

# 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

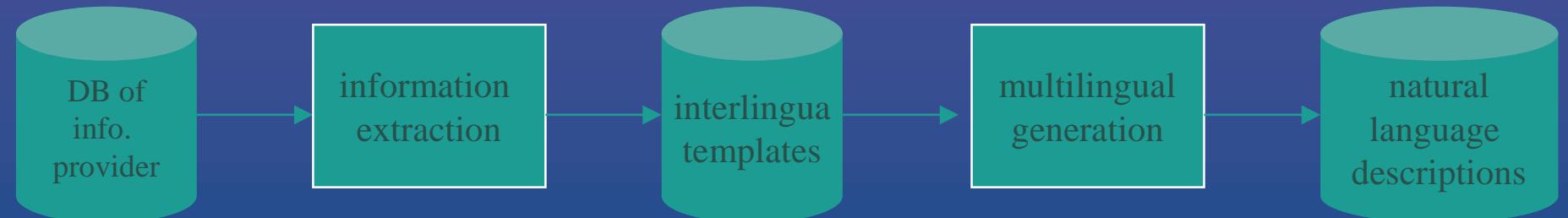
- Integrated access to heterogeneous data sources and make it fully transparent to end users whether they are searching in
  - ★ WWW documents or
  - ★ Databases



# 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

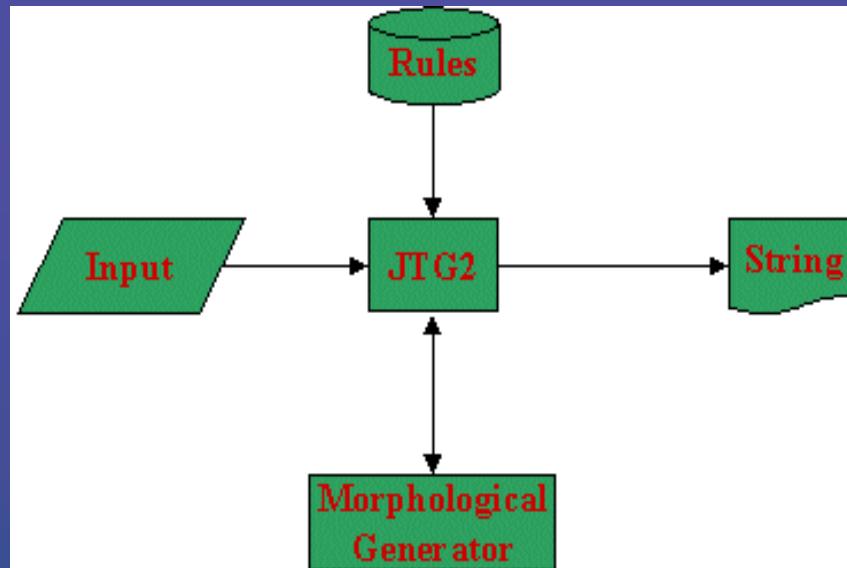
address: Club room

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:

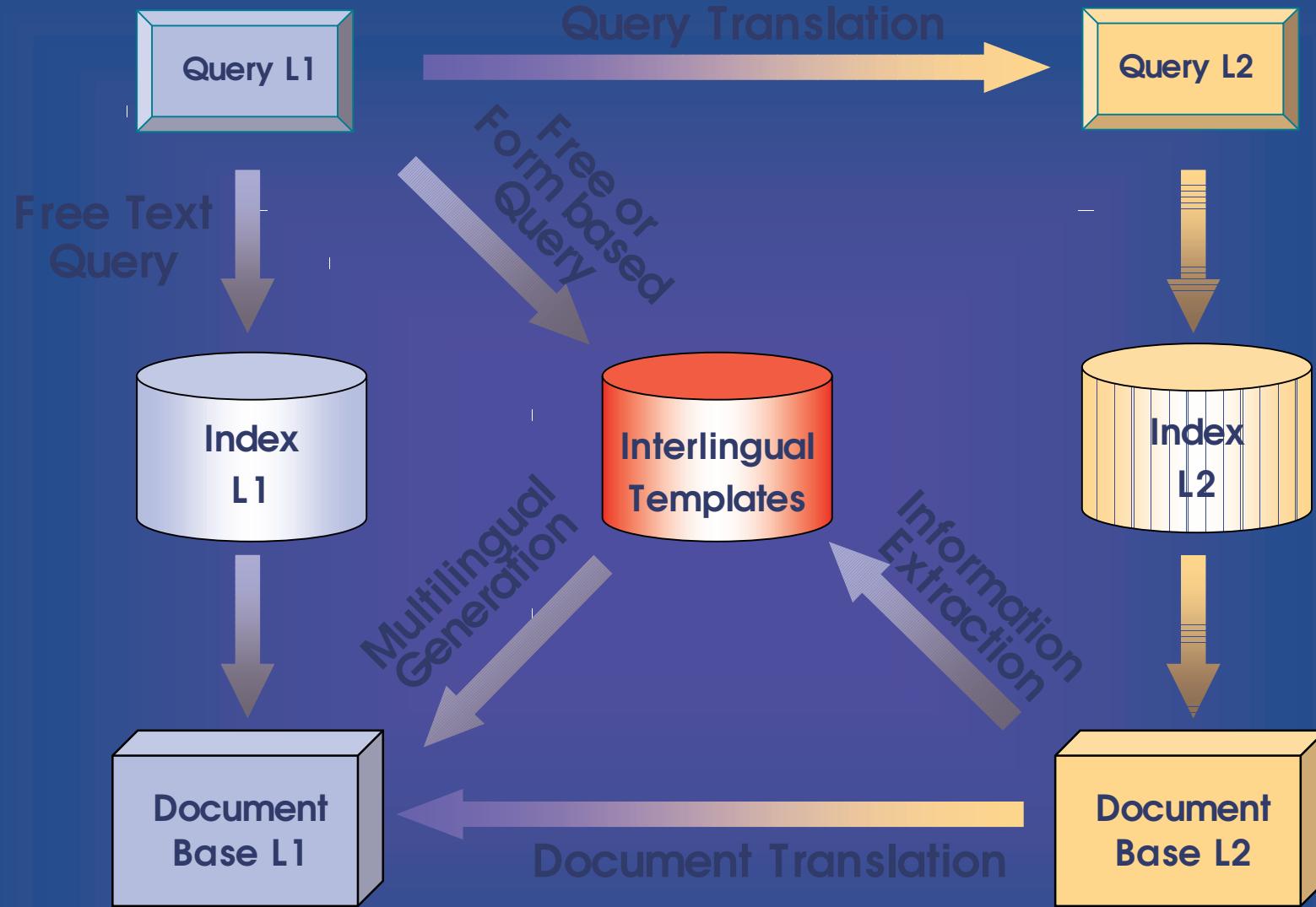
The theater show Faust will take place at the Staatstheater in Schillerplatz 1, 66111 Saarbrücken (in the downtown area).

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

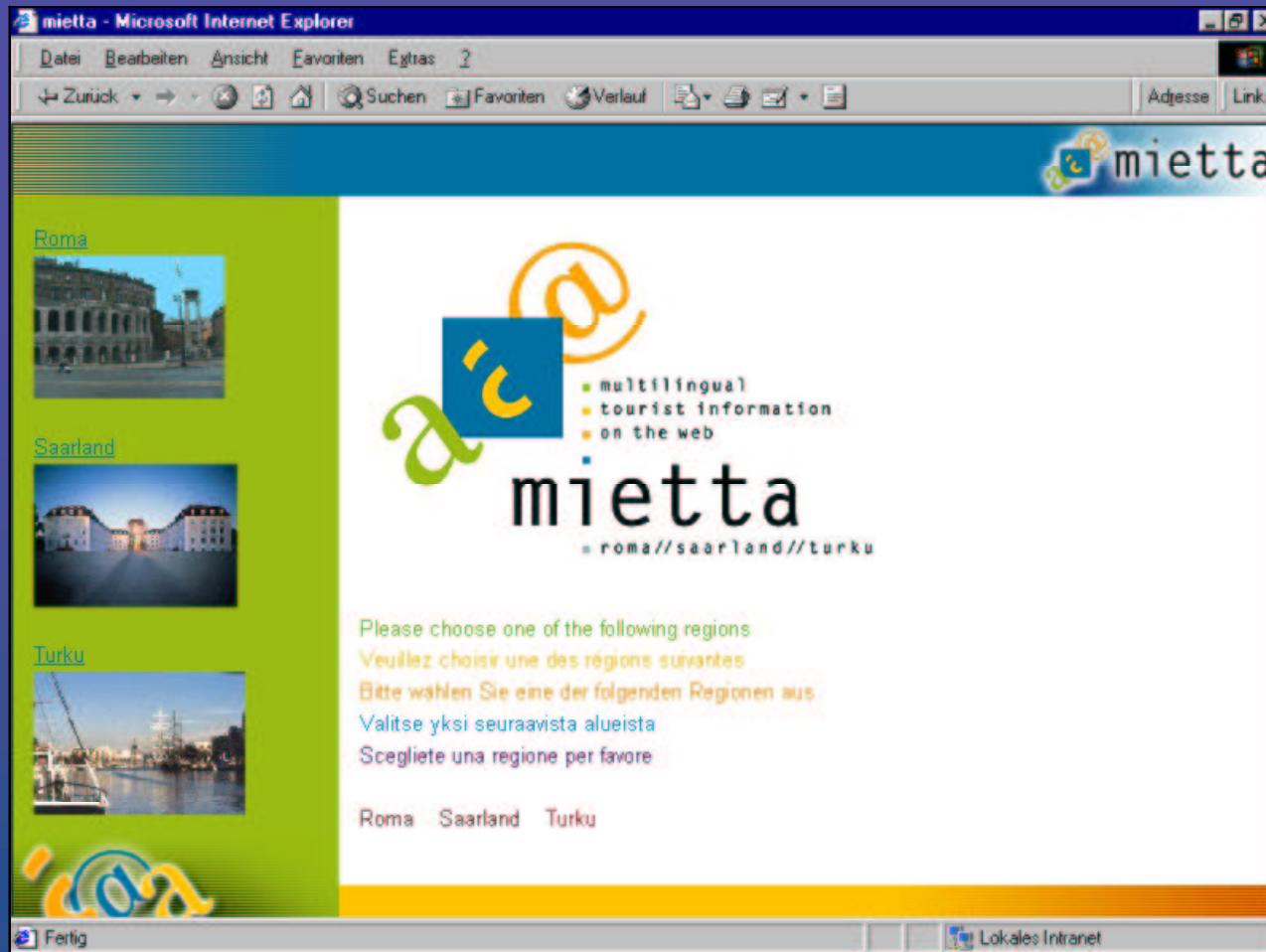
Finnish:

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





# MIETTA Start Page: Choose Region



# Choose Language

The screenshot shows a Microsoft Internet Explorer window with the title bar "mietta - Microsoft Internet Explorer". The menu bar includes "Datei", "Bearbeiten", "Ansicht", "Favoriten", "Eigenschaften", and "?". The toolbar includes "Zurück", "Vorwärts", "Haus", "Suchen", "Favoriten", "Verlauf", and "Adressleiste". The address bar shows the URL "mietta". The main content area features a blue header with the "mietta" logo. On the left, there is a green sidebar titled "Saarland" containing three small images: a building at night, a landscape with a river, and a city skyline. To the right, the text "Willkommen im Saarland" is followed by five language options with their respective flags: German (Germany), Italian (Italy), English (United Kingdom), French (France), and Finnish (Finland). At the bottom, there is a yellow footer bar with icons for "Fertig" (Finish) and "Lokales Intranet".



# MIETTA Search Menu

The screenshot shows the MIETTA search engine interface within a Microsoft Internet Explorer window. The title bar reads "mietta - Microsoft Internet Explorer". The menu bar includes "Datei", "Bearbeiten", "Ansicht", "Favoriten", "Extras", and a question mark icon. The toolbar contains standard icons for Back, Forward, Stop, Home, Search, Favorites, History, and others. The main content area has a blue header with the "mietta" logo featuring a stylized orange and yellow gear-like icon. Below the header, the text "mietta Search page" is displayed. A green sidebar on the left contains a search input field labeled "Type in your query:", a dropdown menu titled "Guided Search. Choose one or more categories:" with the option "What are you searching for?", and three buttons: "Go!", "Expand", and "Reset". Below these is a dropdown menu for "hits per page" set to "10". At the bottom of the sidebar is a yellow "Fertig" button. The main content area describes MIETTA as an advanced multilingual search engine for tourist information, mentioning its ability to search web documents and database information across multiple categories. It also explains how users can perform free text searches or guided searches by category. A note cautions that machine-translated documents may not be checked by a human operator. The bottom of the page features a yellow footer bar with the "Internet" icon.



# MIETTA Free Text Retrieval

mietta - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?  
Zurück Suchen Favoriten Verlauf Adresse Links

Type in your query:  
castles

Guided Search. Choose one or more categories:  
What are you searching for?

hits per page 10

Major Tourist Attractions 6 1  
General Information 2 1  
Nature and Sporting Activities 1 1  
Events 1 1

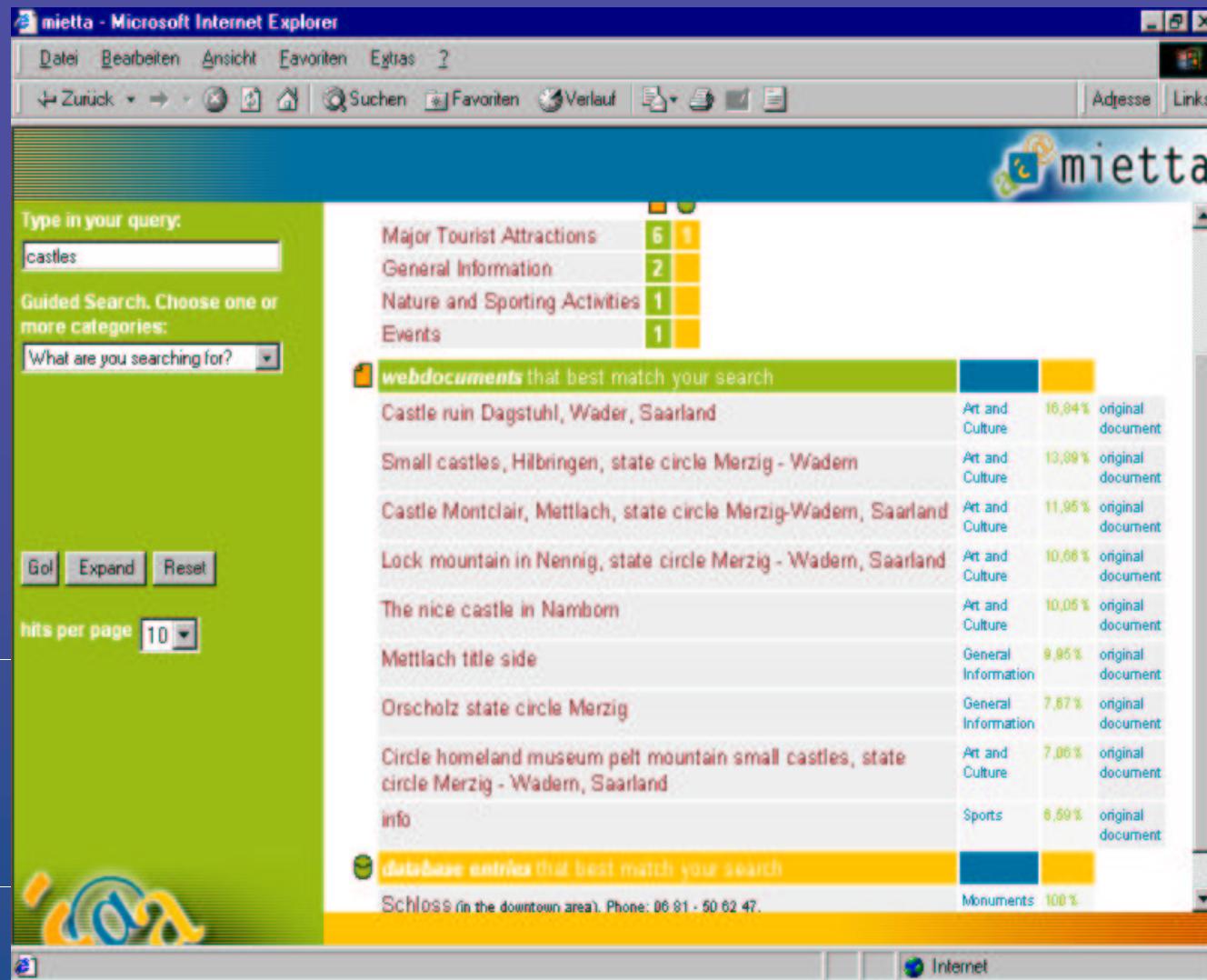
webdocuments that best match your search

Castle ruin Dagstuhl, Wader, Saarland	Art and Culture	16,04%	original document
Small castles, Hilbringen; state circle Merzig - Wadern	Art and Culture	13,99%	original document
Castle Montclair, Mettlach, state circle Merzig-Wadern, Saarland	Art and Culture	11,95%	original document
Lock mountain in Nennig, state circle Merzig - Wadern, Saarland	Art and Culture	10,06%	original document
The nice castle in Namborn	Art and Culture	10,06%	original document
Mettlach title side	General Information	9,95%	original document
Orscholz state circle Merzig	General Information	7,87%	original document
Circle homeland museum pelt mountain small castles, state circle Merzig - Wadern, Saarland	Art and Culture	7,06%	original document
info	Sports	6,59%	original document

database entries that best match your search

Schloss (in the downtown area). Phone: 06 81 - 50 62 47.	Monuments	100%
--	-----------	------

Internet



# MIETTA Class-based Navigation

The screenshot shows a search results page from the MIETTA system. At the top left is a search bar with placeholder text "Type in your query:". Below it is a section titled "Guided Search. Choose one or more categories:" containing three dropdown menus: "Major Tourist Attractions", "Art and Culture", and "Which cultural interest?". To the right of these is a message: "Your search returned 142 webdocuments and 15 database entries." Below this, another message says: "Some of them belong to the following more specific classes. You can click on one of the classnames to get to these specific results." A grid of three items is shown: "Industrial Culture" (with a green icon), "Galleries" (with a yellow icon), and "Monuments" (with a blue icon). To the right of the grid are buttons for "next page" and "11 - 20".

**webdocuments that best match your search**

bauwerke_mfr.htm	Art and Culture	original document
Abbey buildings Mettlach, state circle - Merzig - Wadern, Saarland	Art and Culture	original document
Mettlach: Old tower	Art and Culture	original document
Castle ruin Dagstuhl, Wadern, Saarland	Art and Culture	original document
Lock Dagstuhl, Wadern-Dagstuhl, circle Merzig-Wadern, Saarland	Art and Culture	original document

**database entries that best match your search**

Staatstheater (in the downtown area). Phone: 06 81-32204.	Monuments
Alte Feuerwache (in the downtown area). Phone: 06 81 - 30 92 203.	Art and Culture



# MIETTA Class-based Navigation with Free Text

The screenshot shows a Microsoft Internet Explorer window displaying the MIETTA search results. The title bar reads "mietta - Microsoft Internet Explorer". The menu bar includes "Datei", "Bearbeiten", "Ansicht", "Favoriten", "Extras", and a question mark icon. The toolbar contains standard buttons for Back, Forward, Stop, Home, Search, Favorites, History, and others. The address bar shows the URL "mietta". The logo "mietta" with a stylized orange and green icon is visible in the top right corner.

**Type in your query:**  
church

**Guided Search. Choose one or more categories:**  
Major Tourist Attractions  
What type of place?

Where in the city?

hits per page

Your search returned 4 webdocuments

Some of them belong to the following more specific classes. You can click on one of the classnames to get to these specific results.

Art and Culture 4

webdocuments that best match your search		
Pfarrkirche St.Jacobus, showing churches, circle Merzig - Wadern, Saarland	Art and Culture	30.59 % original document
Curator office: Monument list Saarland, nine churches	Art and Culture	28.72 % original document
Curator office: Monument list Saarland, showing churches	Art and Culture	24.81 % original document
Curator office: Monument list Saarland, Freisen	Art and Culture	15.88 % original document

Fertig Internet



# MIETTA Form based Query

The screenshot shows a Microsoft Internet Explorer window displaying the MIETTA search interface. The title bar reads "mietta - Microsoft Internet Explorer". The menu bar includes "Datei", "Bearbeiten", "Ansicht", "Favoriten", "Extras", and a question mark icon. The toolbar contains icons for Back, Forward, Stop, Home, Search, Favorites, History, and Address/Links.

The main content area features a logo for "mietta" with a stylized blue and yellow "i" icon. A green sidebar on the left contains the text "Type in your query:" followed by a text input field. Below it is a section titled "Guided Search. Choose one or more categories:" with three dropdown menus: "Major Tourist Attractions" (set to "Art and Culture"), "Art and Culture", and "Which cultural interest?". It also includes a dropdown for location ("Center") and buttons for "Go!", "Expand", and "Reset". A dropdown for "hits per page" is set to "10".

The main search results area displays a list of cultural entities in the downtown area of Saarbrücken, each with its name, phone number, and category. The results are as follows:

Entity	Phone Number	Category
Theater Blauer Hirsch	06 81 / 5 84 99 49	seat of cultural events
Kultursaal Hüttigweiler	0681-000000	seat of cultural events
Camera 2	06 81 - 3 49 96	seat of cultural events
Camera	06 81 - 3 49 96	seat of cultural events
Theater Annual	0681-32204/ 32206; Fax: 0681- 3092-316	seat of cultural events
Theater im Viertel	06 81 - 39 04 80 2	theater
Schauplatz, im Filmhaus	06 81 / 37 25 70	movie
Filmhaus Saarbrücken	06 81 - 37 25 70	movie
Kino 8 1/2	06 81 - 390 88 80	movie
Staatstheater	06 81-32204	theater



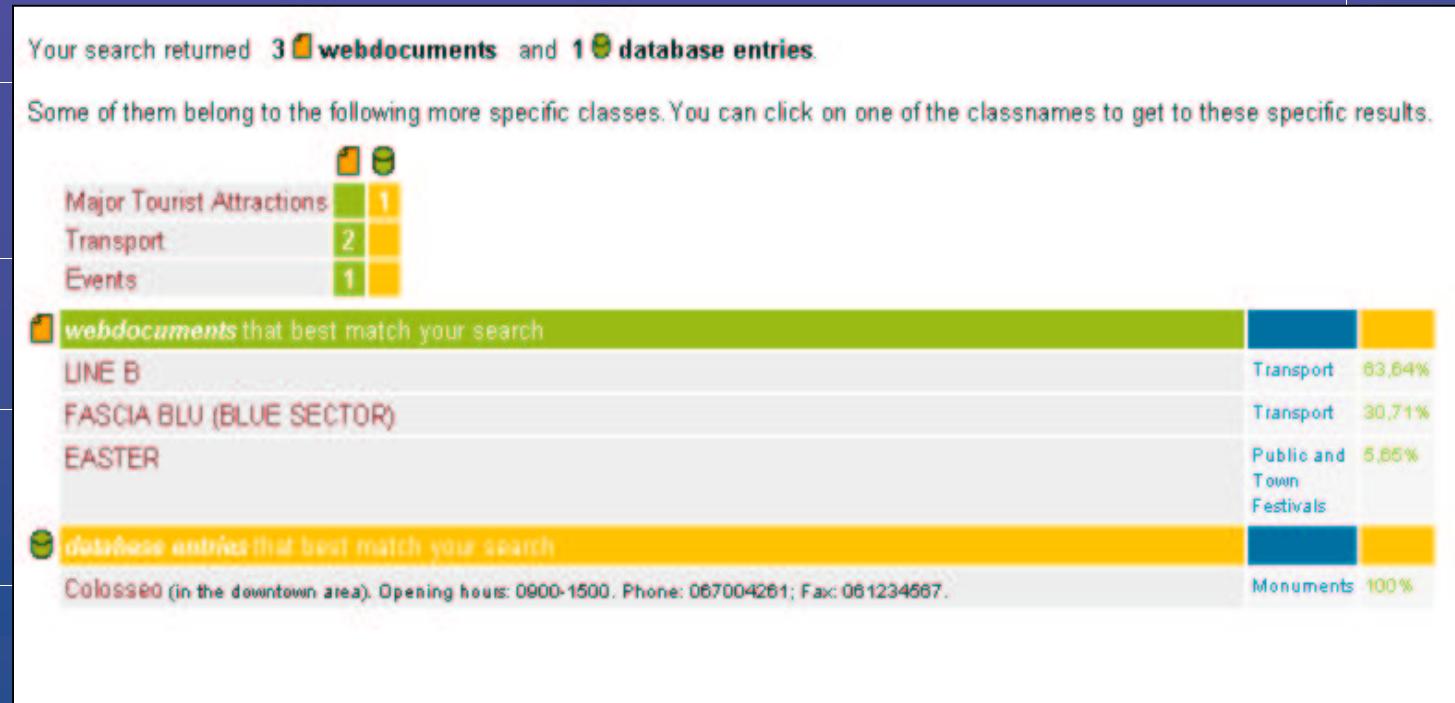
# Online Text Generation

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

LI 2004

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	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:  $\text{gender}(x)$
- time-dependent attribute:  $\text{age}(x)$
- timeless two-place relation:  $\text{father}(x, y)$
- time-dependent two-place relation:  $\text{boss}(x, y)$



# 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  
 $\langle \text{relation} : y_i \rangle$  is called a role assignment (or a feature).
- name  $\langle x, CR \rangle$

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

appointer: Bush  
appointee: Rice  
office: SecurityAdvisor



# 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

The screenshot shows the Protégé-2000 interface for the 'prize-winner-light' ontology. The left pane displays a class hierarchy under the 'Relationship' tab, with 'entity' as the root node. The right pane shows the detailed configuration for the 'prize' class, which is a standard class. The 'Template Slots' table lists the following slots:

Name	Type	Cardinality	Other Facets
S establishedBy	Instance	required single	classes=(organization,human)
S awardedBy	Instance	single	classes=(cognitiveAgent)
S winner	Instance	required multiple	classes=(cognitiveAgent)
S achievement	String	required multiple	
S administratedBy	Instance	required single	
S trophy	String	multiple	
S area	Instance	single	
S monetaryValue	Instance	required single	classes=(currencyMeasure)
S foundationTime	Instance	required single	classes=(timeMeasure)
S x-annual	Float	required single	
S medal	String	single	
S diploma	String	single	

The 'Superclasses' panel indicates that 'prize' is a subclass of 'RepresentationalArtWork' and 'award'.

# 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<sub>1</sub>)  
year(e<sub>1</sub>, "1996")  
recipient(e<sub>1</sub>, x<sub>1</sub>)  
area(e<sub>1</sub>, a<sub>1</sub>)

### to be explicit:

nobel\_prize\_event(e<sub>1</sub>)  
nobel\_prize\_year(e<sub>1</sub>, "1996")  
nobel\_prize\_recipient(e<sub>1</sub>, x<sub>1</sub>)  
nobel\_prize\_area(e<sub>1</sub>, a<sub>1</sub>)

## Questions

$\lambda x.\text{receiving\_nobel\_prize } (\text{e}_1, \text{achievement}, \text{"nobel\_prize"}, \text{"physics"}, \text{"1996"}, x, \text{co-recipients}, \text{location})$   
 $\lambda x.\text{recipient}(\text{e}_1, x)$



# Simple Extensional Denotation of "Nobelpreisträger"

```
nobel_prize_winner' =  
λx [person(x) ∧ ∃e (nobel_prize_event(e) ∧ nobel_prize_recipient(e1, x1))]
```

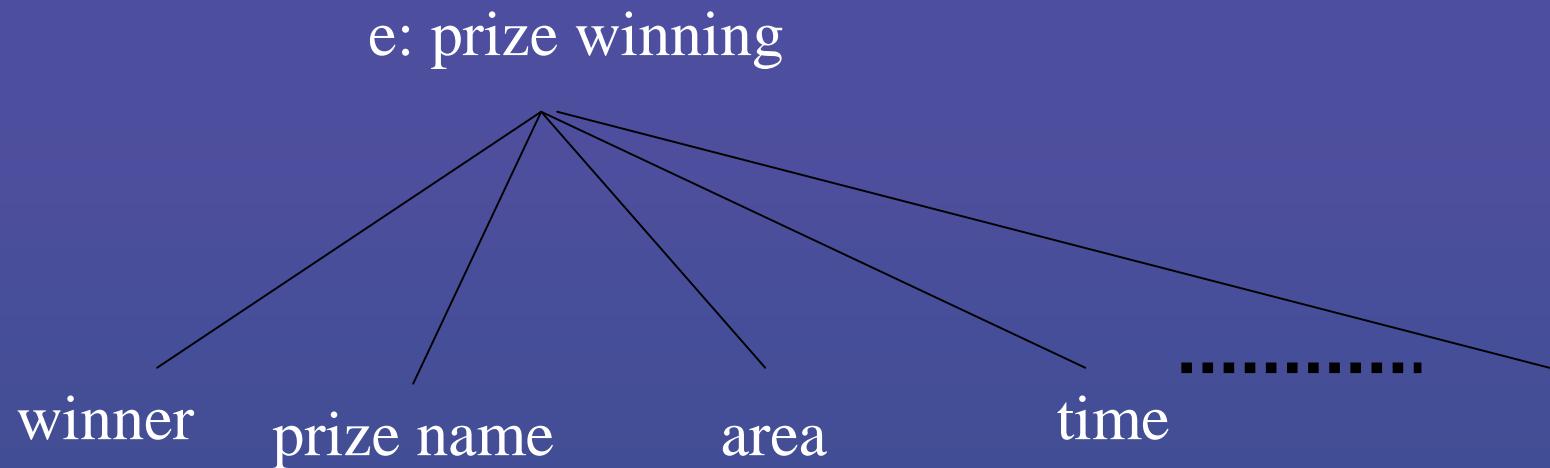


# Pragmatic Approach to Relation Representation

- 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)



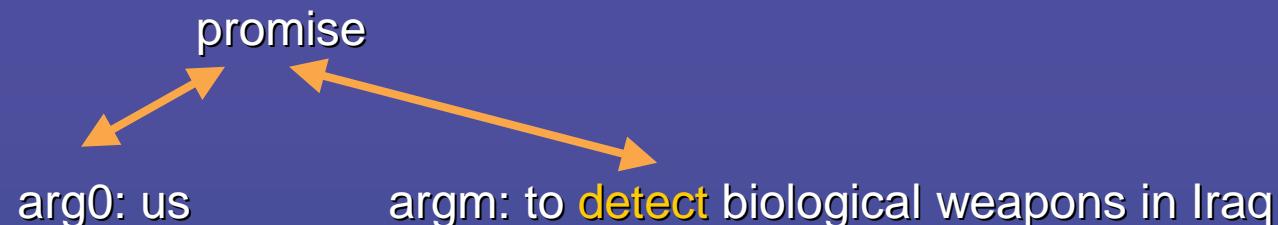
# Semantic Labelling for IE

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



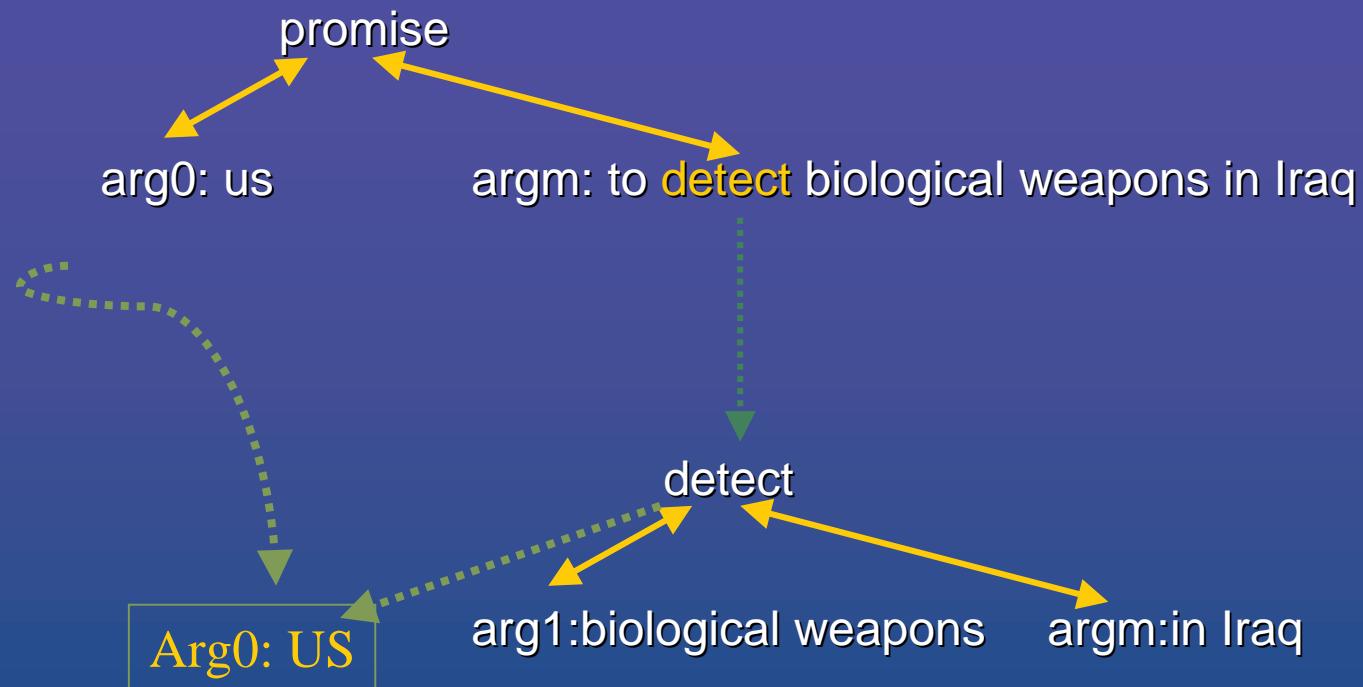
# Flat Predicate Argument Structures

Does US promise to detect biological weapons in Iraq?



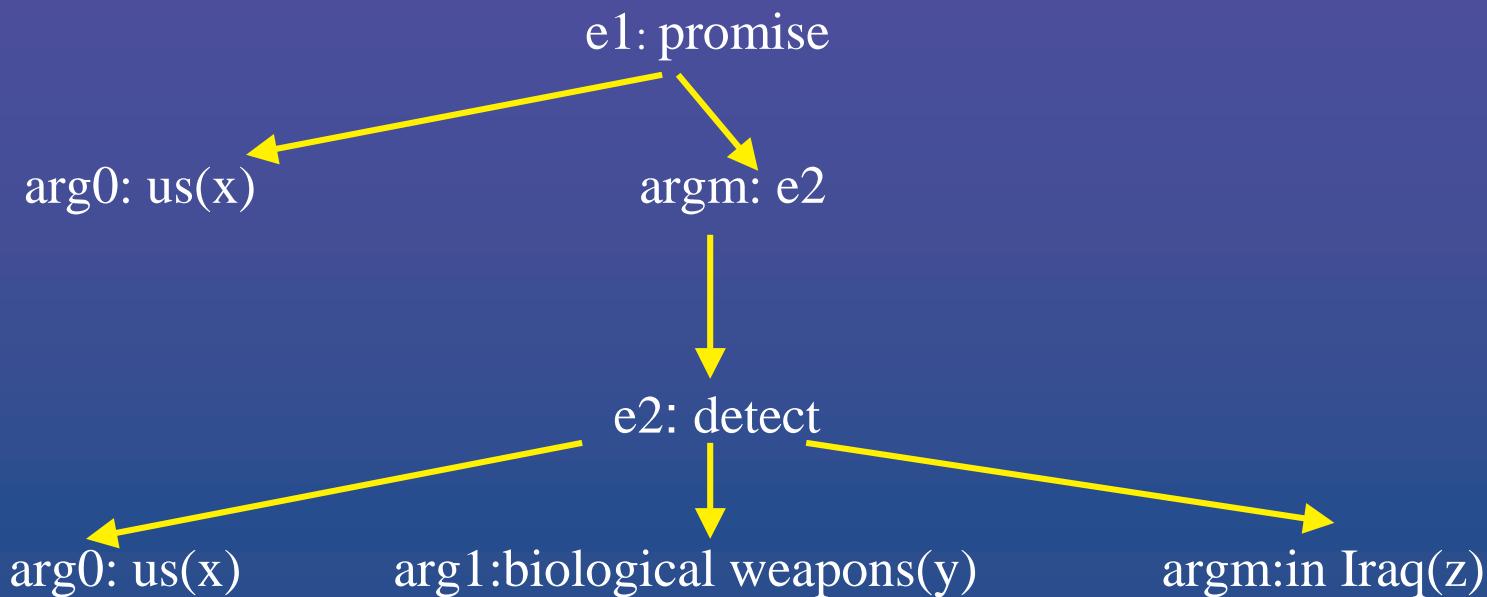
# Flat Predicate Argument Structures

Does US promise to detect biological weapons in Iraq?



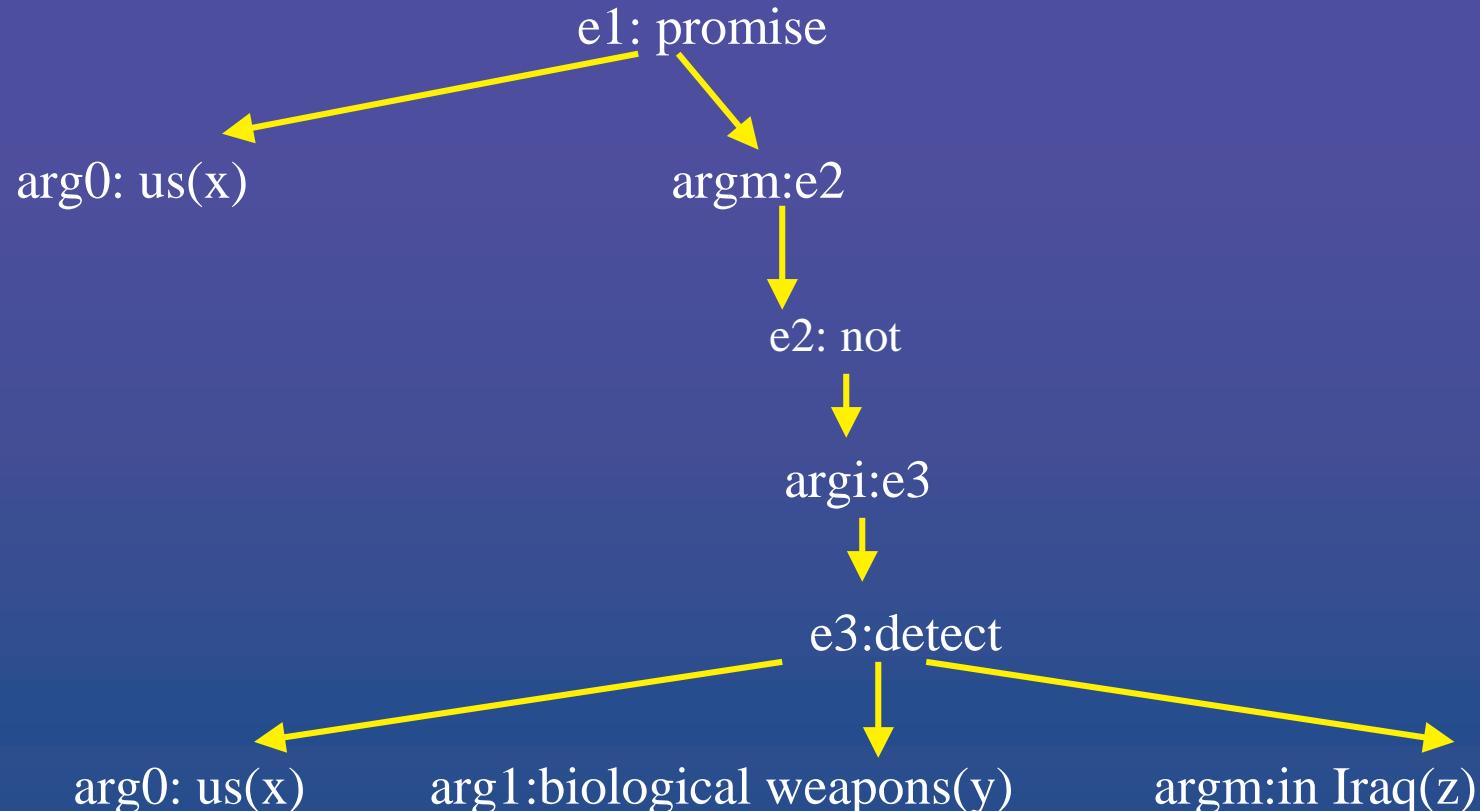
# Linking the Predicate Argument Structures

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)



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



*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



# 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.

