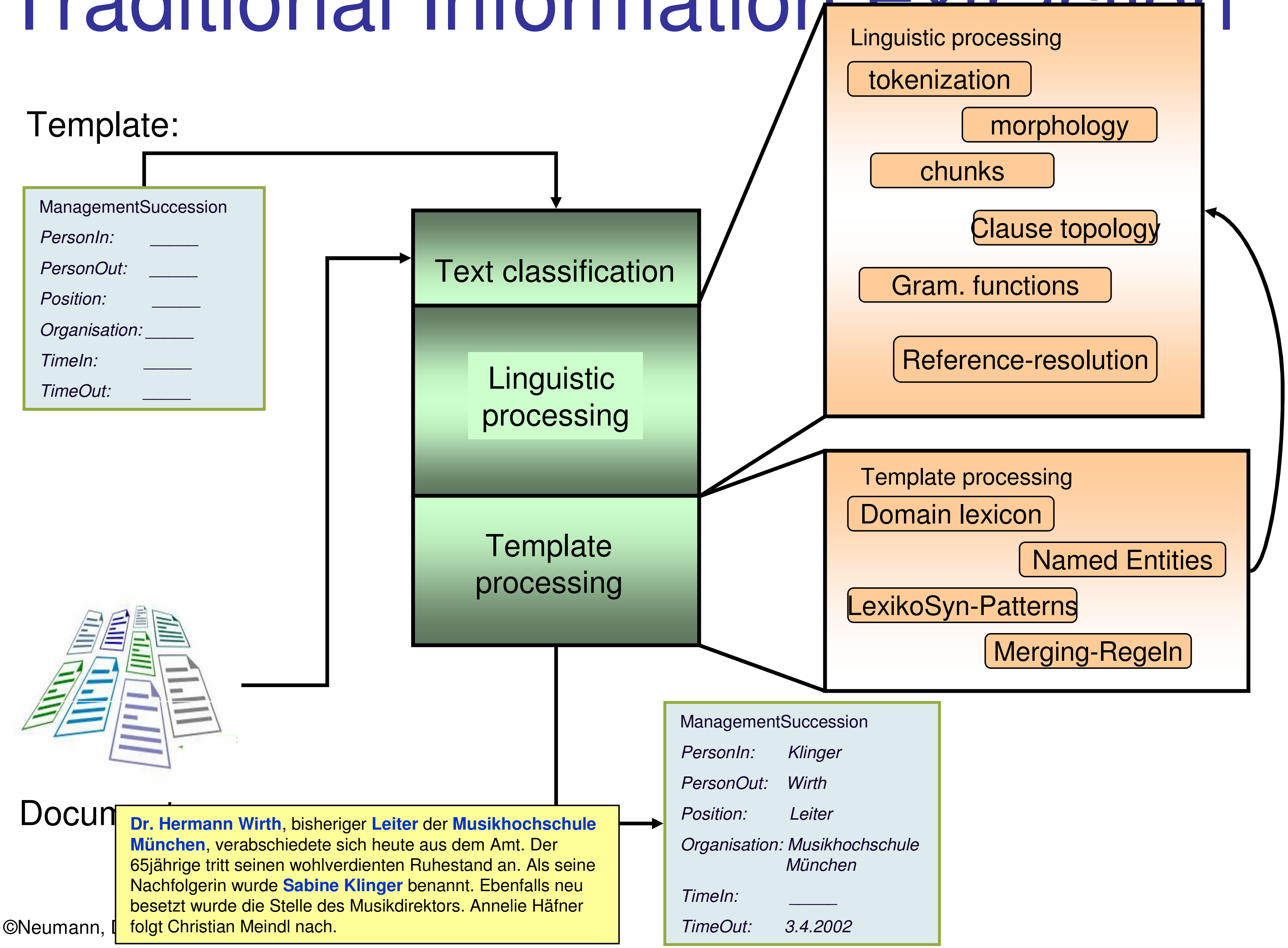
Towards Dynamic Interactive Information Extraction

Günter Neumann

LT-lab, DFKI, Saarbrücken

2008

Traditional Information Extraction



Template:

ManagementSuccession

PersonIn: _____

PersonOut: _____

Position: _____

Organisation: _____

TimeIn: _____

TimeOut: _____

Text classification

Linguistic processing

Template processing

Linguistic processing

- tokenization
- morphology
- chunks
- Clause topology
- Gram. functions
- Reference-resolution

Template processing

- Domain lexicon
- Named Entities
- LexikoSyn-Patterns
- Merging-Regeln

Docum

Dr. Hermann Wirth, bisheriger **Leiter** der **Musikhochschule München**, verabschiedete sich heute aus dem Amt. Der 65jährige tritt seinen wohlverdienten Ruhestand an. Als seine Nachfolgerin wurde **Sabine Klinger** benannt. Ebenfalls neu besetzt wurde die Stelle des Musikdirektors. Annelie Häfner folgt Christian Meindl nach.

ManagementSuccession

PersonIn:	*Klinger*
PersonOut:	*Wirth*
Position:	*Leiter*
Organisation:	*Musikhochschule München*
TimeIn:	_____
TimeOut:	*3.4.2002*

IE for semantic annotation

Identification of IE-sub-tasks:

- basic entities (e.g., proper names)
- binary relations between entities
- n-ary relations/events

Automatic Content Extraction (ACE)

- Spezification of an IE-core-ontology
- Annotation-specification & -tools
- Templates as specializations of the IE-core-ontology (also multi-templates)

IE as core for semantic annotation

- identification
- discovery
- validation
- evaluation

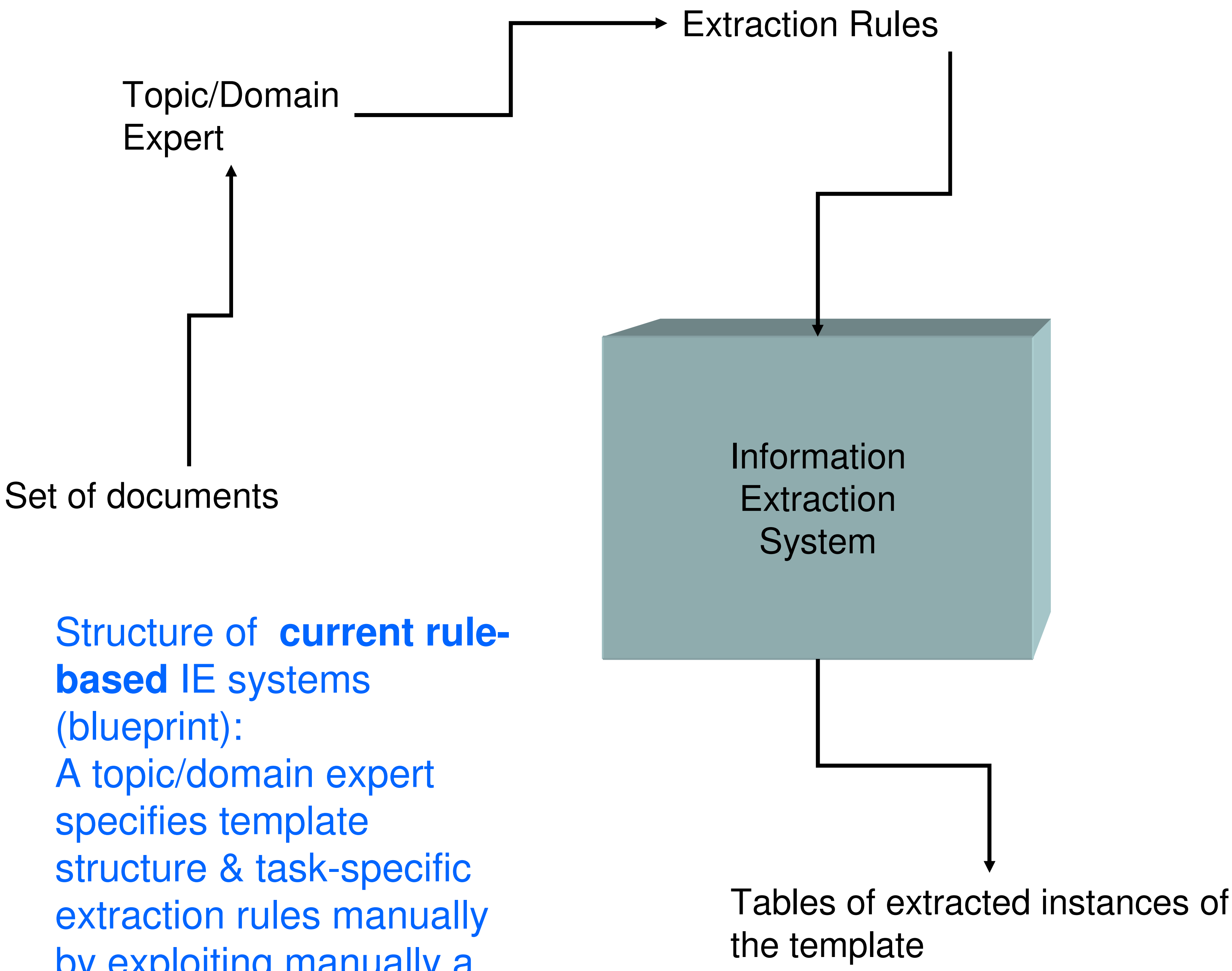
of semantic relationships & as basis for the automatic creation of meta data

An IE system can be seen as an interface between a template and text fragments

- An IE-template is a typed feature structure describing the structure of some information of interest
- An IE system consists of rules/constraints for feature filling & merging
- An IE-template must have an **exact, fixed definition**
- The rules are defined on the basis of a relevant corpus of textual instances of the IE-template

State-of-the-art IE systems

- Offline/static IE:
  - Relevant information in form of templates and relevant corpus is given to the IE system
- Approaches:
  - Manually implemented rule-based IE systems
  - Automatically induced data-driven IE systems



Extraction Rules

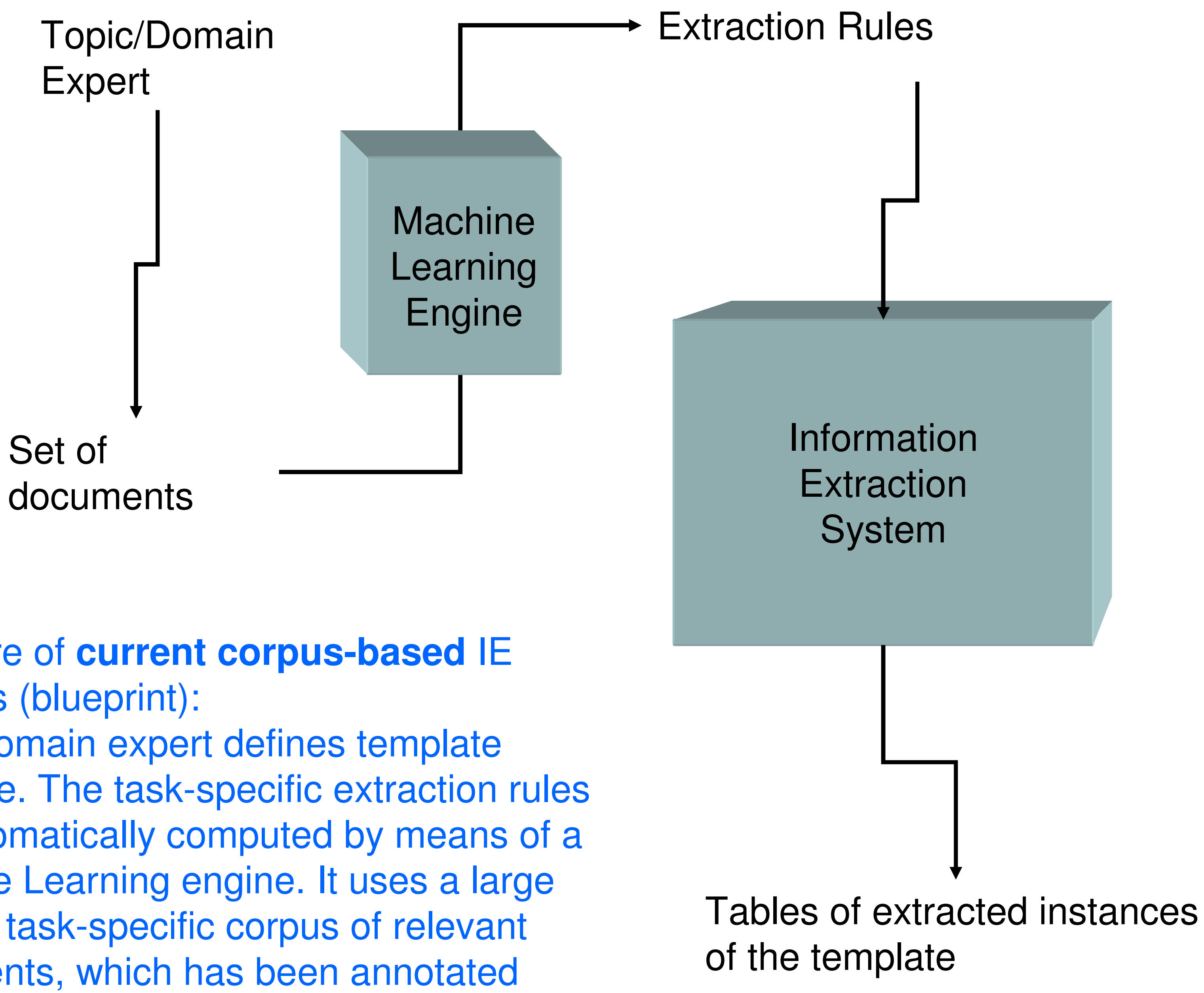
Topic/Domain Expert

Set of documents

Information Extraction System

Tables of extracted instances of the template

Structure of **current rule-based** IE systems (blueprint):
A topic/domain expert specifies template structure & task-specific extraction rules manually by exploiting manually a relevant corpus.


Topic/Domain Expert
Extraction Rules
Machine Learning Engine
Set of documents
Information Extraction System
Tables of extracted instances of the template

Structure of **current corpus-based** IE systems (blueprint):

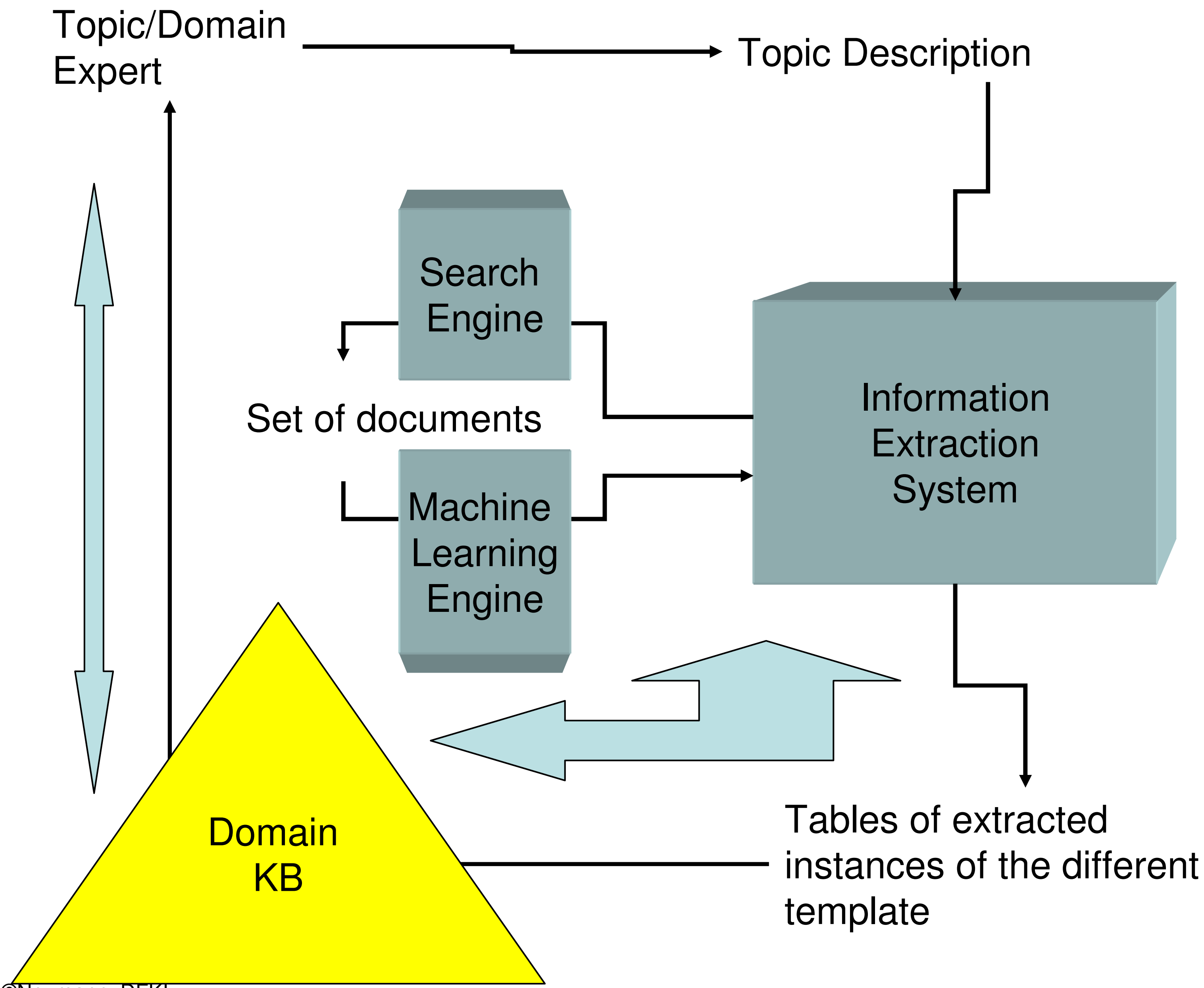
Topic/domain expert defines template structure. The task-specific extraction rules are automatically computed by means of a Machine Learning engine. It uses a large enough task-specific corpus of relevant documents, which has been annotated manually by a topic/domain expert.

Current IE systems are too inflexible

- An IE system needs an exact definition of a template
  - it must be known in advance how information is structured for a certain application AND paraphrased in documents
  - usually one IE system handles one template type
- IE systems are realized by means of a set of sub-components making use of simple and static information flow
- IE systems have no way of adapting themselves to the dynamics in information changes, e.g., to adapt the template structure and mapping rules

We need IE systems which emerge on specific user request

- User and IE system must interact
  - Different users have different interest/knowledge
  - User (goal-directed), IE system (data-oriented)
  - Dynamics of user request and document space
- IE system must be adaptive
  - Open (no fixed template structures, multiple templates)
  - Preemptive (predict all possible interesting template structures)
  - On-line (do on-demand and user-driven/personalized)



©Neumann, DFKI

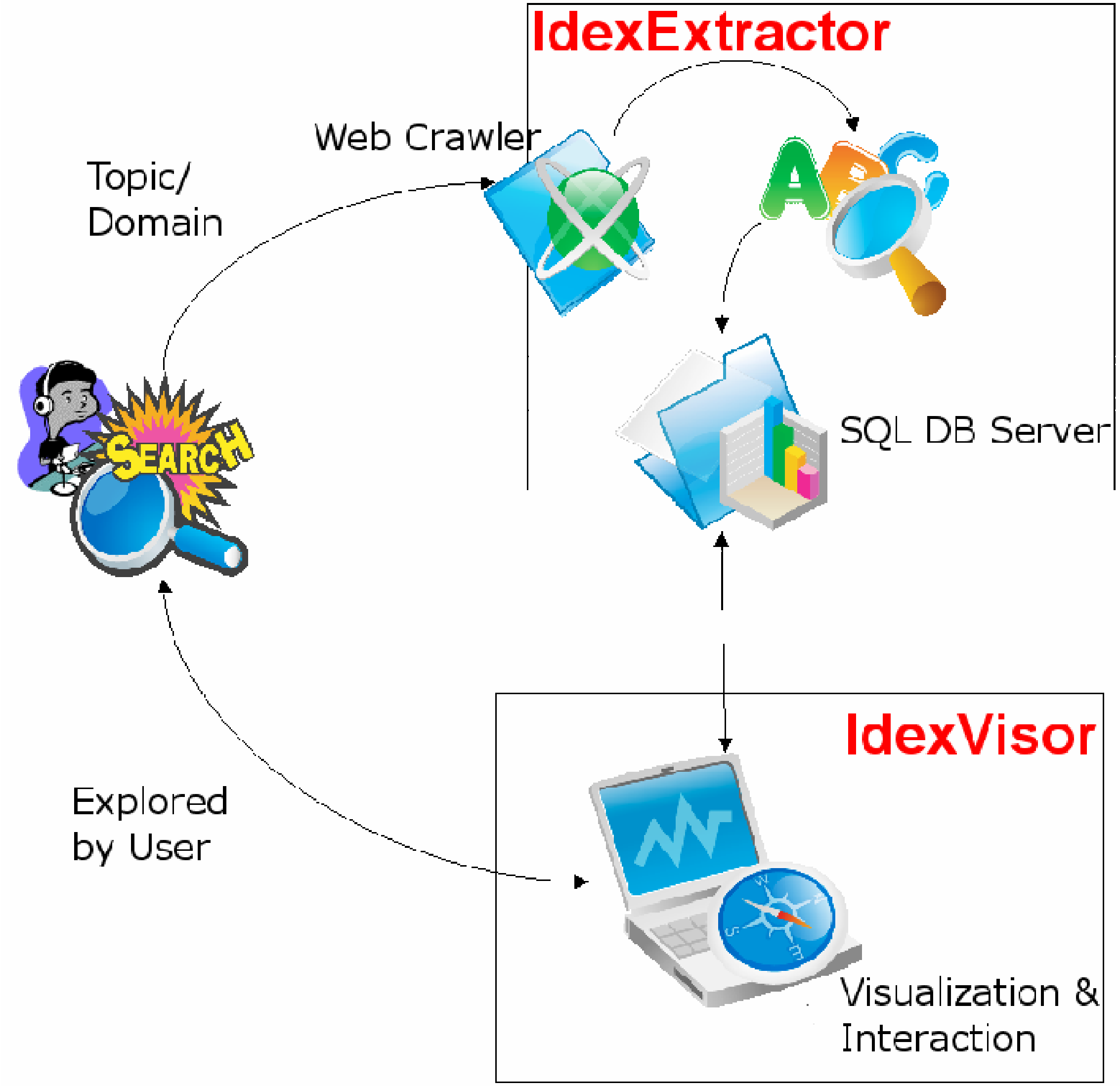
Recently developed new IE paradigms

- University of Washington (Etzioni's group)
  - Open IE from the Web (HLT 2006, IJCAI 2007)
  - Automatically discover possible relations of interest (tuples of for $<e_i, r_{ij}, e_j>$
  - Only make a single parse over the corpus, 9 M web pages
  - Self-supervised learning
- New York University (Sekine's group)
  - On-demand & preemptive IE (Coling 2006, HLT 2006)
  - Automatically identify the most salient structures and extract information on the topic the user demands
  - Unrestricted Relation Discovery
    - Pattern discovery,
    - paraphrase discovery
    - table construction

Some IE trends at LT lab of DFKI

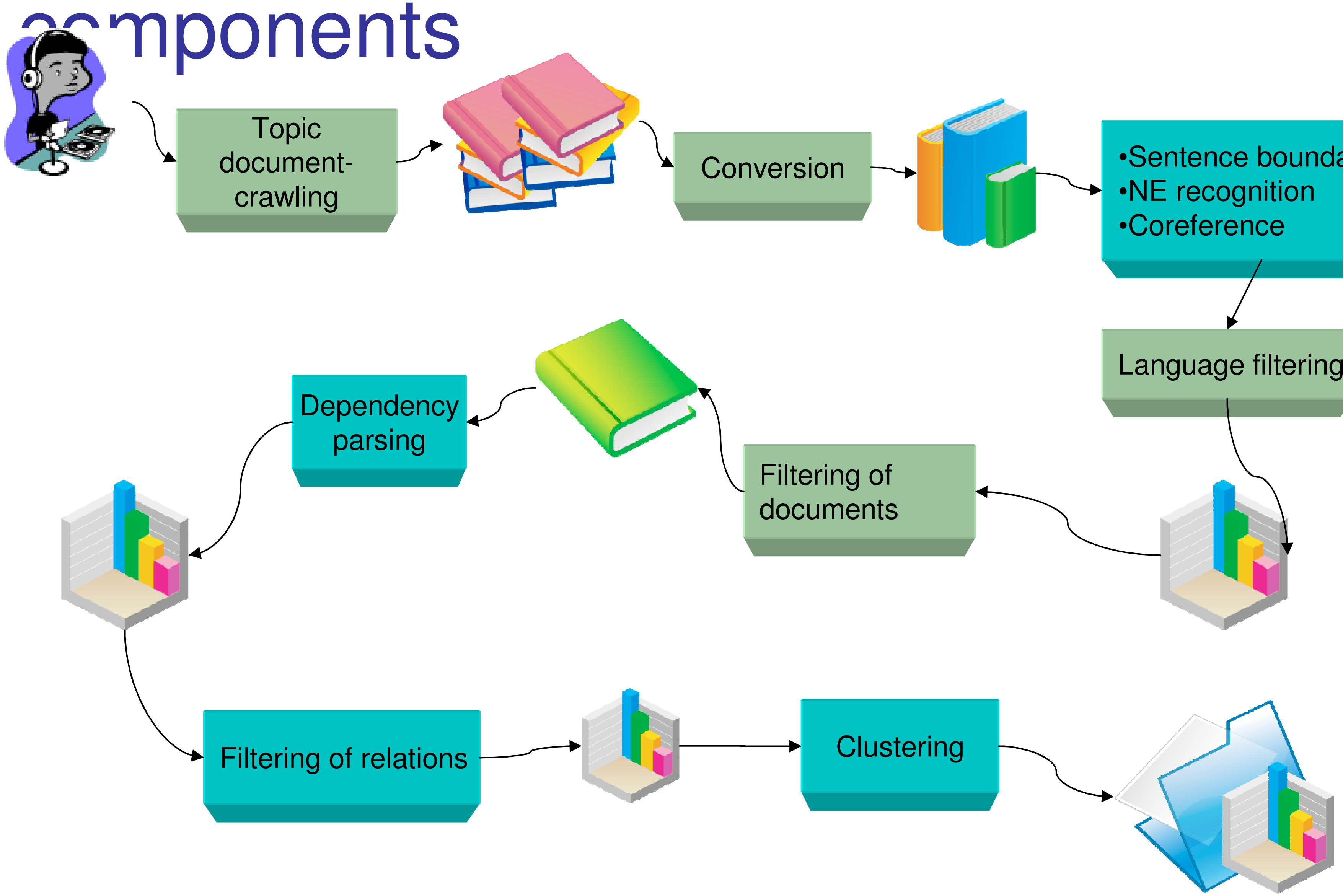
- EU project IDEX (<11.07):
  - Interactive Dynamic IE
  - Risk analysis management
  - Web People Search
- Project proposal DiLiA (>1.08: Digital Library Assistant (with specialization on BioIE))
  - Integrated shallow and deep IE (e.g., text structure, textual inference)
  - Personal virtual digital library (personal views and histories, sharable)
  - BioIE: extraction of protein-protein-interaction or other relations from (full) biomedical texts ⇒ BioCreative-II (active participation, 2007)
- BMBF project HyLaP (<12.08): Web-based open domain IE
  - Definitions (“What is X ?”), Enumerations (“List all instances of X!”)
  - Automatic creation of Search Engine queries & Latent Semantic Analysis
  - Unsupervised basis for ontology population (ongoing work)

The dynamic IE system IDEX



©Neumann, DFKI

IDEX: Language technology components



Topic document-crawling

Conversion

•Sentence boundaries
•NE recognition
•Coreference

Language filtering

Filtering of documents

Dependency parsing

Filtering of relations

Clustering

©Neumann, DFKI

IDEXExtractor: Experiments and results

- Test corpus: „Berlin central station“
  - 1068 web pages
  - 55255 sentences
  - 10773 relations
  - 306 clusters (two or more relations) – 81 clusters with identical relations
    - 121 consistent (i.e., all instances in the cluster express a similar relation)
    - 35 partly consistent (i.e., more than half of the instances in the cluster express a similar relation)
    - 69 not consistent

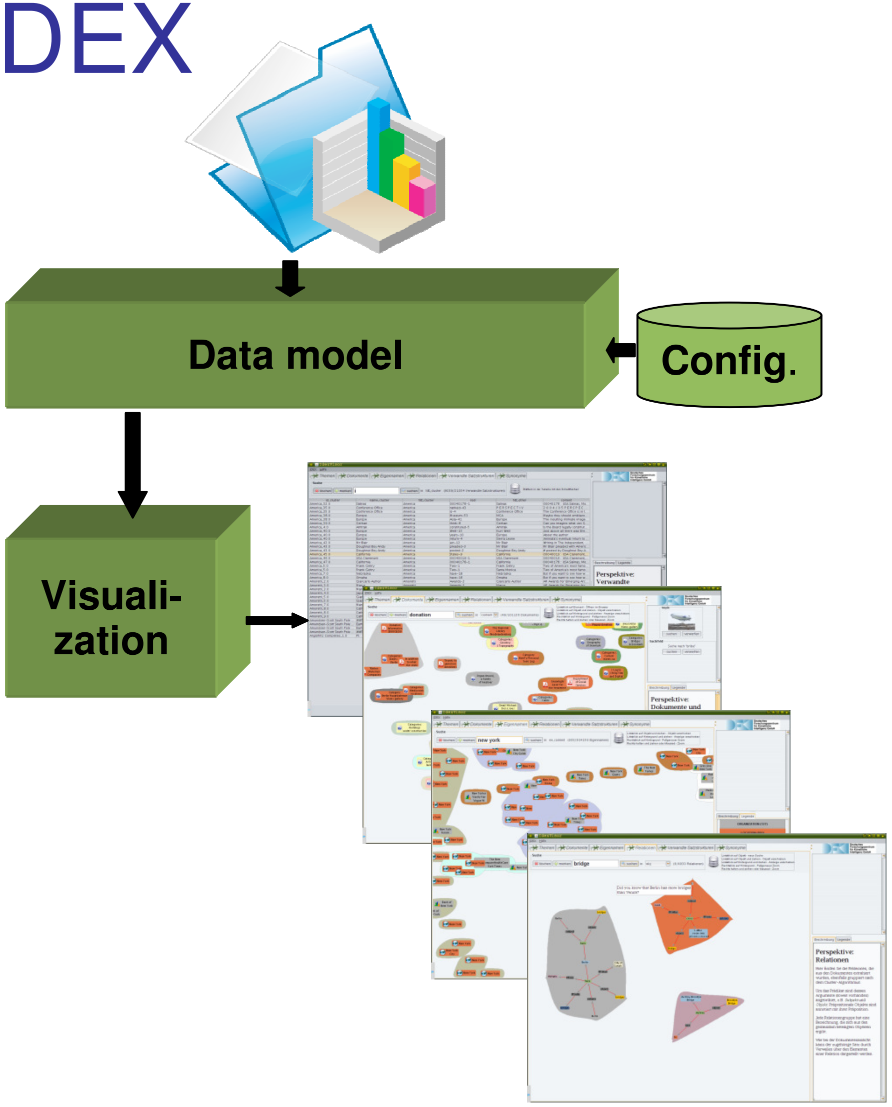
Types of clusters

- Relation paraphrases (18 clusters)
  - *accused(Mr Moore, Disney, In letter)*
  - *accused(Micheal Moore, Walt Disney Company)*
- Different instances of same pattern (76 clusters)
  - *operates(Delta, flights, from New York)*
  - *offers(Lufthansa, flights, from DC)*
- Relations about same topic (27 clusters)
  - *rejected(Mr Blair, pressure, from Labour MPs)*
  - *reiterated(Mr Blair, ideas, in speech, on March)*
  - *created(Mr Blair, doctrine)*

Similarity measures for relation clustering:

- the verbs have the same infinitives, or are in the same synonym set of Word net
- subjects and objects overlap (based on dependency parser information)
- NEs identical and/or NE types of subject and/or object match
  - including coreference resolution

IDEXVisor: Interactive Information Exploration using IDEX



Data model

Config.

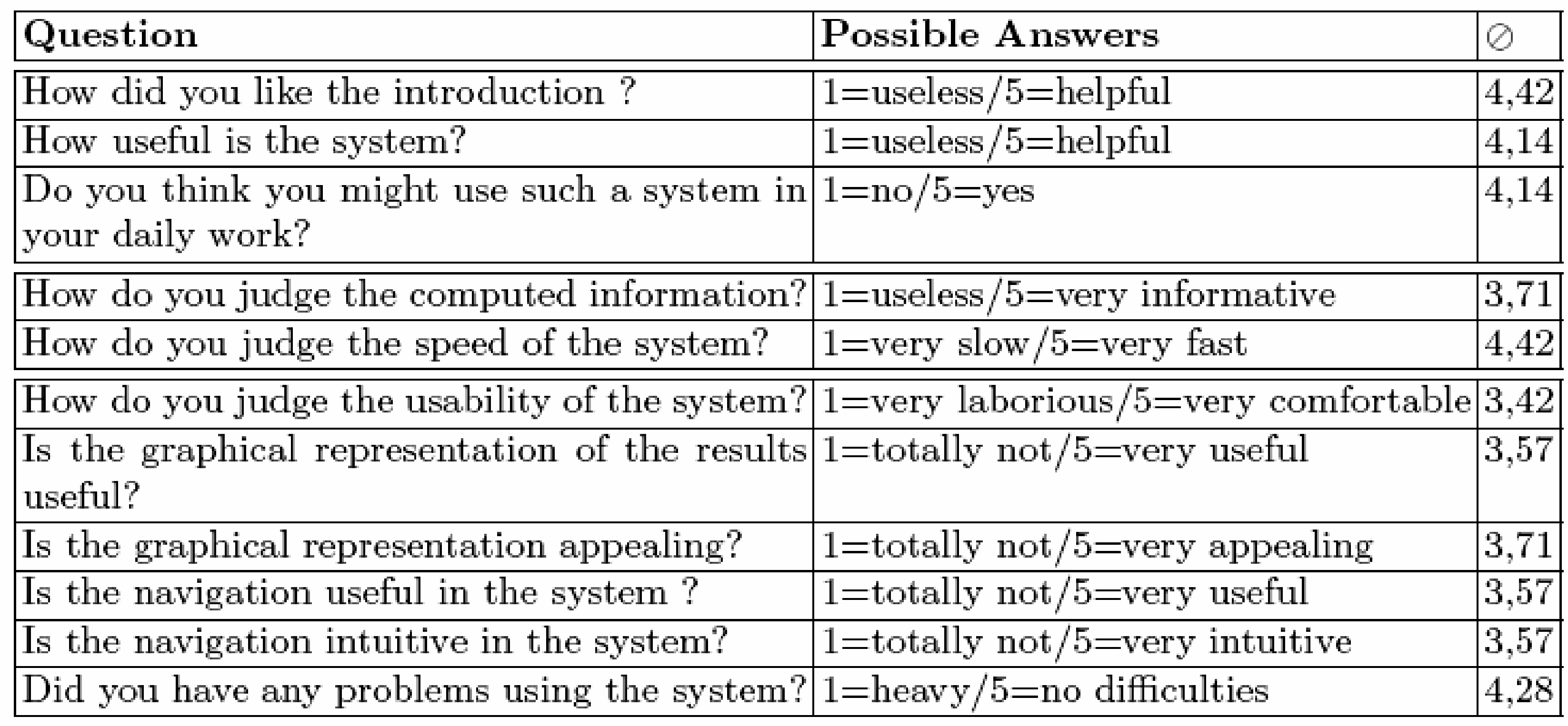
Visuali-zation

- Source
  - the extracted tables
- Goal/function
  - search
  - interaction
  - exploration
- Features
  - separation of the data model from the database
  - interactions and visualizations fitted to the data

©Neumann, DFKI

Evaluation of IDEXVisor

- Qualitative evaluation: 7 users, average age 33 years, 4 male, 3 female
- 4 corpus-related questions had to be solved via interaction with the system



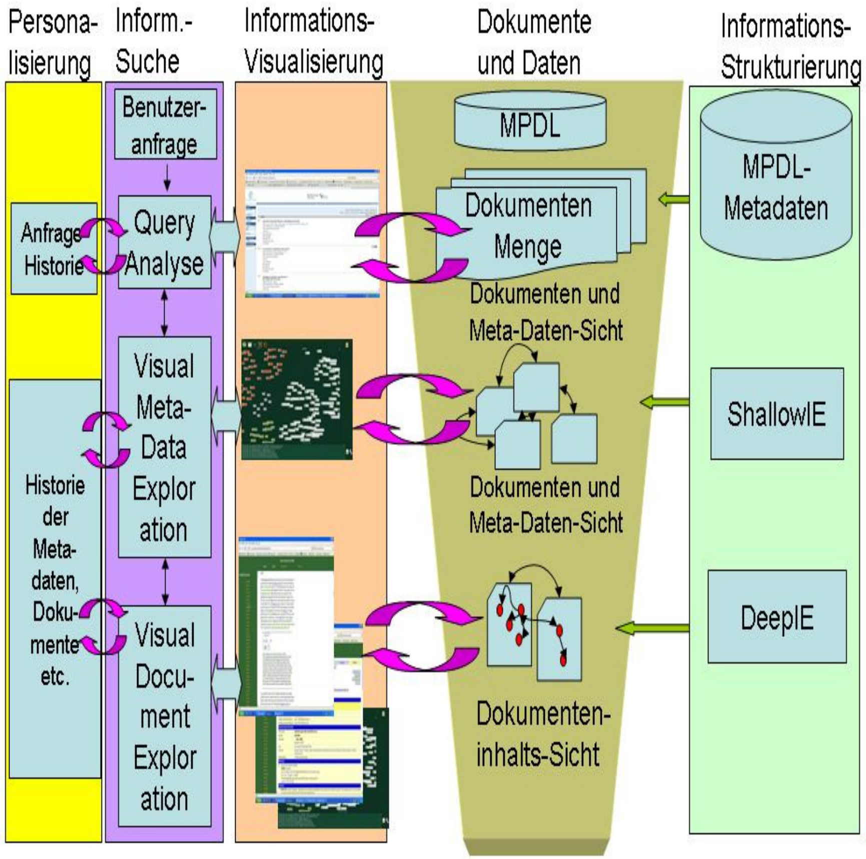
Question	Possible Answers	⊘
How did you like the introduction ?	1=useless/5=helpful	4,42
How useful is the system?	1=useless/5=helpful	4,14
Do you think you might use such a system in your daily work?	1=no/5=yes	4,14
How do you judge the computed information?	1=useless/5=very informative	3,71
How do you judge the speed of the system?	1=very slow/5=very fast	4,42
How do you judge the usability of the system?	1=very laborious/5=very comfortable	3,42
Is the graphical representation of the results useful?	1=totally not/5=very useful	3,57
Is the graphical representation appealing?	1=totally not/5=very appealing	3,71
Is the navigation useful in the system ?	1=totally not/5=very useful	3,57
Is the navigation intuitive in the system?	1=totally not/5=very intuitive	3,57
Did you have any problems using the system?	1=heavy/5=no difficulties	4,28

Results of the Evaluation of IDEXVisor

- All users were able to answer the questions
- The search speed was judged generally as „fast“
- Difficulties with the interaction: more complex interface than current search engines („Google“ syndrome)
  - Parts of the user interface were overlooked or actually not recognized
  - Difficulties to use different perspectives and to coordinate the results of different perspectives.
- Possible improvements:
  - More simple/consistent presentation
    ➔ trade-off between intuitiveness and features
  - Improved clustering through grouping by semantic similarity
    ➔ information has to exist in the database
  - Inclusion of synonyms in the search

DiLiA: Digital Library Assistant

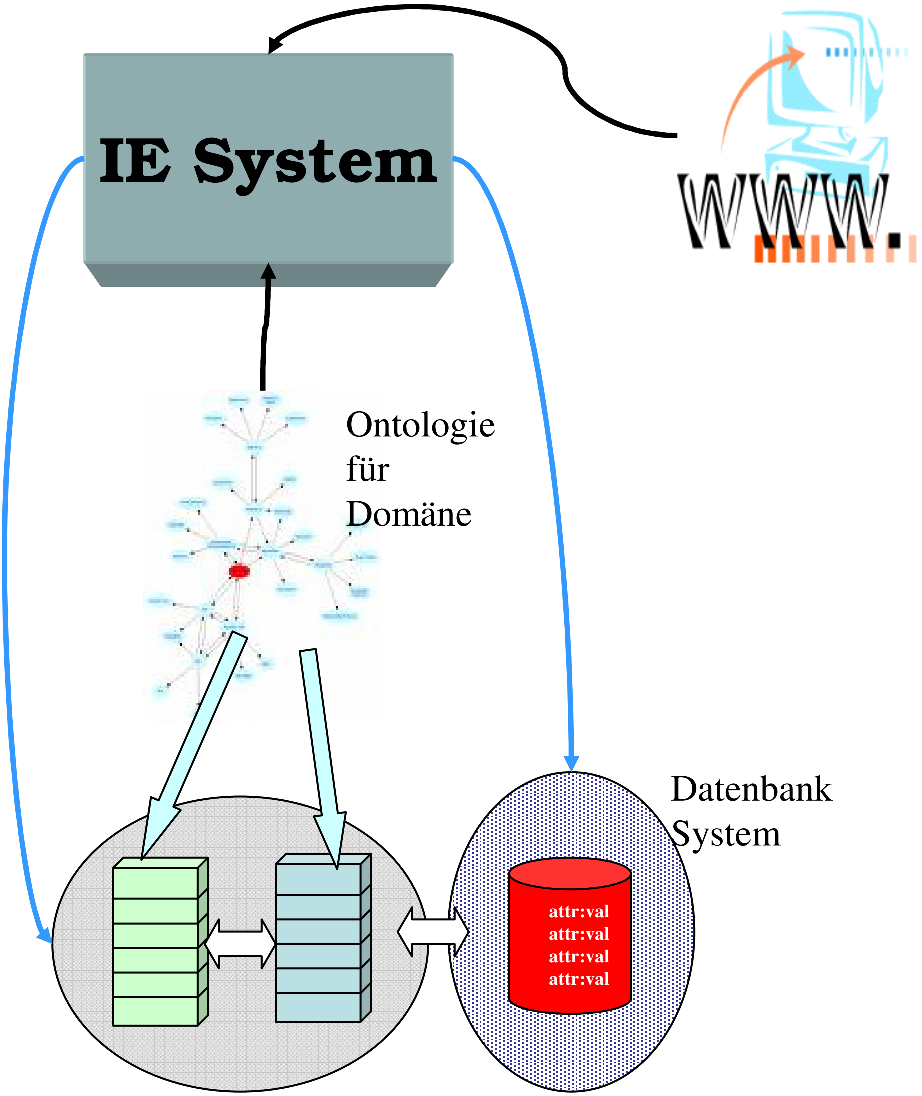
- Combining: IE and QA
- Search as zooming
- Shallow and deep IE
- BioIE as deep IE applications
  - Relation mining in biomedical texts
  - Integration/validation with existing Ontologies (UniProtKB)
- Partners
  - MP digital library
  - Semgine
  - Start: Jan 2008


Persona-lisierung
Inform.-Suche
Informations-Visualisierung
Dokumente und Daten
Informations-Strukturierung
Benutzer-anfrage
Anfrage Historie
Query Analyse
Visual Meta-Data Exploration
Visual Docu-ment Exploration
Historie der Meta-daten, Doku-mente etc.
MPDL
Dokumenten Menge
Dokumenten und Meta-Daten-Sicht
Dokumenten und Meta-Daten-Sicht
Dokumenten-inhalts-Sicht
MPDL-Metadaten
ShallowIE
DeepIE

©Neumann, DFKI

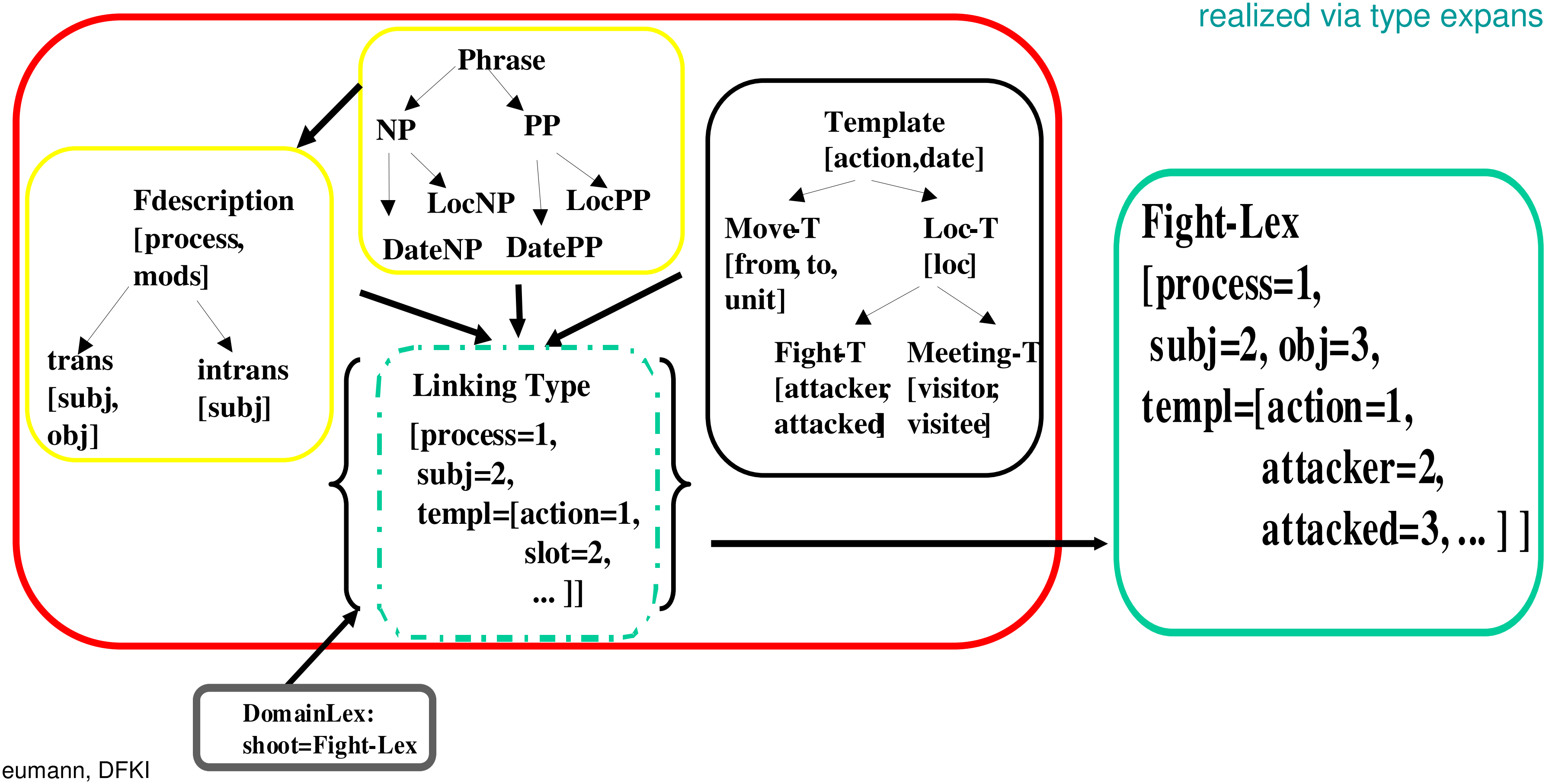
Ontologie-basierte Informationsextraktion

- Extraktion von relevanten Informationen aus textuellen Quellen (Web Seiten)
- Integration der extrahierten Daten mit der aktuellen Datenbank
- Domänen-Ontologie als Ausgangspunkt
  - Relevanz
  - Normalisierung
  - Abbildung



Domain modeling in DFKI system SMES is realised using typed feature structures

- Domain modeling via hierarchy of templates (black box), using the formalism TDL, which is also used to model hierarchies of linguistic objects (yellow boxes).
- The interface between domain knowledge and linguistic entities is specified via *linking types* (green box), which represent a close connection between concepts of the different layers, and which are accessible via the domain lexicon (brown & green box). Template-filling is then realized via type expansion.



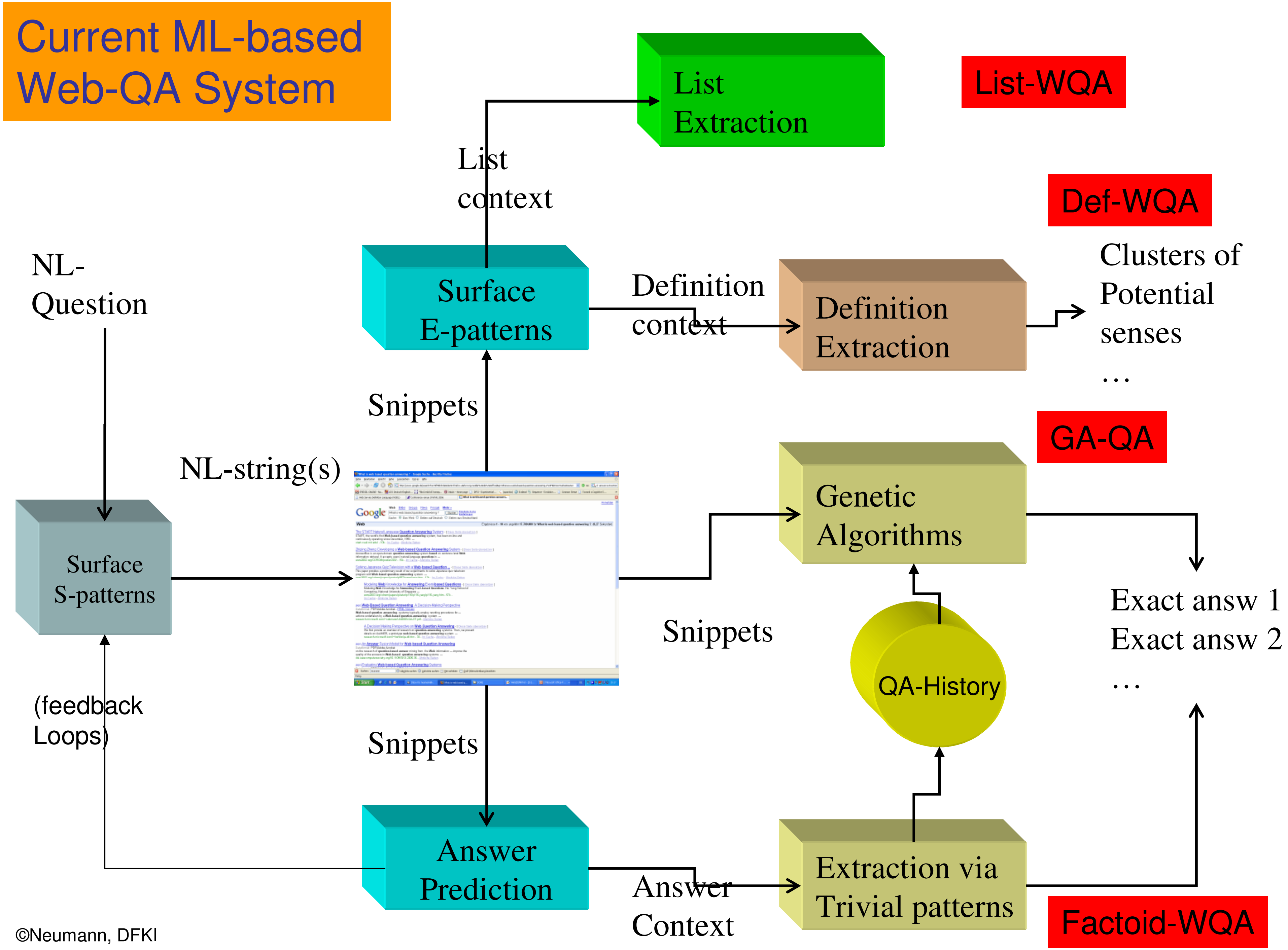
©Neumann, DFKI

HyLaP-QA: Machine Learning for web-based QA

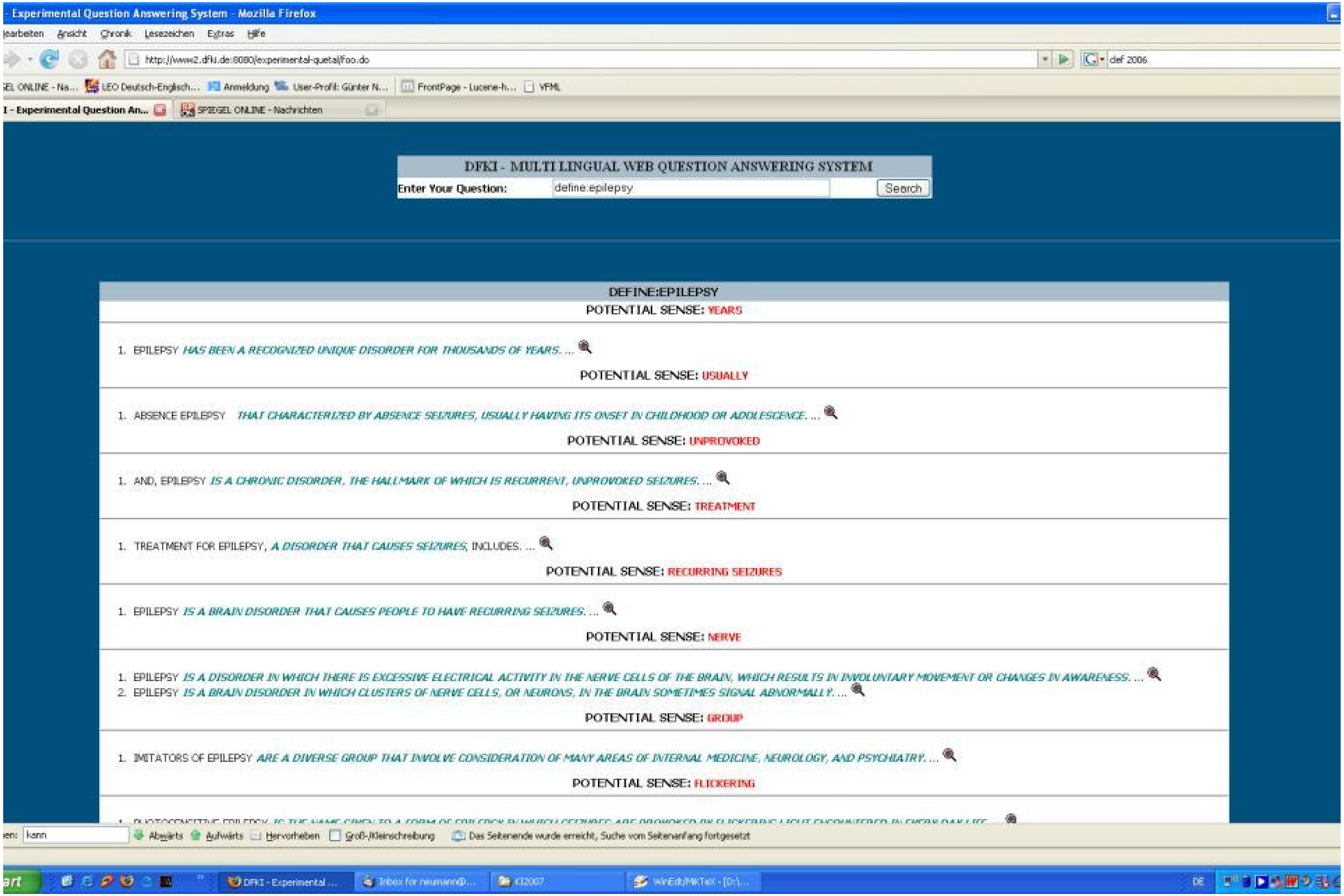
- Our goal:
  - Development of ML-based strategies for complete **end-to-end** question answering for different type of questions and the open domain.
- Our perspective:
  - Extract exact answers for different types of questions **only** from web snippets
  - Use strong data-driven strategies
  - Evaluate them with Trec/Clef Q-A pa
- Our current results:
  - ML-based strategies for open domain **factoid**, **definition** and **list** questions
  - Question type specific query expansion for controlling web search
  - **Unsupervised** learning for answer extraction
  - Promising performance (~ 0.5 MRR on Trec/Clef data)

F: When was Madonna born?
D: What is Ubuntu?
L: What movies did James Dean appear in?

Current ML-based Web-QA System



Example: What is epilepsy?


- Experimental Question Answering System - Mozilla Firefox
earbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www2.dfki.de:8080/experimental-quetal/foo.do
def 2006
EL ONLINE - Na... LEO Deutsch-Englisch... Anmeldung User-Profil: Günter N... FrontPage - Lucene-h... VFML
I - Experimental Question An... SPIEGEL ONLINE - Nachrichten
en: kann Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung
Das Seitenende wurde erreicht, Suche vom Seitenanfang fortgesetzt
art DFKI - Experimental ... Inbox for neumann@... KI2007 WinEdt/MiKTeX - [D:\...
DE

DFKI - MULTI LINGUAL WEB QUESTION ANSWERING SYSTEM

Enter Your Question: define:epilepsy Search

DEFINE:EPILEPSY

POTENTIAL SENSE: YEARS

1. EPILEPSY *HAS BEEN A RECOGNIZED UNIQUE DISORDER FOR THOUSANDS OF YEARS.* ...

POTENTIAL SENSE: USUALLY

1. ABSENCE EPILEPSY *THAT CHARACTERIZED BY ABSENCE SEIZURES, USUALLY HAVING ITS ONSET IN CHILDHOOD OR ADOLESCENCE.* ...

POTENTIAL SENSE: UNPROVOKED

1. AND, EPILEPSY *IS A CHRONIC DISORDER. THE HALLMARK OF WHICH IS RECURRENT, UNPROVOKED SEIZURES.* ...

POTENTIAL SENSE: TREATMENT

1. TREATMENT FOR EPILEPSY, *A DISORDER THAT CAUSES SEIZURES*, INCLUDES. ...

POTENTIAL SENSE: RECURRING SEIZURES

1. EPILEPSY *IS A BRAIN DISORDER THAT CAUSES PEOPLE TO HAVE RECURRING SEIZURES.* ...

POTENTIAL SENSE: NERVE

1. EPILEPSY *IS A DISORDER IN WHICH THERE IS EXCESSIVE ELECTRICAL ACTIVITY IN THE NERVE CELLS OF THE BRAIN, WHICH RESULTS IN INVOLUNTARY MOVEMENT OR CHANGES IN AWARENESS.* ...
2. EPILEPSY *IS A BRAIN DISORDER IN WHICH CLUSTERS OF NERVE CELLS, OR NEURONS, IN THE BRAIN SOMETIMES SIGNAL ABNORMALLY.* ...

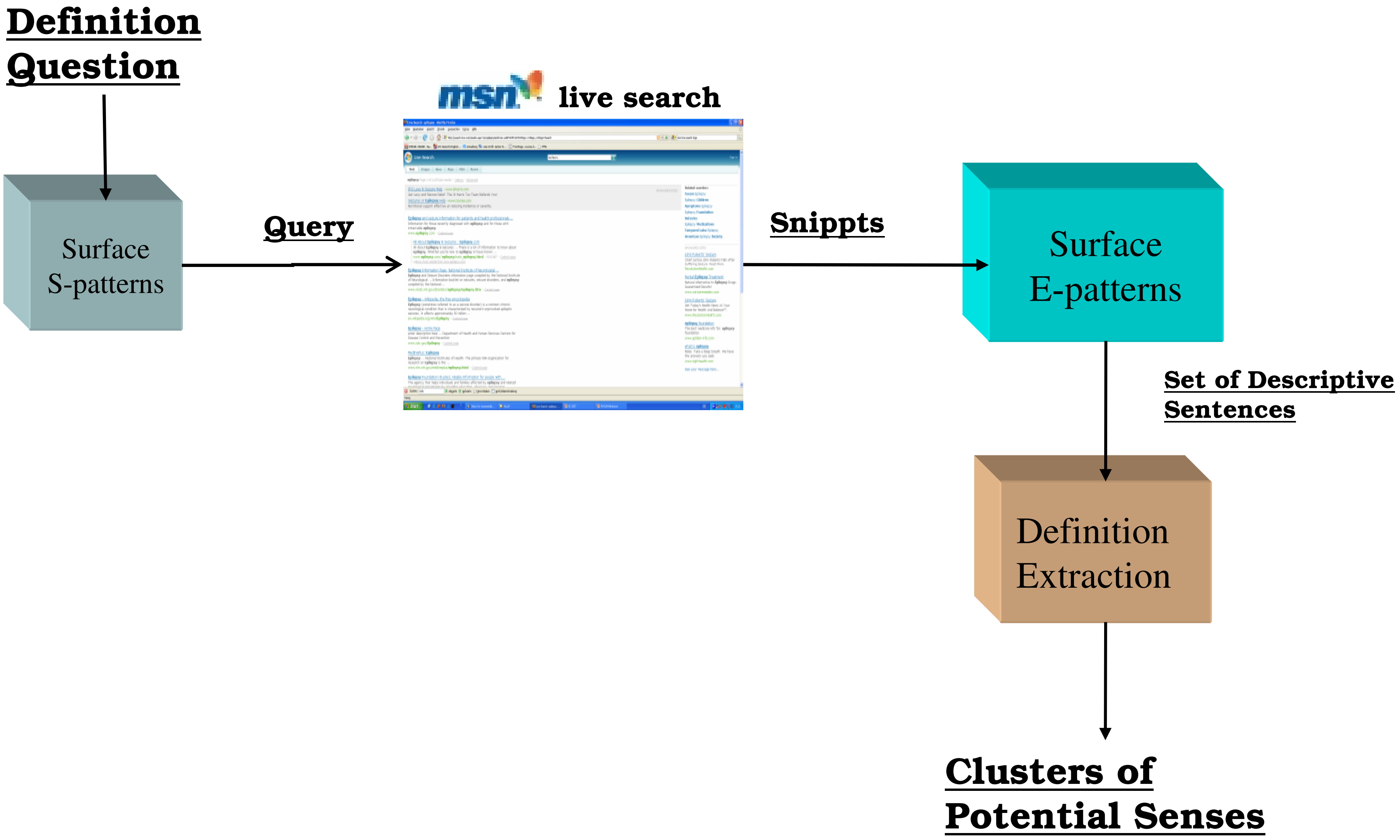
POTENTIAL SENSE: GROUP

1. IMITATORS OF EPILEPSY *ARE A DIVERSE GROUP THAT INVOLVE CONSIDERATION OF MANY AREAS OF INTERNAL MEDICINE, NEUROLOGY, AND PSYCHIATRY.* ...

POTENTIAL SENSE: FLICKERING

1. PHOTOSENSITIVE EPILEPSY *IS THE NAME GIVEN TO A FORM OF EPILEPSY IN WHICH SEIZURES ARE PROVOKED BY FLICKERING LIGHT ENCOUNTERED IN EVERY DAY LIFE.* ...

Language Independent Architecture



Definition Question

Surface S-patterns

Query

msn live search

Snippts

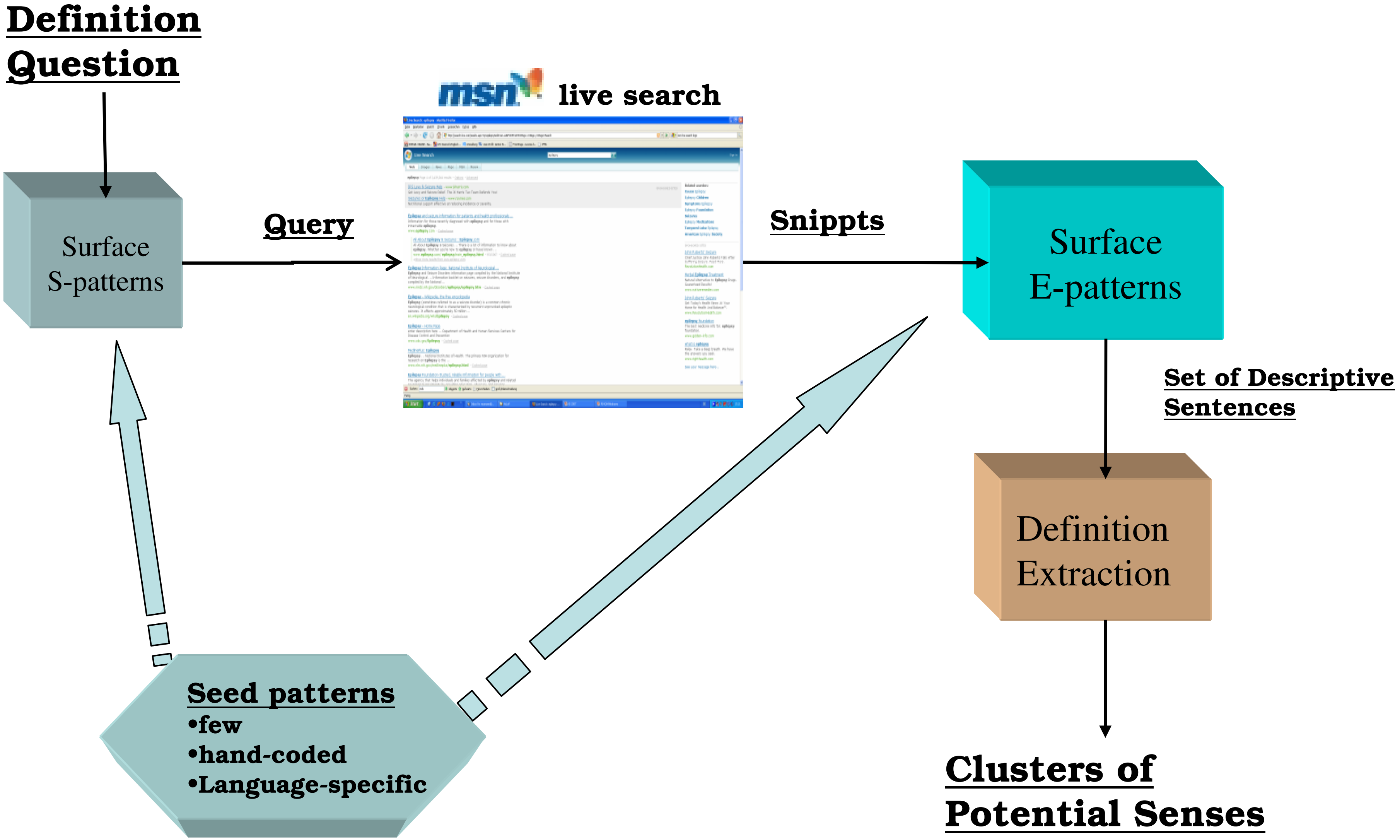
Surface E-patterns

Set of Descriptive Sentences

Definition Extraction

Clusters of Potential Senses

Language Independent Architecture



Definition Question

Surface S-patterns

Query

msn live search

Snippts

Surface E-patterns

Set of Descriptive Sentences

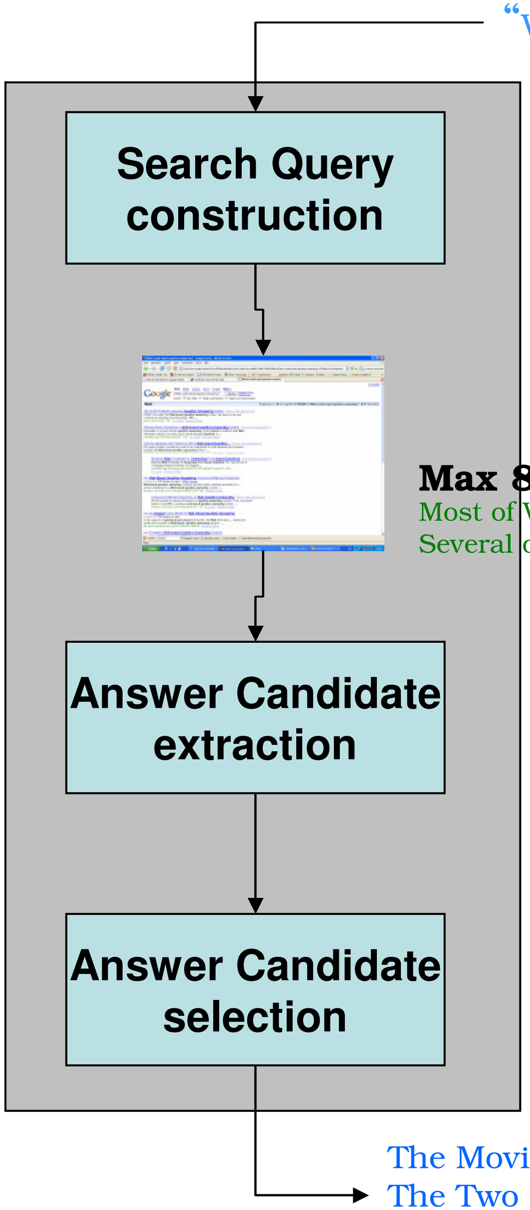
Definition Extraction

Clusters of Potential Senses

Seed patterns
- **few**
- **hand-coded**
- **Language-specific**

List-WQA – Overview

"What are 9 works written by Judith Wright?"



Search Query construction

Answer Candidate extraction

Answer Candidate selection

Qfocus → inbody
NPs → intitle
Apply 4 patterns Qi

Q1: (intitle:"*Judith Wright*") AND (inbody:"works" OR inbody:"written")

Max 80 snippets:

Most of Wright's poetry was **written** in the mountains of southern Queensland. …
Several of her early **works such as** 'Bullocky' and 'Woman to Man' became standard …

Apply 8 patterns πi (hyponym, possessive, copula, quoting, etc.)

π4: **entity** is \w+ **qfocus** \w*

*Chubby Hubby is …. Ben and Jerry's **ice cream** brand.*

Use Semantic kernel & Google N-grams

The Moving Image, Woman to Man, The Gateway, The Two Fires, Birds, The Other Half, City Sunrise, The Flame three and Shadow.

List-WQA – Results

- **Answer Selection**:
  - Two measures **Accuracy** and $\mathbf{F_1}$ score.
  - Two values
    - All questions
    - Only questions where at least one answer was found in the fetched snippets.
  - Duplicate answers have also an impact on the performance. For instance:
    - "*Maybelline*" (also found as "*Maybellene*" and "*Maybeline*").
    - John Updike's novel "*The Poorhouse Fair*" was also found as "*Poorhouse Fair*".

Systems\Trec	2001	2002	2003	2004
ListWebQA(F_1)	0.35/0.46	0.34/0.37	0.22/0.28	0.30/0.40
ListWebQA(Acc)	0.5/0.65	0.58/0.63	0.43/0.55	0.47/0.58
Top one(Acc.)	0.76	0.65	-	-
Top two(Acc.)	0.45	0.15	-	-
Top three(Acc.)	0.34	0.11	-	-
Top one(F_1)	-	-	0.396	0.622
Top two(F_1)	-	-	0.319	0.486
Top three(F_1)	-	-	0.134	0.258
Yang & Chua 04 (F_1)	-	-	.464 ~.469	-

We conclude:
Encouraging results, competes well with 2nd best;
Still creates too much noise;

Summary

- Dynamic, interactive IE
  - Expert and IE system together explore data pool
- Combining IR, QA and IE
- Highly scalable Language Technology needed
- Relation mining and clustering
- Prob. needs probabilistic reasoning