VP

* Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
RichardStallman	founder	Free Soft..

IE: A pragmatic approach to Semantic Theory [Appelt, 2003]

- Let application requirements drive semantic analysis
  - Motivation for a semantic theory is a practical one driven by database filling needs
- Pick a limited ontology of core concepts, and build out, motivated by application needs
- Identify the types of entities that are relevant to a particular task
- Identify the range of facts that one is interested in for those entities
- Ignore everything else

The ACE Program

- "Automated Content Extraction"
- Develop core information extraction technology by focusing on extracting specific semantic entities and relations over a very wide range of texts.
- Corpora: Newswire and broadcast transcripts, but broad range of topics and genres.
  - Third person reports
  - Interviews
  - Editorials
  - Topics: foreign relations, significant events, human interest, sports, weather
- Discourage highly domain- and genre-dependent solutions

Components of a Semantic Model

- Entities - Individuals in the world *that are mentioned in a text*
  - Simple entities: singular objects
  - Collective entities: sets of objects of the same type *where the set is explicitly mentioned in the text*
- Relations – Properties that hold of tuples of entities.
- Complex Relations – Relations that hold among entities and relations
- Attributes – one place relations are attributes or individual properties

Components of a Semantic Model

- Temporal points and intervals
- Relations may be timeless or bound to time intervals
- Events – A particular kind of simple or complex relation among entities involving a change in at least one relation

Relations in Time

- timeless attribute: gender(x)
- time-dependent attribute: age(x)
- timeless two-place relation: father(x, y)
- time-dependent two-place relation: boss(x, y)

Relations vs. Features or Roles in AVMs

- Several two place relations between an entity *x* and other entities y_i can be bundled as properties of x.
- In this case, the relations are called roles (or attributes) and any pair
  <relation : y_i> is called a role assignment (or a feature).
- name <x, CR>

name: Condoleezza Rice
office: National Security Advisor
age: 49
gender: female

Relations vs. Features or Roles in AVMs

- any many-place relation can be expressed as a set of two-place relations

appoint (x,y,z) e.g., appoint(Bush, Rice, SecurityAdvisor)
appoint-security-advisor(Bush, Rice)
appoint-rice(Bush, SecurityAdvisor)

- appoint-relation

$$\begin{bmatrix} \text{appointer: Bush} \\ \text{appointee: Rice} \\ \text{office: SecurityAdvisor} \end{bmatrix}$$

Relations vs. Features or Roles in AVMs

- in this way appointer, appointee and office become attributes of the appoint relation
- since IE templates are special cases of AVMs, the mapping between IE templates and our relations is rather straightforward

Semantic Analysis: Relating Language to the Model [Appelt, 2003]

- Linguistic Mention
  - A particular linguistic phrase
  - Denotes a particular entity, relation, or event
    - A noun phrase, name, or possessive pronoun
    - A verb, nominalization, compound nominal, or other linguistic construct relating other linguistic mentions
- Linguistic Entity
  - Equivalence class of mentions with same meaning
    - Coreferring noun phrases
    - Relations and events derived from different mentions, but conveying the same meaning

Relations as Nodes in an Ontology

receiving_award

reason : *achievement* (accomplishment, service, skills, ...)

award : *award_type* (medal, prize, title, ...)

recipient : *person*

time : *time* (interval, date)

location : *place* (place, region,..)

receiving_prize

reason : *achievement* (accomplishment, service, skills, ...)

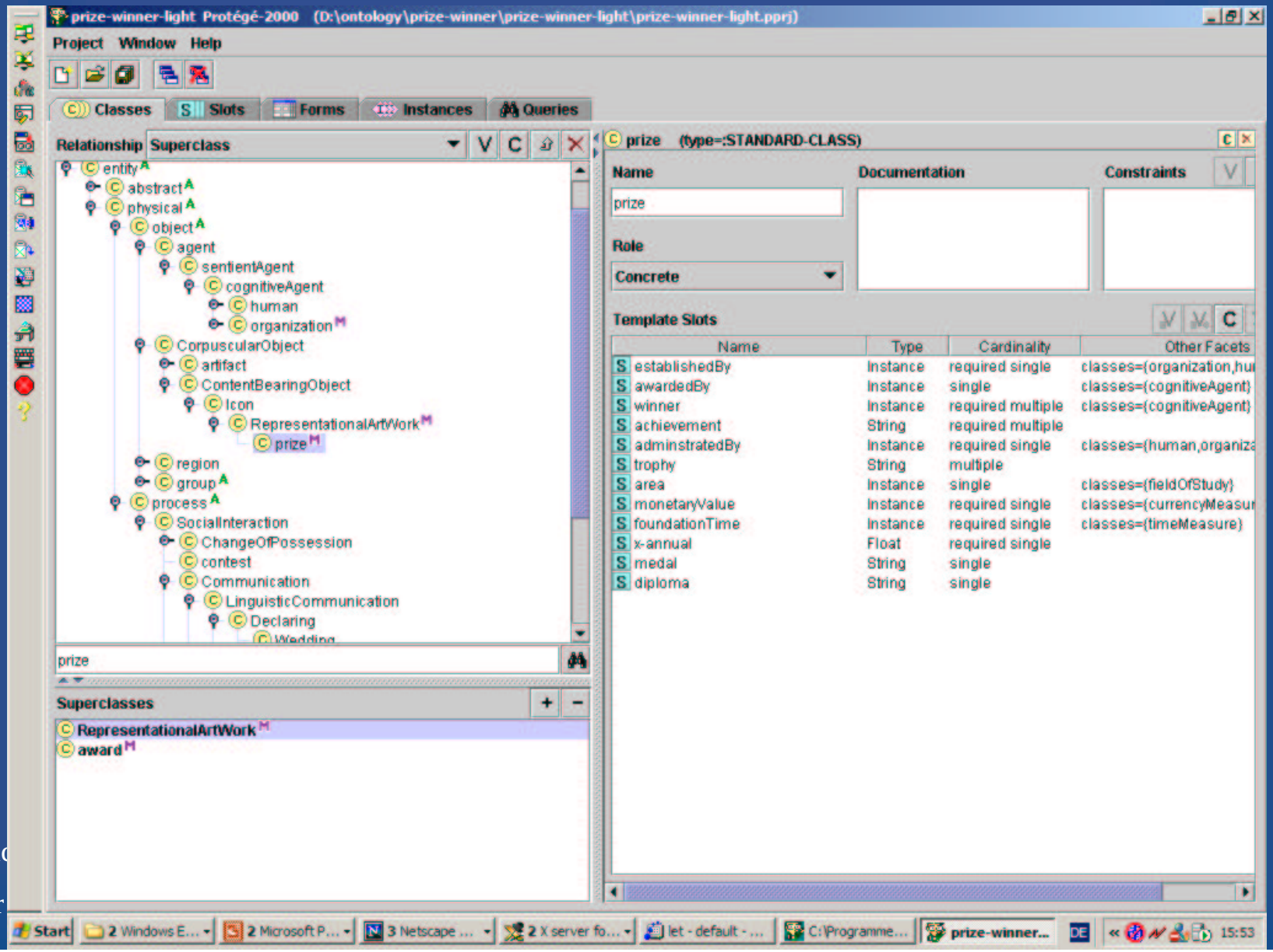
award : *prize*

recipient : *person*

time : *time* (interval, date)

location : *place* (place, region,..)

Modelling Ontology with SUMO, WordNet



Advanc

Günter

From Generic to Domain Specific Relations

receiving_nobel_prize
reason : *achievement* (accomplishment, service, skills, ...)
award : *nobel_prize*
recipient : *person*
time : *time* (interval, date)
location : *place* (place, region,..)

nobel_prize
area : *nobel_prize_area* (medicine, physics, literature, peace, ...)
year: *year*
recipient: *person*
co-recipients: *persons*

Scenario Template View of A Complex Relation

receiving_nobel_prize
reason : *achievement* (accomplishment, service, skills, ...)
award : **nobel_prize**
area : *nobel_prize_area* (medicine, physics, literature, peace, ...)
year: *year*
recipient: *person*
co-recipients: *persons*
recipient : *person*
time : *time* (interval, date)
location : *place* (place, region,..)

Scenario Template to a Flat Relation

receiving_nobel_prize

reason :	*achievement* (accomplishment, service, skills, ...)
award :	nobel_prize
area :	*nobel_prize_area* (medicine, physics, literature, peace, ...)
year:	*year*
recipient:	*person*
co-recipients:	*persons*
location:	*place*

Representation of an Event

receiving_nobel_prize
event : *event*
reason : *achievement* (accomplishment, service, skills, ...)
award : nobel_prize
area : *nobel_prize_area* (medicine, physics, literature, peace, ...)
year: *year*
recipient: *recipient*
co-recipients: *person*
location: *place*

receiving_nobel_prize (event, achievement, "nobel_prize", nobel_prize_area, year, recipient, co-recipients, location)

Neo-Davidsonian View of Events

receiving_nobel_prize (event, achievement, "nobel_prize", nobel_prize_area, year, recipient, co-recipients, location)

Neo-Davidsonian view

receiving_nobel_prize (event, achievement, "nobel_prize", "physics", "1996", recipient, co-recipients, location)

polymorphic relations:

event(e_1)

year(e_1, "1996")

recipient(e_1, x_1)

area(e_1, a_1)

to be explicit:

nobel_prize_event(e_1)

nobel_prize_year(e_1, "1996")

nobel_prize_ recipient(e_1, x_1)

nobel_prize_ area(e_1, a_1)

Questions

λx.receiving_nobel_prize (e_1, achievement, "nobel_prize", "physics", "1996", x, co-recipients, location)

λx.recipient(e_1, x)

Simple Extensional Denotation of "Nobelpreisträger"

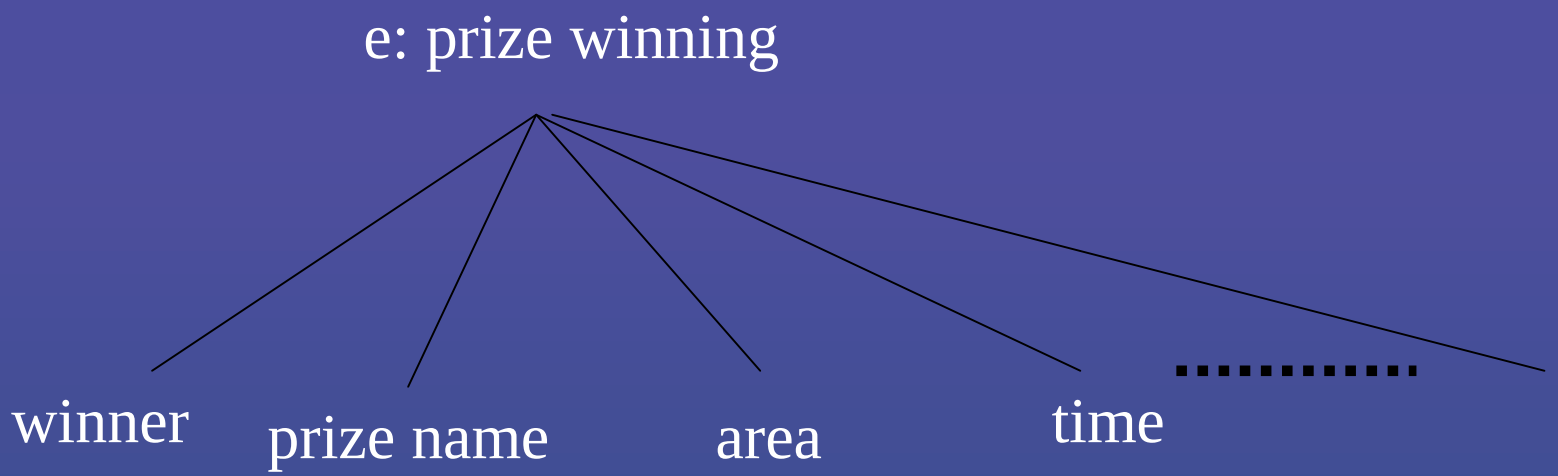
nobel_prize_winner' =

$\lambda x\ [person(x) \wedge \exists e\ (nobel\_prize\_event(e) \wedge\ nobel\_prize\_recipient(e_1, x_1))]$

Pragmatic Approach to Relation Representation

- N-ary to binary/elementary relations
  - Neo-Davidsonian view
- Nested relations to a flat list of elementary relations
  - Collapsing
  - Meta structure for representation of nested relations

Neo-Davidsonian View of Relations



prizewinning(e), winner(e,x), prizename(e,y), area(e,z), time(e,w)

Semantic Labelling for IE

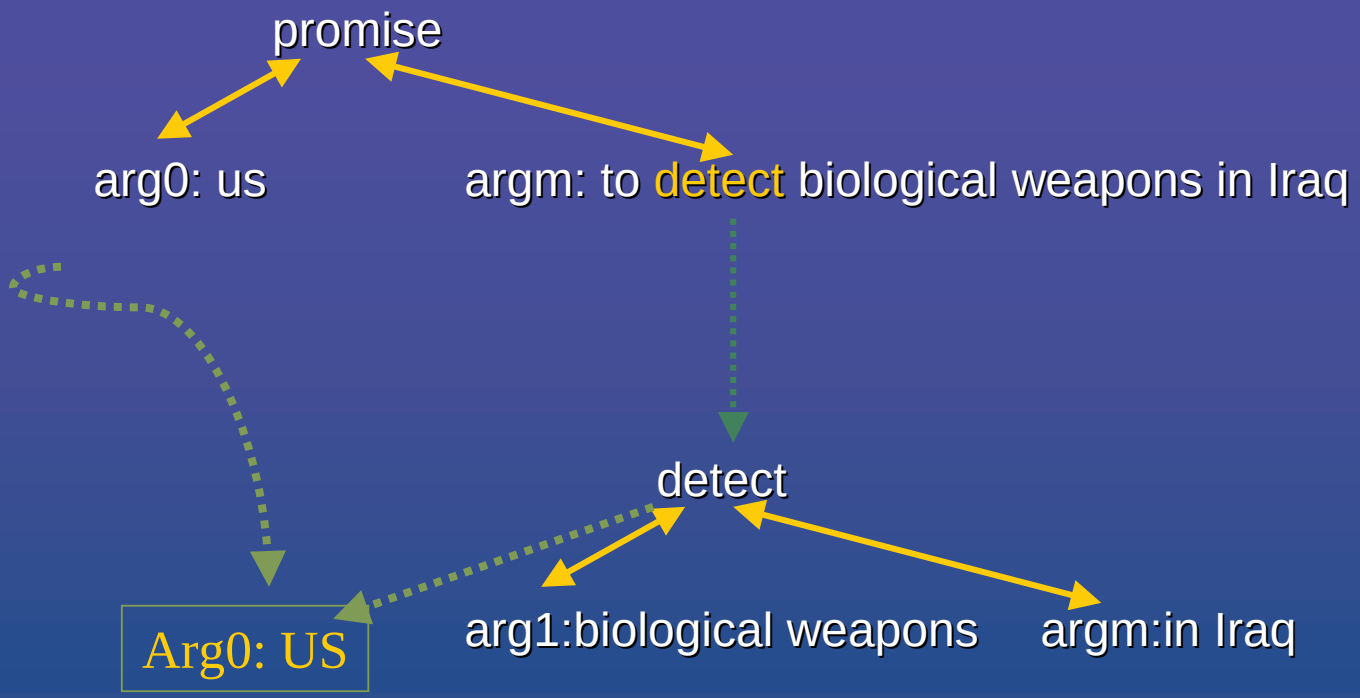
- Automatic recognition and classification of predicate argument structures
- A new IE paradigam [Surdeanu et al., 2003]
  - Mapping predicate argument structures to domain specific relations
- Introduction to Semantic Labelling
  - CONLL 2004 (NAACL 2004)

Flat Predicate Argument Structures

Does US promise to detect biological weapons in Iraq?

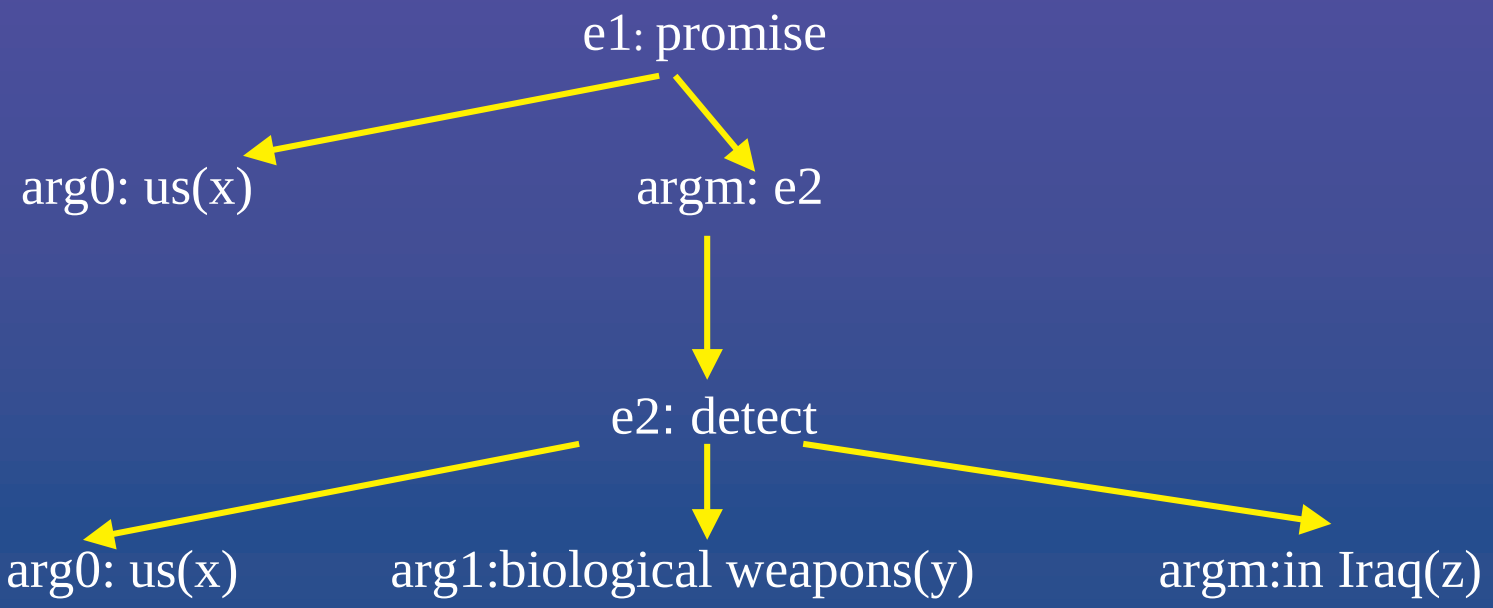
Flat Predicate Argument Structures

Does US promise to detect biological weapons in Iraq?



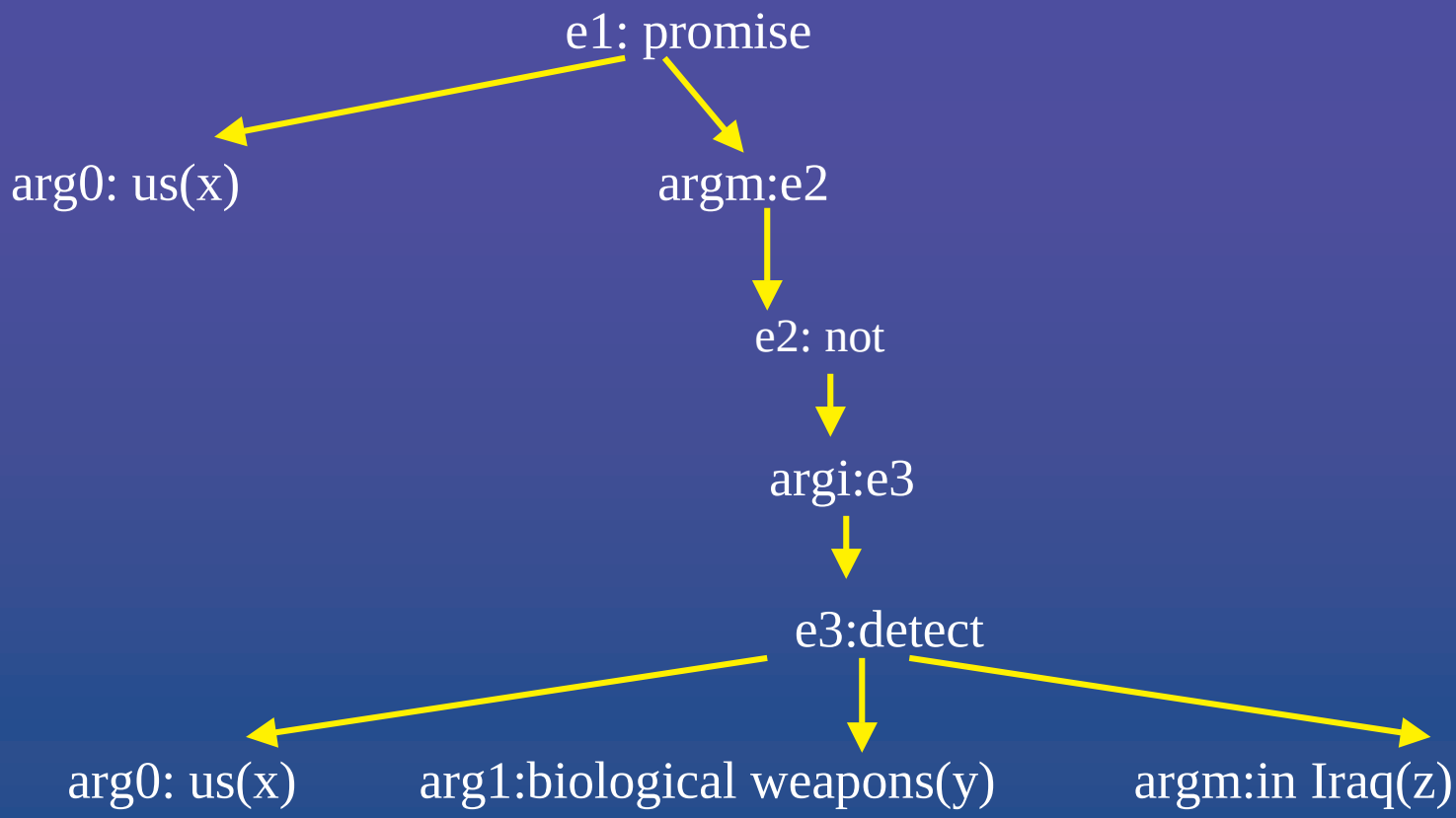
Linking the Predicate Argument Structures

Does US promise to detect biological weapons in Iraq?



Modality, Scope and Context Information

Does US promise not to detect biological weapons in Iraq?



Richer Semantics for QA and IE

Question:

What did the researchers report about asbestos?

⇒

[PRED: report,
ARG0: researchers,
ARG1: ?/asbestos]

Answer Text:

A form of asbestos … has caused a high percentage of cancer deaths …, researchers reported …

⇒

[PRED: report,
ARG0: researchers,
ARG1: [PRED: cause,
ARG0: asbestos,
ARG1: a high percentage of cancer deaths]]

Answer Extraction/Generation

?(asbestos)

= cause(arg0:asbestos, arg1: a high percentage of cancer deaths)

semantic resolution

?=λ x. cause(arg0:x, arg1: a high percentage of cancer deaths)

NL generation

Researchers reported that asbestos are something, that cause a high percentage of cancer deaths

Necessity of Richer Semantics

After the retirement of Peter Smith,
Mary Hopp was asked to take over the development sector

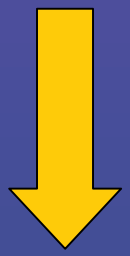
Flat Predicate Argument Structures

{
[PRED: *ask*,
ARG0: __,
ARG1: *Mary Hopp*,
ARG2: *take over the development sector*],

[PRED: *take_over*,
ARG0: **?**
ARG1: *the development sector*]
}

Modality and Truth Conditions

After the retirement of Peter Smith, *Mary Hopp was asked to take over the development sector*



IE

PERSON-IN (Mary Hopp)

Modality and Exact Answer

After the retirement of Peter Smith,
Mary Hopp was asked to take over the development sector

Who took over the development sector after the retirement of Peter Smith?

Information Merging and Fusion

(NYT16) NEW YORK -- Oct. 13, 1998 -- SCI-NOBEL-PHYSICS-CHEMISTRY, 10-13 –
The Nobel Prizes in Physics and Chemistry were announced Tuesday by the Royal Swedish Academy of Sciences.
Dr. Horst Stoermer, 49, a German-born professor who works at both Columbia University in New York and at Bell Laboratories in Murray Hill, N.J., is one of the three winners of the physics prize. (Suzanne DeChillo/New York Times Photo)

<PrizeAnnouncement, Nobel, {Physics,Chemistry}, {Tuesday, 1998}, Royal Swedish Academy of Sciences>

< PrizeWinner, Dr. Horst Stoermer, Nobel, Physics, 1998>

Outlook

- IE emerged as an inferior but achievable alternative to full text understanding.
- However, we believe that IE is not just an shortcut to doable applications but also another research strategy in our quest for language understanding.
- IE equipped with a pragmatic but solid semantic foundation and increasing contributions from deep processing methods will serve as a controlled and well-understood stepwise approximation to language understanding.