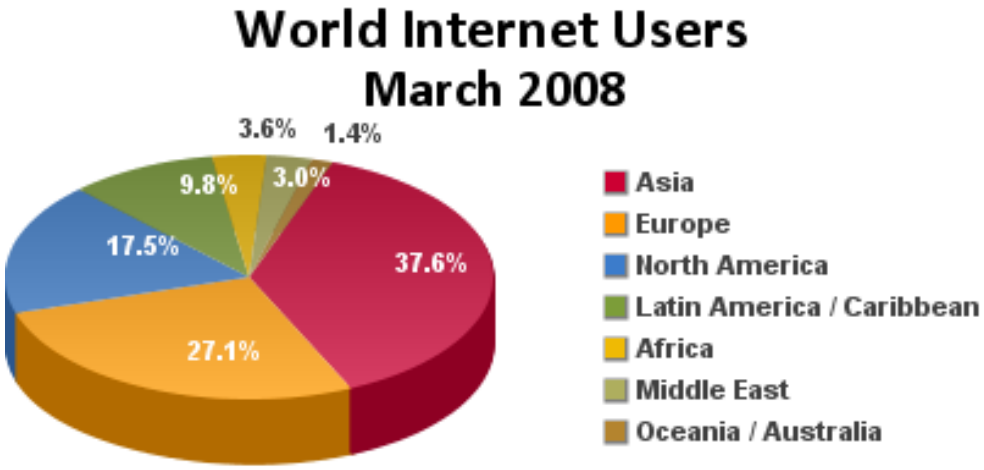
The Multilingual and Cross-lingual Web

PD Dr. Günter Neumann
LT lab
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
November, 2009

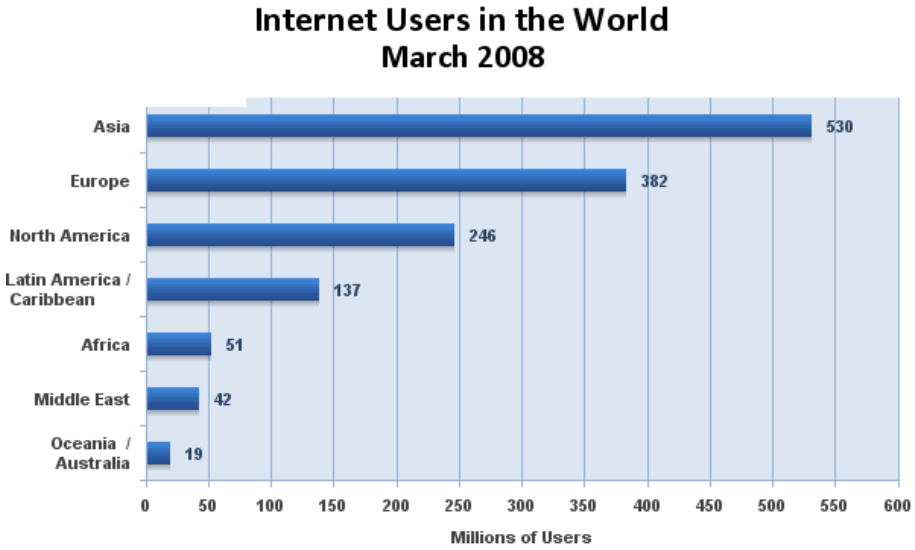
Outline

- Why Multilingual/crosslingual Web
- Key technologies
- HLT directions

Why Multilingual Web ?

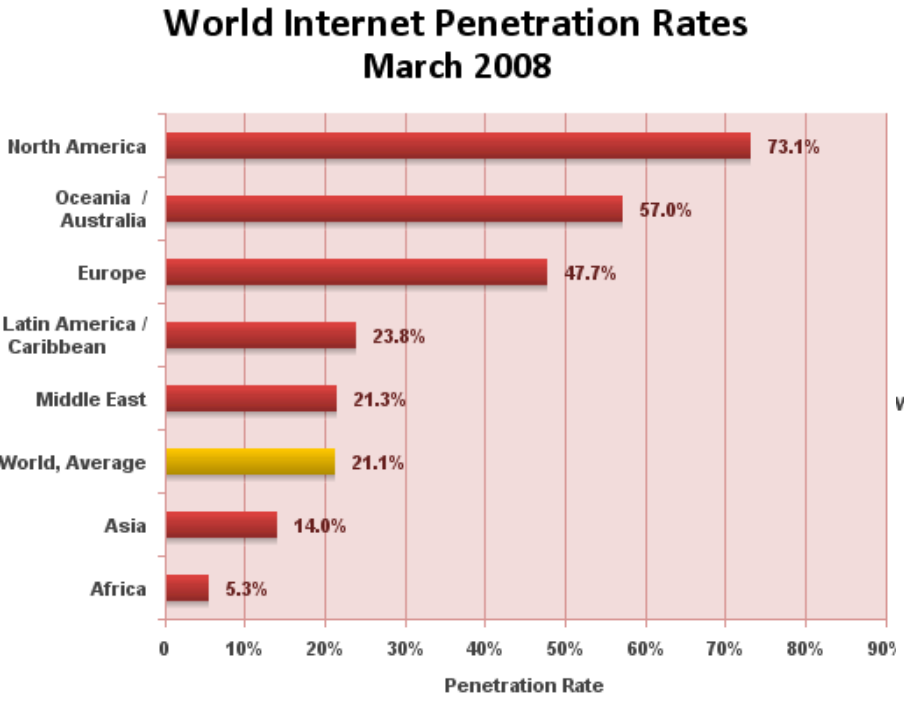


Source: www.internetworldstats.com/stats.htm
Copyright © 2008, Miniwatts Marketing Group

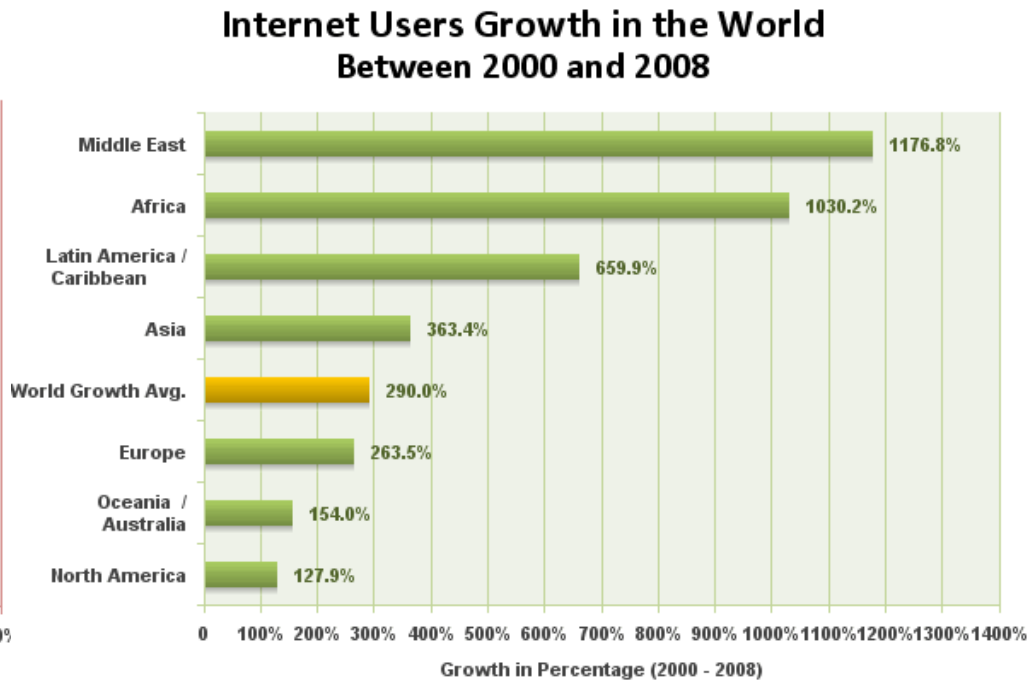


Note: World Internet Users estimate is 1,407,724,920 for Q1 2008
Copyright © 2008, Miniwatts Marketing Group - www.internetworldstats.com

The number of Internet Users is still growing



Note: Penetration Rates are based on a world population of 6,676,120,288 for mid-year 2008
Copyright © 2008, Miniwatts Marketing Group - www.internetworldstats.com



Note: World Internet Users estimate is 1,407,724,920 for Q1 2008.
Copyright © 2008, Miniwatts Marketing Group - www.internetworldstats.com

The Web is still evolving

What is Web 2.0 ?

A description from Tim O‘Reilly:

"Web 2.0 is the business revolution in the computer industry caused by the move to the **internet as platform**, and an attempt to understand the rules for success on that new platform. Chief among those rules is this: Build applications that harness **network effects** to **get better the more people use them**.“

Tim O'Reilly (2006-12-10). Web 2.0 Compact Definition: Trying Again

Tim Bernes-Lee:

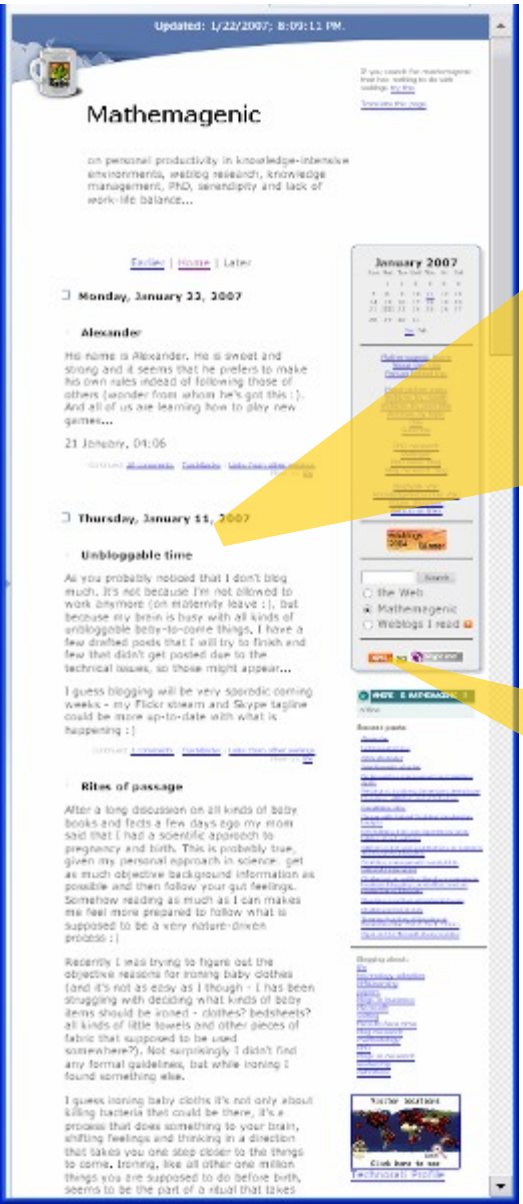
Web 1.0 was all about connecting people. It was an interactive space, and I think **Web 2.0 is of course a piece of jargon**, nobody even knows what it means. If Web 2.0 for you is blogs and wikis, then that is people to people. But that was what the Web was supposed to be all along.

developerWorks Interviews: Tim Berners-Lee (7-28-2006)

Key Web 2.0 services/applications

- Blogs
- Wikis
- Tagging and social bookmarking
- Multimedia sharing
- RSS and syndication
- Podcasting
- P2P

Anatomy of a Blog



Thursday, January 11, 2007

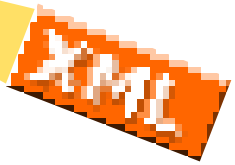
Unbloggable time

As you probably noticed that I don't blog much. It's not because I'm not allowed to work anymore (on maternity leave :), but because my brain is busy with all kinds of unbloggable baby-to-come things. I have a few drafted posts that I will try to finish and few that didn't get posted due to the technical issues, so those might appear...

I guess blogging will be very sporadic coming weeks - my Flickr stream and Skype tagline could be more up-to-date with what is happening :)

Continued: 1 comments | TrackBacks | Links from other weblogs
More on: life

Continued: 1 comments | TrackBacks | Links from other weblogs
More on: life



Add your comment:

Name:

Email:

Web site:

Comment:

WIKIPEDIA

English
The Free Encyclopedia
2 117 000+ articles

Deutsch
Die freie Enzyklopädie
674 000+ Artikel

Français
L'encyclopédie libre
591 000+ articles

Polski
Wolna encyklopedia
448 000+ haseł

日本語
フリー百科事典
444 000+ 記事

Nederlands
De vrije encyclopedie
384 000+ artikelen

Italiano
L'enciclopedia libera
381 000+ voci

Português
A enciclopédia livre
343 000+ artigos

Español
La enciclopedia libre
306 000+ artículos

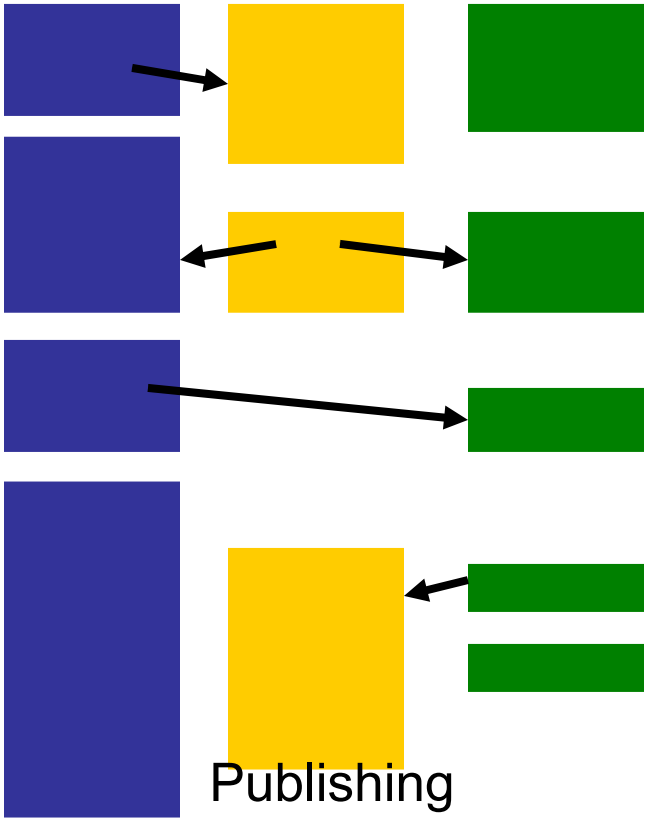
Svenska
Den fria encyklopedin
264 000+ artiklar

search • suche • rechercher • szukaj • 検索 • zoeken
ricerca • busca • buscar • sök • поиск • 搜索 • søk • haku • suk

English ›

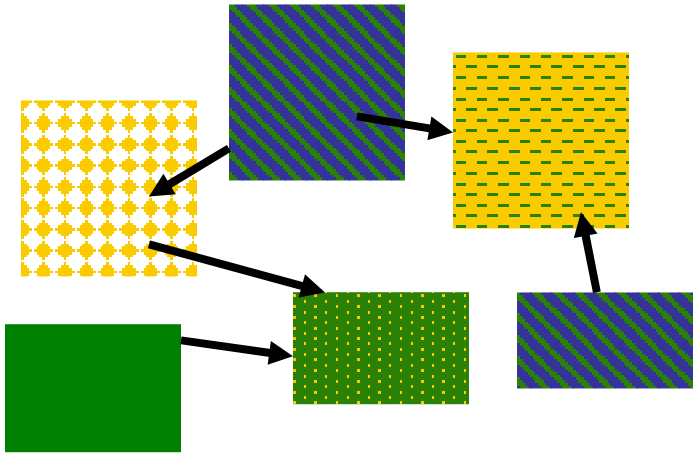
Blogs versus Wikis

Blogs
„Collective Thinking,
individual writing“



Publishing

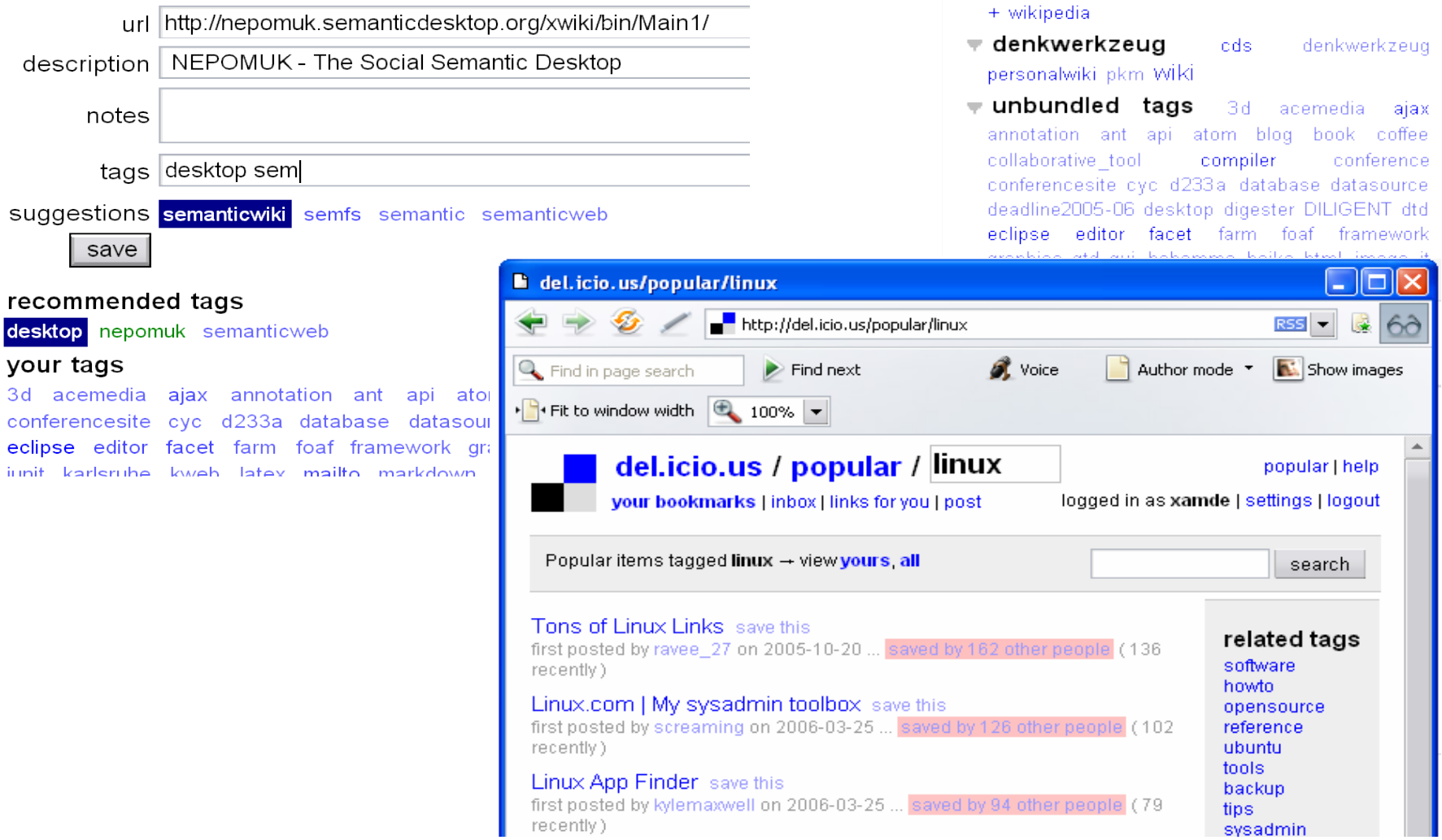
Wikis
„ Collective Thinking,
collective writing“



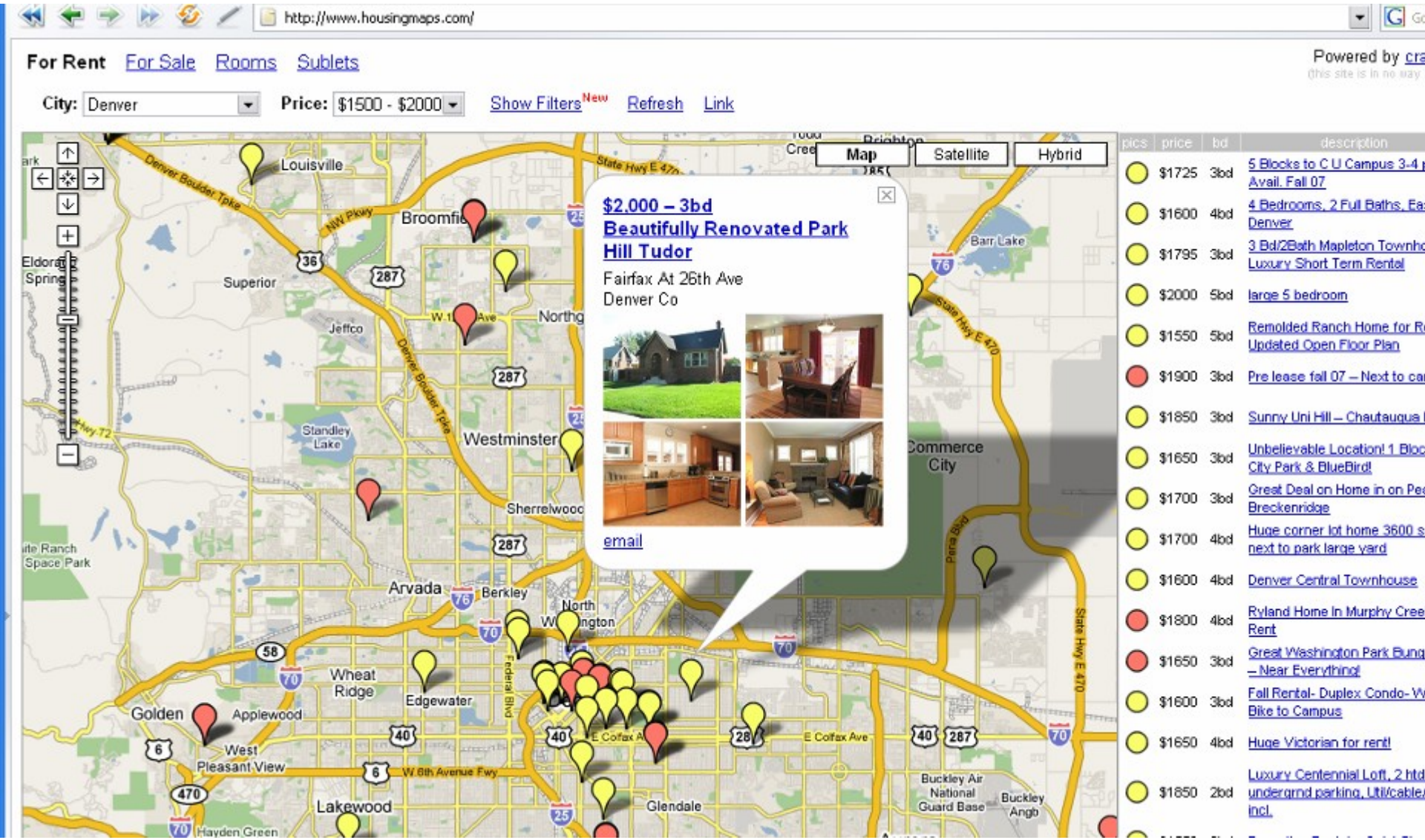
Organising

Social bookmarking

is a web-based service to share Internet bookmarks.

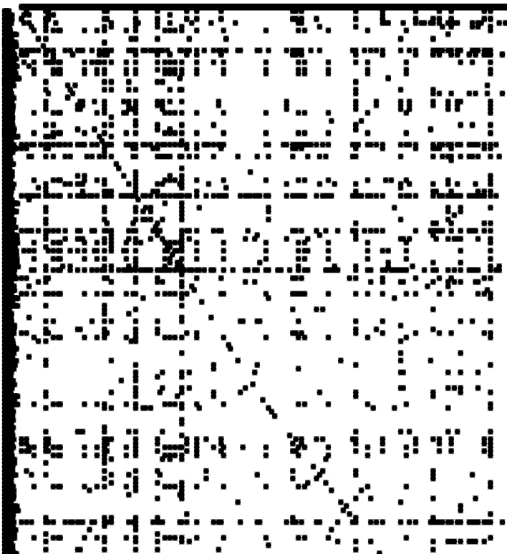


Mash-Up: Example

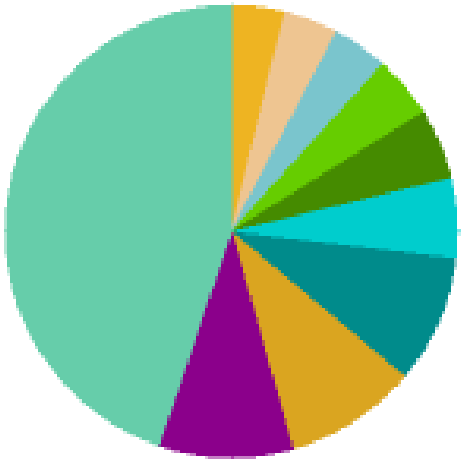


Mash-Ups

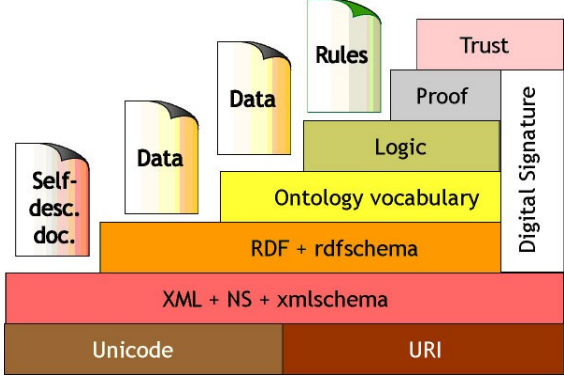
- „From two (web pages) make one“
  - Craigs List: Google Maps & real estate ads
- Programmableweb.com: 755 web-APIs



» Amazon
» Delicious
» Flickr
» Google
» GoogleMaps
» Technorati
» Yahoo
» YouTube



Semantic Web



- Idea: Web pages which are enriched with machine readable annotations
  - Search using **unique concepts** than ambiguous keywords
  - **Structural search** instead of bag of kewyowds
    - Ex: <*, *located_in, Europe*> instead of „`located in Europe`“
  - **Inference** finds implict knowledge
    - Ex: <Karlsruhe, located_in, Germany> and
<Germany, located_in, Europe>
→ <Karlsruhe, located_in, Europe>

- State of the art:
  - Exchange formats RDF, OWL are W3C-Standards (HTML, CSS, XML)
  - RDF & OWL Tools incl. inference exist

- Trend:
  - Information extraction is being considered as a basic functionality for automatically enriching/learning ontologies from Web sources
  - Question Answering as a means for semantic search and answer extraction

Semantic Web + Web 2.0 = Web 3.0?

	Web 2.0	Web 3.0
Tagging	• Annotation with mit **ambiguous** keywords • Singular/Plural-problem • Synonyms • No inference	• annotation with **unique** keywords • inference (tag „dog“ deduces tag „animal“)
Recombinaton of data from different sources	• Mesh-Ups manually programmed in advance	• Dynamic tagging through end user (cf. Piggybank)
Search	• Keyword search or tag-based search *finds* documents	• Structural search combines data and *creates* documents
Time horizon	• 2004 - 2007	• 2007 – 2010

Summary: The Web Changes in Several Dimensions

- Semantics
- Dynamics
- Heterogeneity
- Collaboration
- Composition
- Socialization
- Mobility

- Increasing demands on HLT technology
- Cross-lingual and multilingual HLT in order to further drive evolution of the Web

Key technological areas – Information Retrieval Perspective

- **Cross-lingual information retrieval**: enables users to enter queries in languages they are fluent in, and uses language translation methods to retrieve documents originally written in other languages.
- **Cross-lingual question answering**: Find precise answers in documents of one language for a complete Natural Language question formulated in another language.

Knowledge Extraction Perspective

- **Cross-lingual information extraction**: The extraction and merging of relevant facts from Web documents from different languages.
- **Cross-lingual ontology population:** The acquisition of domain specific ontologies automatically from Web sources of different languages. This will also help to share and exchange content expressed in different countries and languages.

Semantic Web Perspective

- **Cross-lingual services:** The technology behind the Web2.0 has made it easily possible to create regional specific service providers almost everywhere and for almost anything, be it business, cultural, public or administrative. With the increasing mobility of citizens and the emergence of the Mobile Web, we can expect that users of different languages will have direct access to such regional specific information services.
- **Cross-lingual service composition:** The integration of diverse local services data into larger, globally operating services or chains of services provided through automatic service composition with user interfaces in different languages (e.g., travel agencies, online market places, Internet television).

Web 2.0 Perspective

- **Cross-lingual wikis**: In Wikipedia, for example, there are several articles written in several languages on the same topic, but contents are different by languages. By comparing these differences among languages, we can find various viewpoints of the same topic.
- **Cross-lingual blogosphere**: Find differences of concerns and opinions about a topic in blogs of different countries and languages. It is useful not only for mutual understanding, but also for the analysis of social and political problems.

Current Research Activities

- Information Retrieval on Blogs
  - NTCIR-7 CLIRB (Cross-Lingual Information Retrieval for Blog)
- Question Answering on Blogs
  - TREC 2007 QA Track
- Question Answering on Wikipedia
  - QA@CLEF 2007
- CLEF 2006 WiQA
  - given a Wikipedia page, locate information snippets in Wikipedia
- CoNLL challenges on multilingual dependency parsing, 2006, 2007
- ACE (Automatic Content Extraction)
  - Multilingual Named Entity Extraction and Relation Extraction
- PASCAL Ontology Learning Challenge
  - Ontology construction
  - Ontology extension
  - Ontology population
  - Concept naming

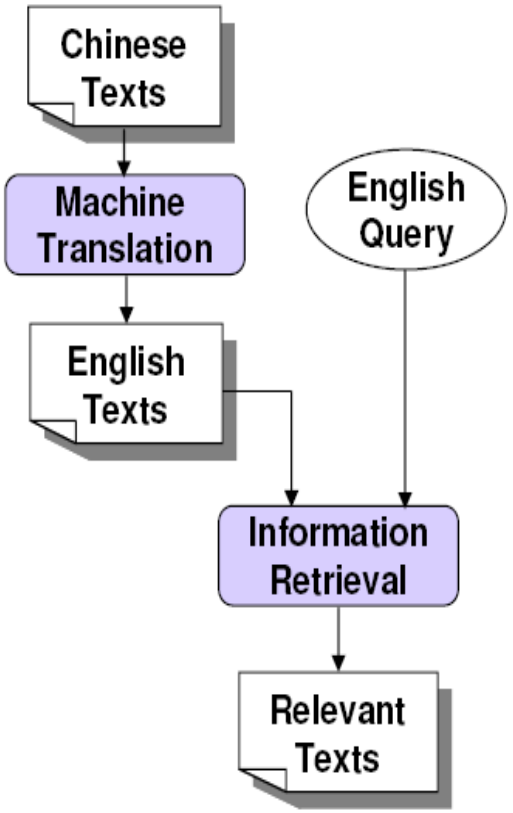
Human Language Technology

- Core applications
  - Cross-lingual Document Retrieval
  - Multilingual IE
  - Multilingual QA
  - …
- Core Technologies
  - Language resources
    - Grammars, lexicon
    - Corpora
    - …
  - Technologies
    - Machine Learning
    - Multilingual Parsing
    - Machine Translation
    - …

CLDR: Crosslingual Document Retrieval

- A baseline MT based approach ala Dilek Hakkani-Tür (ICSI, Berkeley) & Heng Ji and Ralph Grishman (NYU), 2007

Baseline CLDR

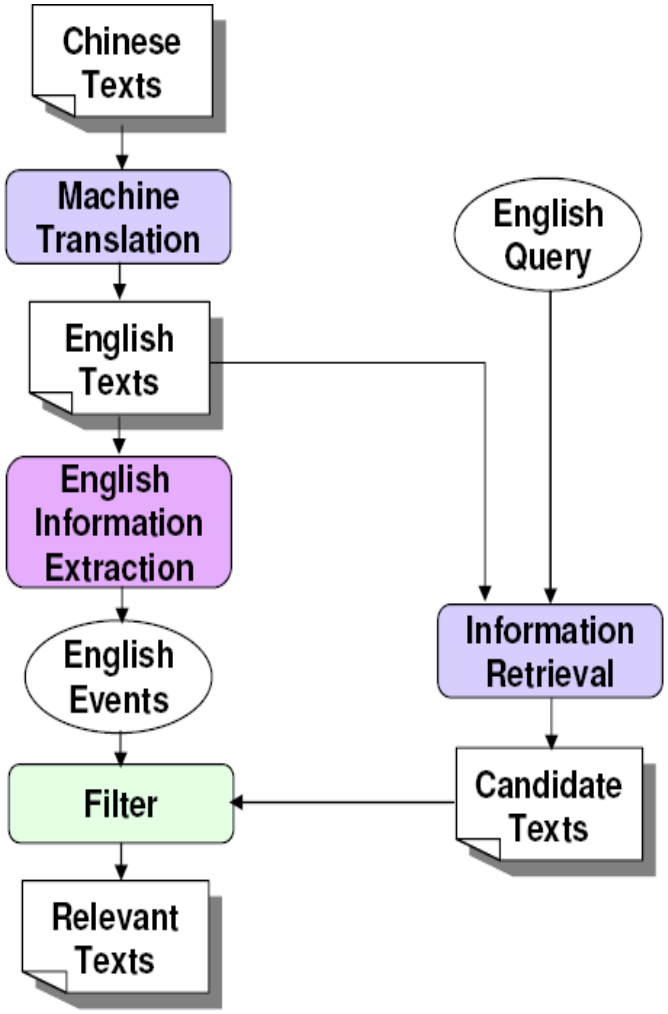


- Problem: High Recall but Low Precision

Baseline CLDR + IE

Motivation:
Events in a IR query overlap
With event types from IE (ACE)

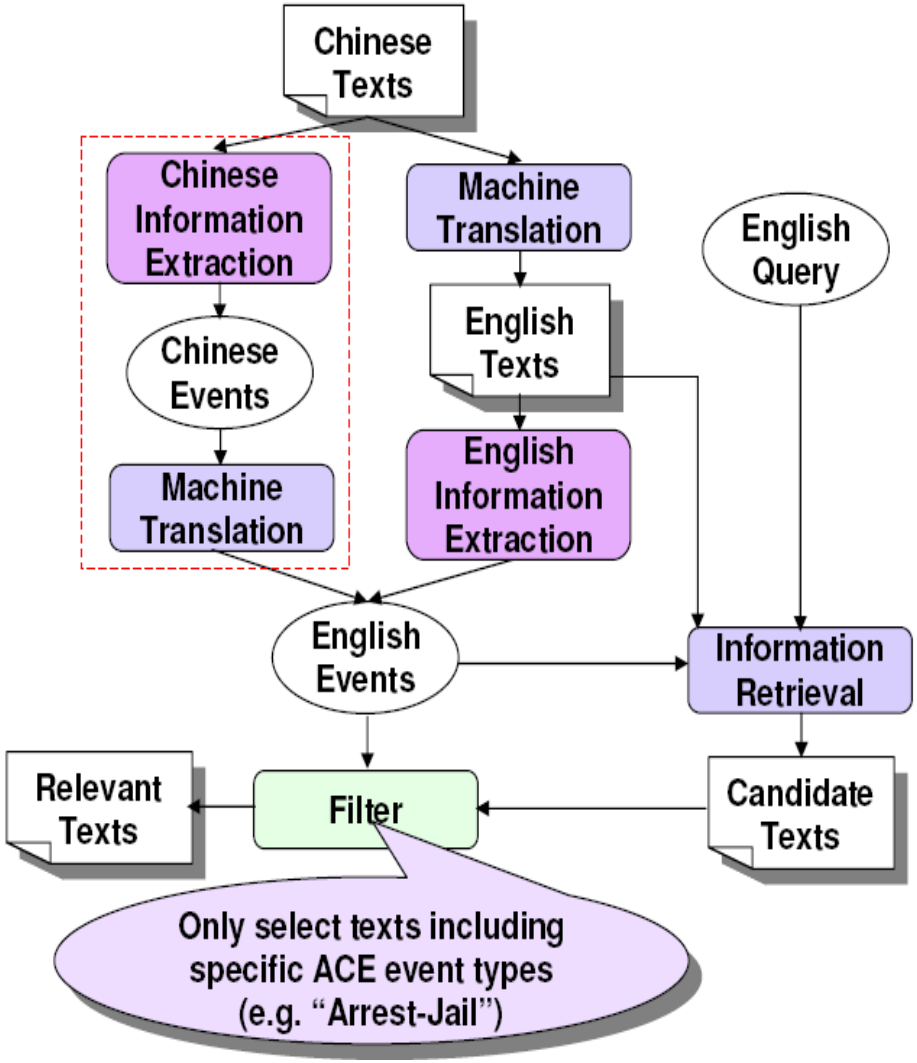
ACE event type	*Example*
Life/Die	**Kurt Schork** died in **Sierra Leone yesterday**
Transaction/Transfer	**GM** sold **the company** in **Nov 1998** to **LLC**
Movement/Transport	**Homeless people** have been moved to **schools**
Business/Start-Org	**Schweitzer** founded **a hospital** in **1913**
Conflict/Attack	the attack on **Gaza** killed **13 people**
Contact/Meet	**Arafat's cabinet** met for **4 hours**
Personnel/Start-Position	**Cornell Medical Center** recruited **12 nursing students**
Justice/Arrest	**Zawahiri** was arrested in **Iran**



- **Problem: Significantly Improves precision but with noticeable loss in recall**
- **Reason: Events were missed by Chinese-to-English machine translation**

Major problem:
Events might be lost by MT

Solution: Use Chinese IE to Find more Events



- RWTH Chinese-to-English Machine Translation (Zens and Ney, 2004)
  - Statistical, phrase-based system
  - Computes best translation using a weighted log-linear combination of various statistical models
  - The model scaling factors are optimized on development corpus with respect to BLEU score (Och, 2003)
- University of Massachusetts INDRI Baseline Cross-lingual Document Retrieval (Strohman et al., 2005)
  - Combines language modeling and inference network
  - Task independent, no emphasis on event information
- NYU Information Extraction (Grishman et al., 2005)
  - English system combines pattern matching with statistical models
  - Chinese system is based on semi-automatically extracted pattern matching

IE for semantic annotation

Identification of IE-sub-tasks:

- named 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

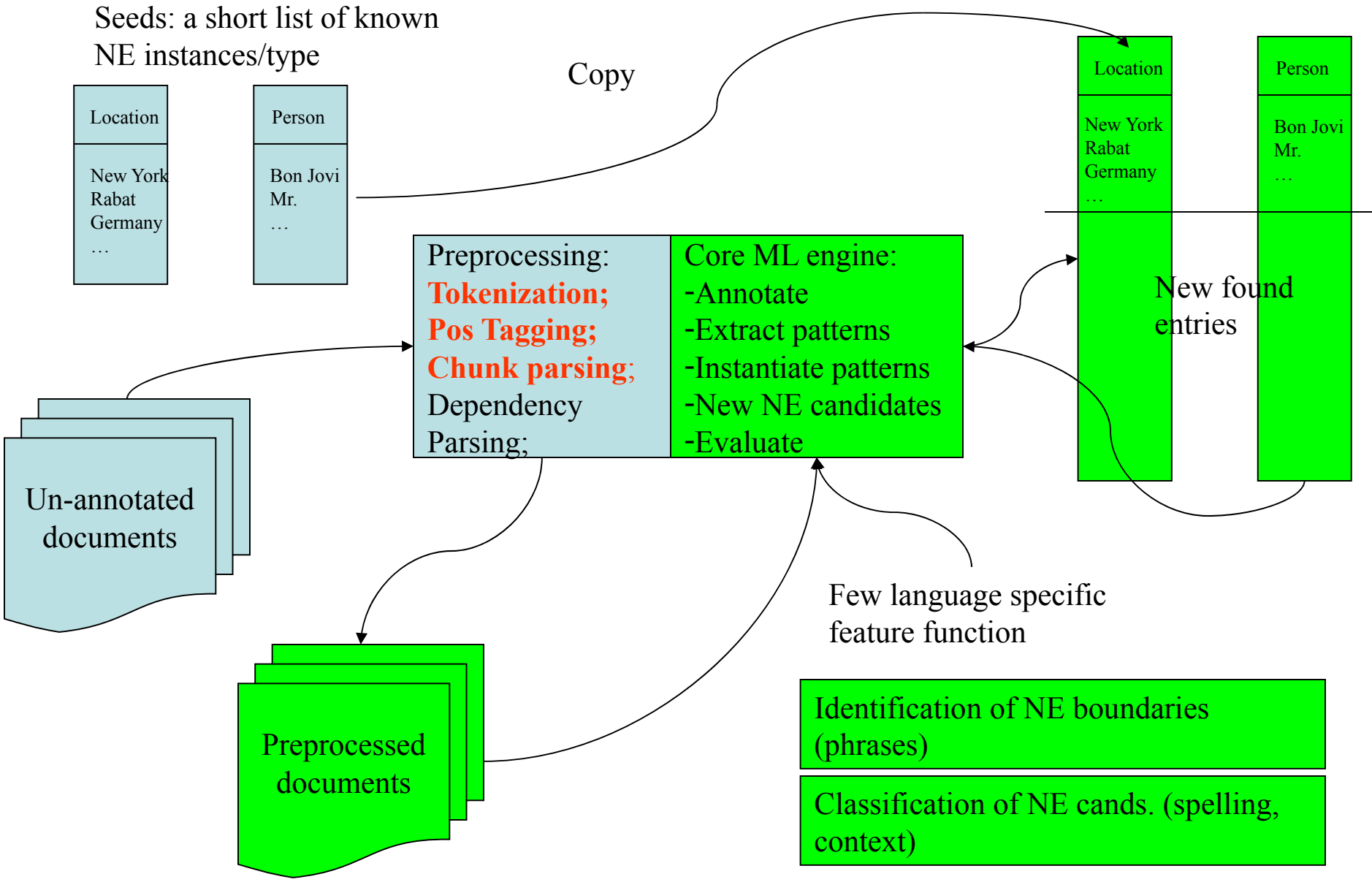
- identification
- discovery
- validation
- evaluation

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

Multilingual Information Extraction

- Relevance of NER/RE
  - NEs are major types of relation arguments
    - Born_in(Person,Location)
  - NER/RE important for a number of other applications, e.g., QA, ontology learning, semantic search
    - Where was Wolfgang Amadeus Mozart born ?
- Machine Learning (ML) approaches are dominating
  - Language independent processing
  - Language dependent feature engineering
- Particular promising: seed-based ML
  - RELFEX: a recent approach for multilingual NER and transliteration for 50 languages, cf. Sproat et al. 2005
  - Recent approaches for seed-based relation extraction

Seed-based Machine Learning: NER



Motivation for Seed Rules

"The only supervision is in the form of 7 seed rules (namely, that *New York, California* and *U.S. are* locations; that any name containing Mr. is a person; that any name containing *Incorporated* is an organization; and that *I.B.M.* and *Microsoft* are organizations)."

[Collins and Singer, 1999]

Seed Rules: Thai

- Something including and to the right of นาย is likely to be a person
  Something including and to the right of นาง is likely to be a person
  Something including and to the right of นางสาว is likely to be a person
  Something including and to the right of น.ส. is likely to be a person
  Something including and to the right of คุณ is likely to be a person
  Something including and to the right of เด็กหญิง is likely to be a person
  Something including and to the right of ด.ญ. is likely to be a person
- Something including and to the right of พ.ต.อ. is likely to be a person
  Something including and to the right of พล.ต.ต. is likely to be a person
  Something including and to the right of พล.ต.ท. is likely to be a person
  Something including and to the right of พล.ต.อ. is likely to be a person
  Something including and to the right of ส.ส. is likely to be a person
- ทักษิณ ชินวัตร is a person
  ทักษิณ is likely a person
  ชวน หลีกภัย is a person
  บรรหาร ศิลปอาชา is a person

Seed Rules: Persian

- Lexicon TITLE
  ياقآ
  رتكد
  مناخ
  بانج
  ونياب
  سدنهم
- Lexicon OrgDesc
  يرادناتسا
  ترازو
  تلود
  ميژر
  يرادرهش
  نمجنا

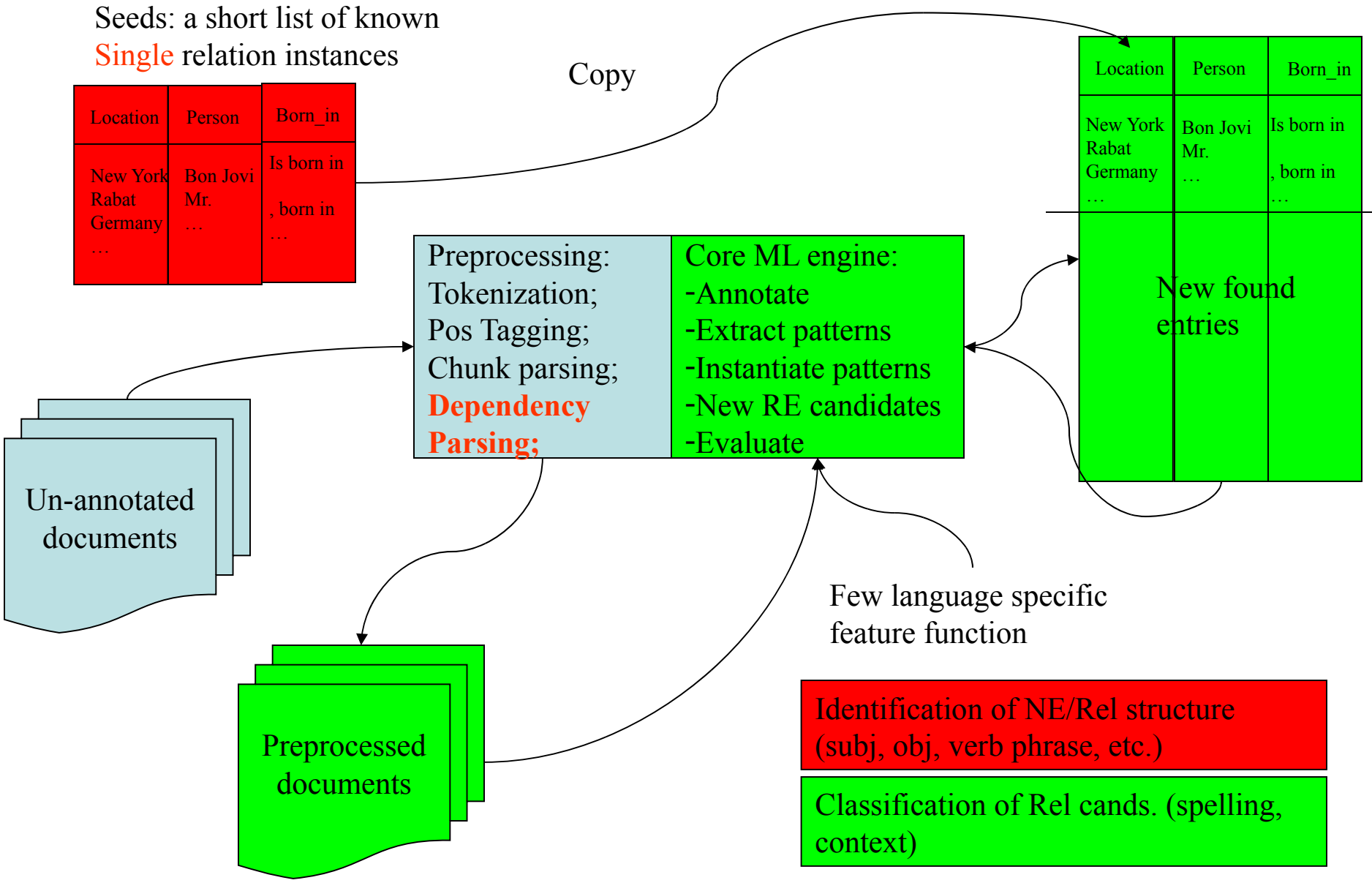
- Lexicon POSITION
  روهمج سيئر
  يروهمج سييري
  تندیزرپ
  تاملپيد

- Descriptors for named entities
  Lexicon PerDesc
  قباس
  هدنيآ
  Lexicon CityDesc
  رهش
  كرهش
  تختياپ
  Lexicon CountryDesc
  روشک

Seed rules for German (DFKI System BiQueNER)

- <rule contains="**Bush**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r001**" id="**r0**"> <type ne-type="**PERSON**" /> </rule>
- <rule contains="**Mitterrand**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r002**" id="**r1**"> <type ne-type="**PERSON**" /> </rule>
- <rule contains="**Kohl**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r003**" id="**r2**"> <type ne-type="**PERSON**" /> </rule>
- <rule contains="**Berlin**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**101**" id="**r3**„> <type ne-type="**LOCATION**" /> </rule>
- <rule contains="**Deutschland**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r102**" id="**r4**"> <type ne-type="**LOCATION**" /> </rule>
- <rule contains="**Frankreich**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r103**" id="**r5**"> <type ne-type="**LOCATION**" /> </rule>
- <rule contains="**Lufthansa**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r201**" id="**r6**"> <type ne-type="**ORGANIZATION**" /> </rule>
- <rule contains="**Karstadt**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r202**" id="**r7**"> <type ne-type="**ORGANIZATION**" /> </rule>
- <rule contains="**CDU**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r203**" id="**r8**"> <type ne-type="**ORGANIZATION**" /> </rule>
- <rule contains="**Sonntag**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r401**" id="**r9**"> <type ne-type="**DATE**" /> </rule>
- <rule contains="**Juni**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r402**" id="**r10**"> <type ne-type="**DATE**" /> </rule>
- <rule contains="**Uhr**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r501**" id="**r11**"> <type ne-type="**TIME**" /> </rule>
- <rule contains="**vormittags**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r402**" id="**r12**"> <type ne-type="**TIME**" /> </rule>
- <rule contains="**nachmittags**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r403**" id="**r13**"> <type ne-type="**TIME**" /> </rule>
- <rule contains="**Euro**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r601**" id="**r14**"> <type ne-type="**MONEY**" /> </rule>
- <rule contains="**Dollar**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r602**" id="**r15**"> <type ne-type="**MONEY**" /> </rule>
- <rule contains="**Prozent**" nonalpha="" weight="**1.0**" count1="**0**" count2="**0**" seed-id="**r701**" id="**r16**"> <type ne-type="**PERCENTAGE**" /> </rule>

Seed-based Machine Learning: Relation Extraction

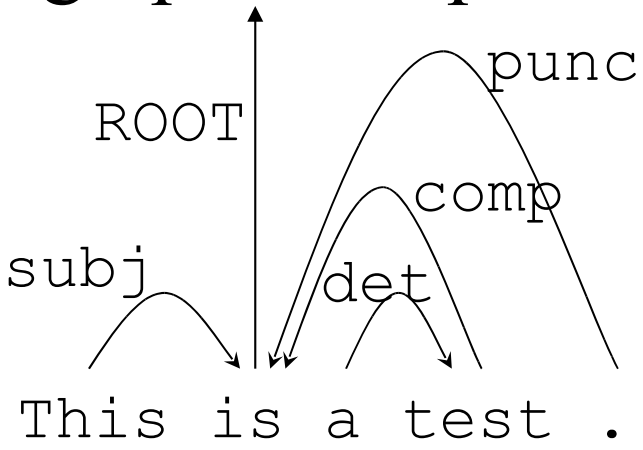


Summary: MLIE

- Seed-based approaches are promising basis for MLIE
  - No annotated corpora are needed
  - Small sets of seed examples are sufficient
  - Few language specific features
- BUT:
  - the richer the information to be extracted should be, the more complex the preprocessing has to be
- We need sufficiently deep & accurate multilingual HLT

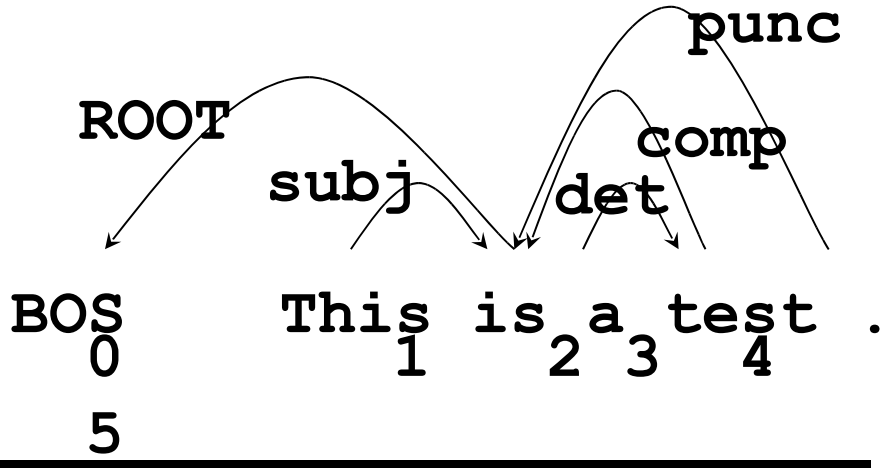
Multilingual Dependency Parsing

- No constituents (unlike phrase structure)
- Dependency relations between two lexical items (tokens)
- One possible graphical representation:



CoNLL shared tasks on multilingual depdency parsing (DP)

- Goal: evaluate current **data-driven** approaches for DP using standard representation for many languages
- Data: dependency tree banks
- Parsing means: compute HEAD & DEPREL (i.e., learn statistical models)



ID	FORM	LEMMA	CPOS TAG	POS TAG	FEATS	HEAD	DEPREL
1	This	this	pronoun	demon	sg	2	subj
2	is	be	v	v-fin	3|sg|pres	0	ROOT
3	a	a	art	art	indef	4	det
4	test	test	n	nc	sg	2	comp
5	.	.	punc	punc	_	2	punc

Treebanks used in CoNLL 2006

- Czech: Prague Dependency Treebank (PDT)
- Arabic: Prague Arabic Dependency Treebank (PADT)
- Slovene: Slovene Dependency Treebank (SDT)
- Danish: Danish Dependency Treebank (DDT)
- Swedish: Talbanken05
- Turkish: Metu-Sabancı treebank
- German: TIGER treebank
- Japanese: Japanese Verbmobil treebank
- Portuguese: The Bosque part of the Floresta sintá(c)tica
- Dutch: Alpino treebank
- Chinese: Sinica treebank
- Spanish: Cast3LB
- Bulgarian: BulTreeBank

Depen-dency format

Consti-tuents and functions

Constituents and some functions

Example for Arabic PADP Treebank

1 اِتِّفاقٌ_Ait~ifAqN اِتِّفاق_Ait~ifAq N N case=1|def=I 0 ExD _ _
2 بَيْنَ_bayona بَيْنَ_bayona P P _ 1 AuxP _ _
3 لُبْنانِ_lubonAni لُبْنان_lubonAn Z Z case=2|def=R 4 Atr _ _
4 وَ_wa وَ_wa C C _ 2 Coord _ _
5 سُورِيَّةٍ_suwriy~apK سُوري_suwriyA Z Z gen=F|num=S|case=2|def=I 4 Atr _ _
6 عَلَى_EalaY عَلَى_EalaY P P _ 1 AuxP _ _
7 رَفْعِ_rafoEi رَفْع_rafoE N N case=2|def=R 6 Atr _ _
8 مُسْتَوَى_musotawaY مُسْتَوَى_musotawaY N N _ 7 Atr _ _
9 التَّابُلِ_AltabAduli تَابُل_tabAdul N N case=2|def=D 8 Atr _ _
10 التِّجارِيِّ_AltijAriy~i تِجارِيّ_tijAriy~ A A case=2|def=D 9 Atr _ _
11 إِلَى_<ilaY إِلَى_<ilaY P P _ 7 AuxP _ _
12 500_500 500_500 Q Q _ 11 Atr _ _
13 مِلْيُونِ_miloyuwni مِلْيُون_miloyuwn N N case=2|def=R 12 Atr _ _
14 دُولارٍ_duwlArK دُولار_duwlAr N N case=2|def=I 13 Atr _ _

Results for CoNLL 2006

	Ar	Ch	Cz	Da	Du	Ge	Ja	Po	Sl	Sp	Sw	Tu	Tot	SD	Bu
McD	66.9	85.9	80.2	84.8	79.2	87.3	90.7	86.8	73.4	82.3	82.6	63.2	80.3	8.4	87.6
Niv	66.7	86.9	78.4	84.8	78.6	85.8	91.7	87.6	70.3	81.3	84.6	65.7	80.2	8.5	87.4
O'N	66.7	86.7	76.6	82.8	77.5	85.4	90.6	84.7	71.1	79.8	81.8	57.5	78.4	9.4	85.2
Rie	66.7	90.0	67.4	83.6	78.6	86.2	90.5	84.4	71.2	77.4	80.7	58.6	77.9	10.1	0.0
Sag	62.7	84.7	75.2	81.6	76.6	84.9	90.4	86.0	69.1	77.7	82.0	63.2	77.8	9.0	0.0
Che	65.2	84.3	76.2	81.7	71.8	84.1	89.9	85.1	71.4	80.5	81.1	61.2	77.7	8.7	86.3
Cor	63.5	79.9	74.5	81.7	71.4	83.5	90.0	84.6	72.4	80.4	79.7	61.7	76.9	8.5	83.4
…															
Av	59.9	78.3	67.2	78.3	70.7	78.6	85.9	80.6	65.2	73.5	76.4	56.0			80.0
SD	6.5	8.8	8.9	5.5	6.7	7.5	7.1	5.8	6.8	8.4	6.5	7.7			6.3

Labeled accuracy score: correct dependency relation (HEAD) and type (DEPREL) between words

Crosslingual Question Answering

Find exact answers written in any language

– Using NL questions expressed in a single language

Cross Language QA

- Similar task as TREC QA but with Questions and documents in different languages.
- Open domain: no restrictions of topic or domain of possible questions (question can be about anything)
- CLEF: European initiative
  - Multiple Languages QA
    - 2003 preliminary task
    - 2004, 2005, 2006, 2007
- NTCIR: Asian initiative
  - Question Answering Challenge:
    - NTCIR 3 (QAC1 Oct 2001-Oct 2002)
    - NTCIR 4 (QAC2 Apr 2003 – June 2004)
    - NTCIR 5 (QAC3 Nov 2004 – June 2005)

Multilingual QA Track at Clef

	2003	2004	2005	2006	2007
Target languages	3	7	8	9	10
Collections	News 1994		+News 1995		+Wikipedia Nov. 2006
Type of questions	200 Factoid		+ temporal restrictions + Definitions	-Type of questions + Lists	+ Linked questions + Closed lists
Supporting information	Doc.	Doc.	Doc.	Snippet	Snippet
Pilots and exercises		-Temporal restrictions - Lists		-AVE - RealTime - WiQA	- AVE - QAST

Clef 2006: 200 Questions

- FACTOID (150): loc, mea, org, oth, per, tim
- DEFINITION (40): per, org, object, oth
  - Person: Who is Josef Paul Kleihues?
  - Object: What is a router?
  - Other: What is a tsunami?
- LIST (10): "Name works by Tolstoy."
- Temporally restricted (40): by date, by period, by event
- NIL questions (without known answer in the collection
- Input format: question type (F, D, L) not indicated

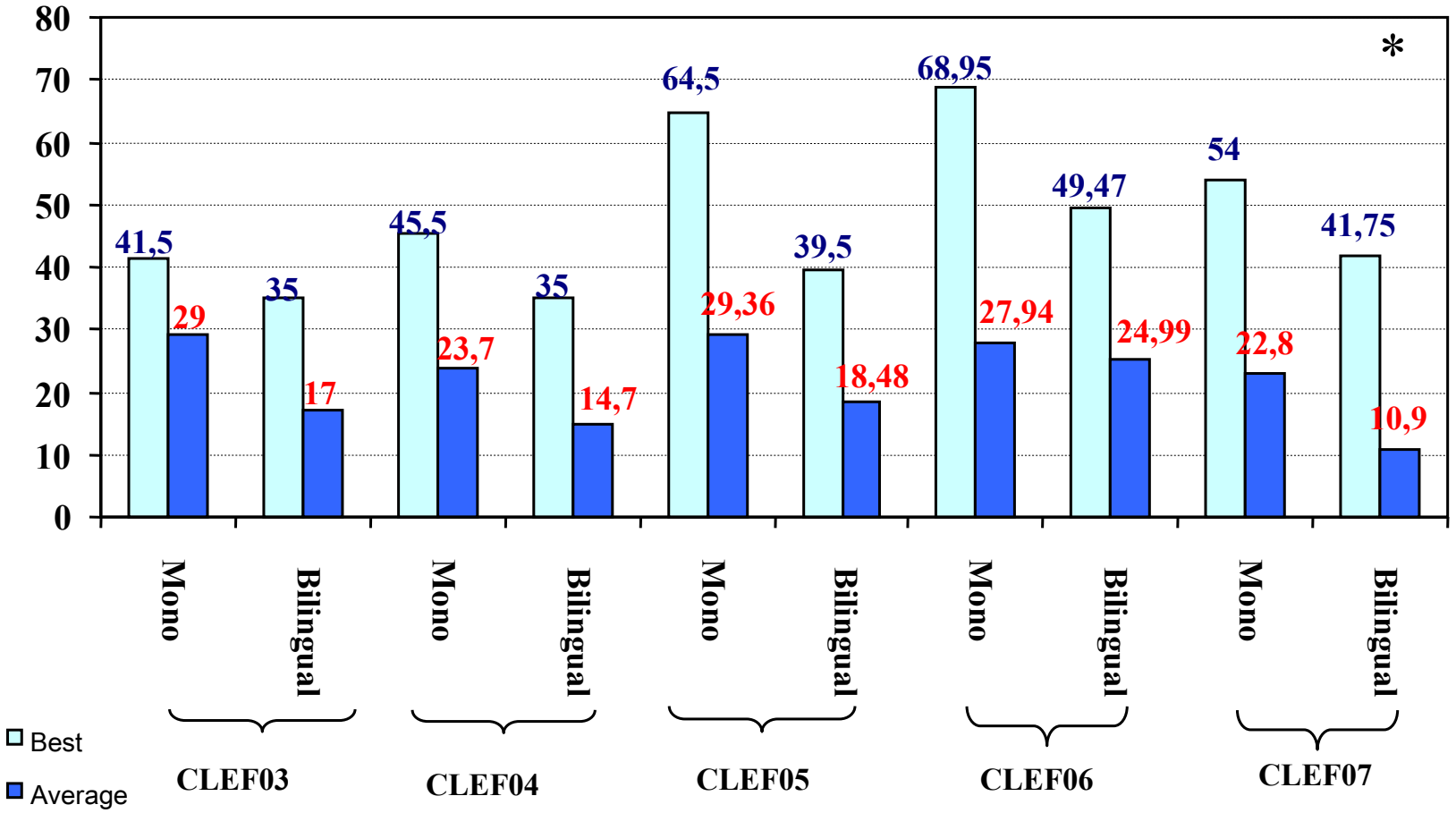
Clef 2007: Clef 2006 plus

- Closed lists:
  - Who were the components of the Beatles?
  - Who were the **last three** presidents of Italy?
- Linked questions
  - Topic: Otto von Bismarck
    - Who was called the “Iron-Chancellor”?
    - When was he born?
    - Who was his first wife?
  - Topics
    - Person or Event
    - Not provided to participants
    - Only a portion of the questions (from 15% depending on the languages)

Run format

- Clef 2006:
  - Multiple answers: from one to ten *exact* answers per question
  - *exact* = neither more nor less than the information required
  - each answer has to be supported by
    - docid
    - one to ten text snippets justifying the answer (substrings of the specified document giving the actual context)
- Clef 2007:
  - News articles
  - Wikipedia dump from November 2006 (→ caused critical decrase of performance)

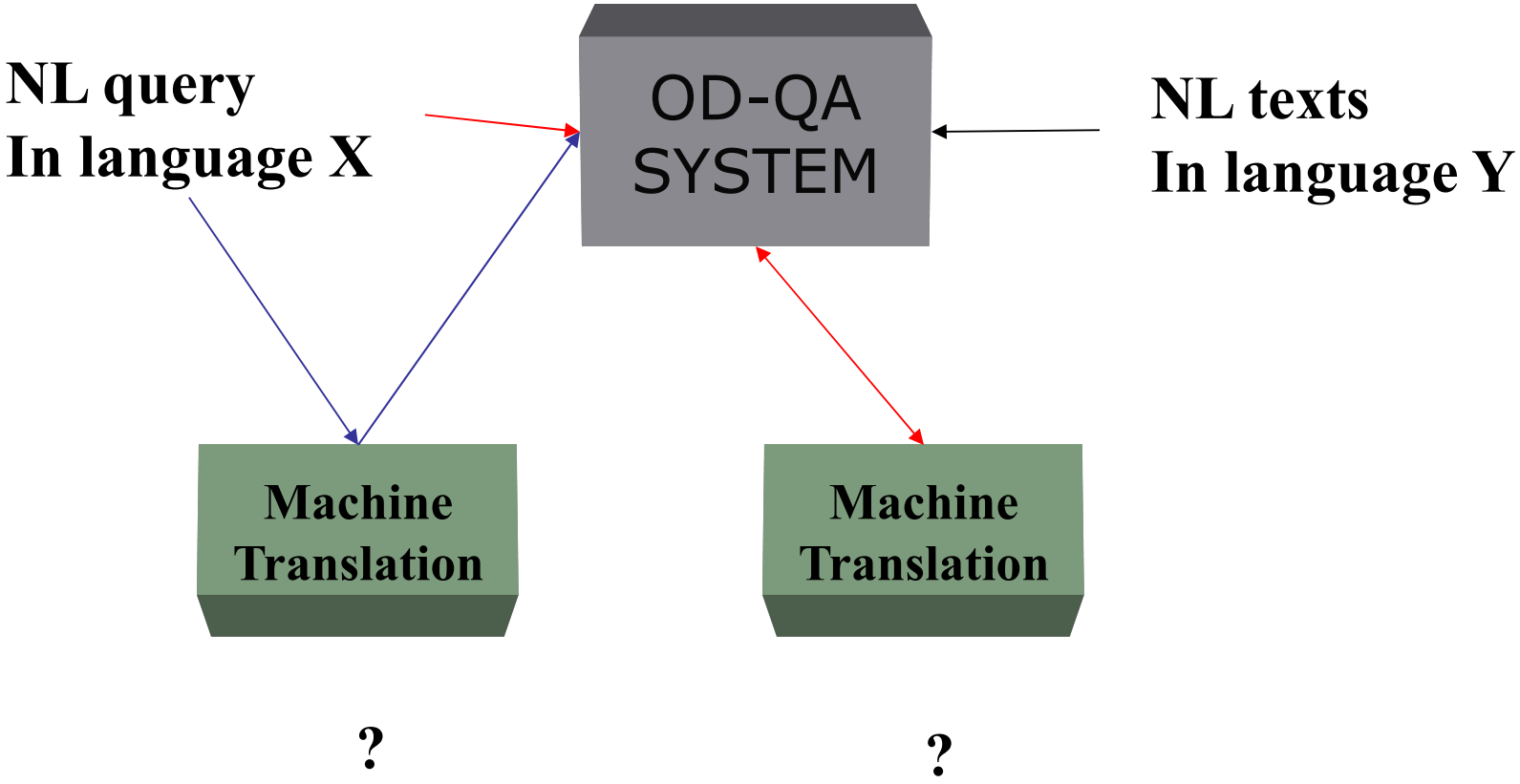
Results: Best and Average scores



Lower results in 2007

- Some answers only in Wikipedia
- Closed lists
  - Almost no answers
- Temporal restrictions
  - Still very difficult
- Linked questions
  - Topic not provided
  - Fail the first, fail the rest
  - Co-reference resolution

Cross-Lingual ODQA - Approaches



Approaches in CL QA

Two main different approaches used in Cross-Language QA systems:

Before Method

1

translation of the question into the target language (i.e. in the language of the document collection)

question processing

answer extraction

After Method

2

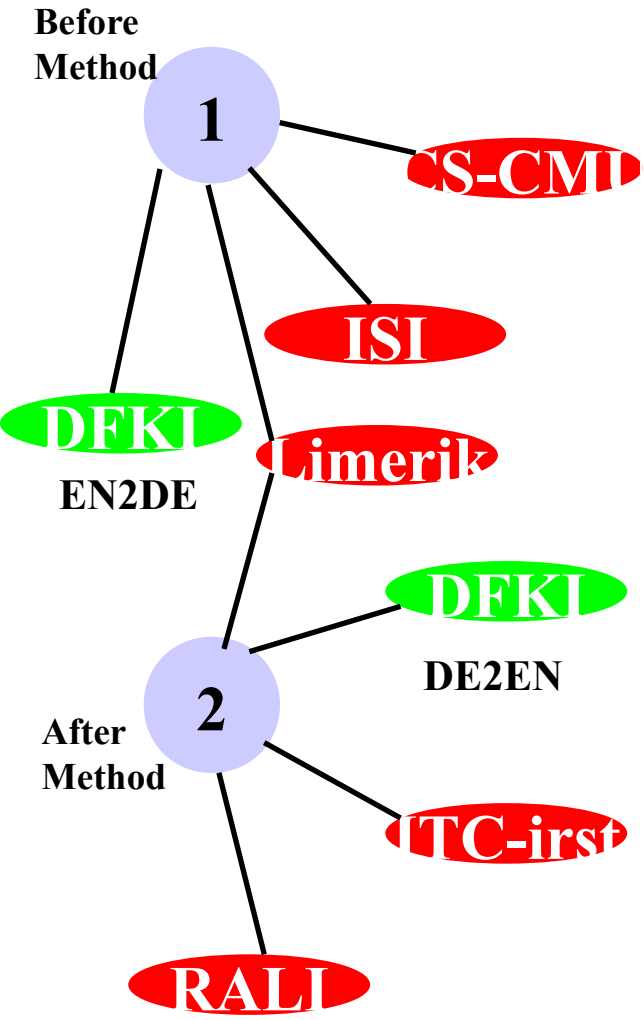
question processing in the source language to retrieve information (such as keywords, question focus, expected answer type, etc.)

translation and expansion of the retrieved data

answer extraction

Approaches in CL QA

Two main different approaches used in Cross-Language QA systems:



translation of the question into the target language (i.e. in the language of the document collection)

question processing

answer extraction

question processing in the source language to retrieve information (such as keywords, question focus, expected answer type, etc.)

translation and expansion of the retrieved data

answer extraction

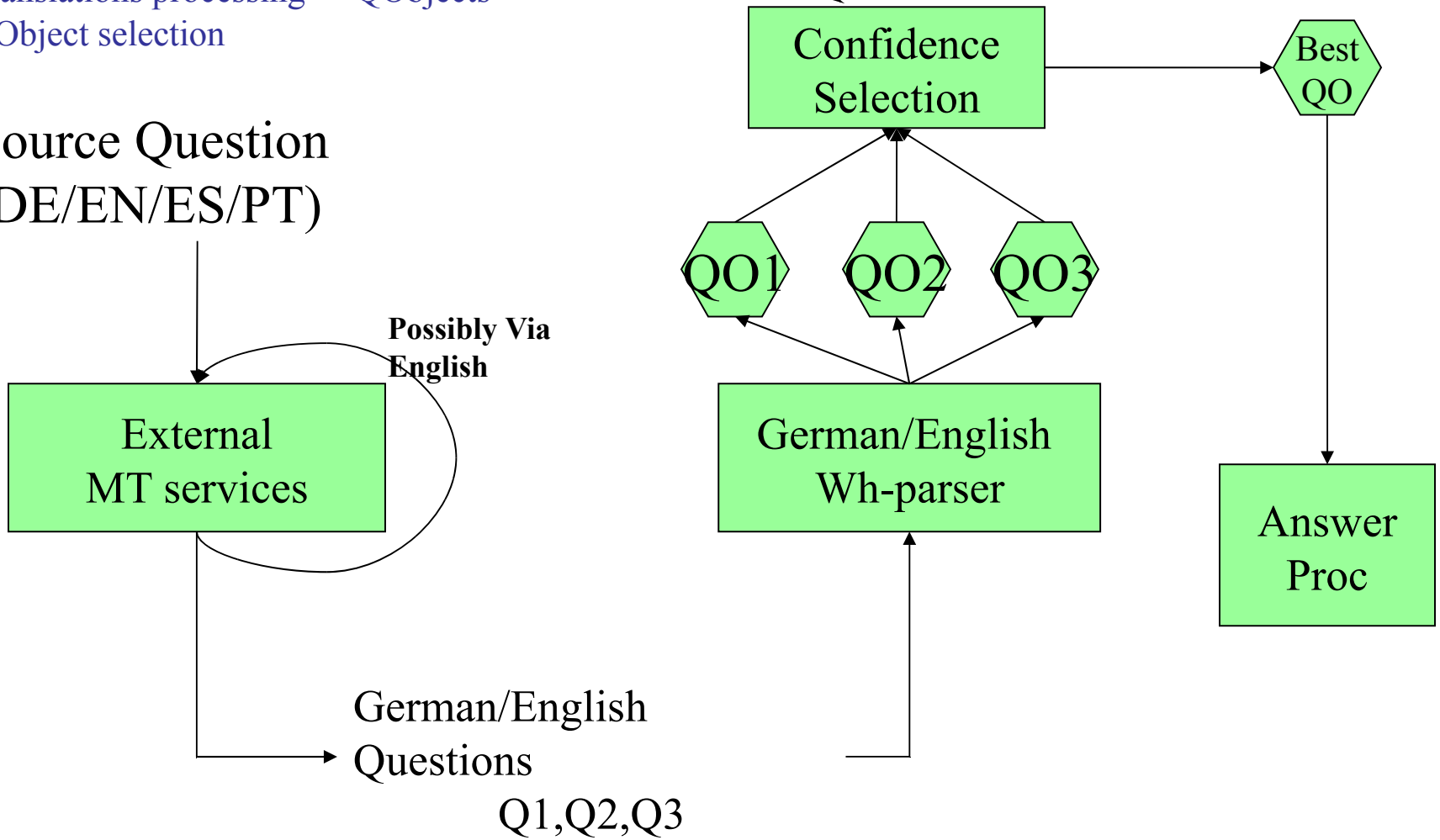
DFKI's Cross-[illegible]ach to ODQA

Assumption: the better the query analysis of a translated question is done the better was the translation being made

Before Method

- Question translation
- Translations processing -> QObjects
- QObject selection

Completeness wrt.
-Parse tree
-major semantic Wh-types



Project Idea

This is Bernardo, a DFKI guest from Trento just visiting Saarbrücken.
He wants to have a dinner tonight in a Spanish restaurant.
He calls the QALL-ME QA service provider:

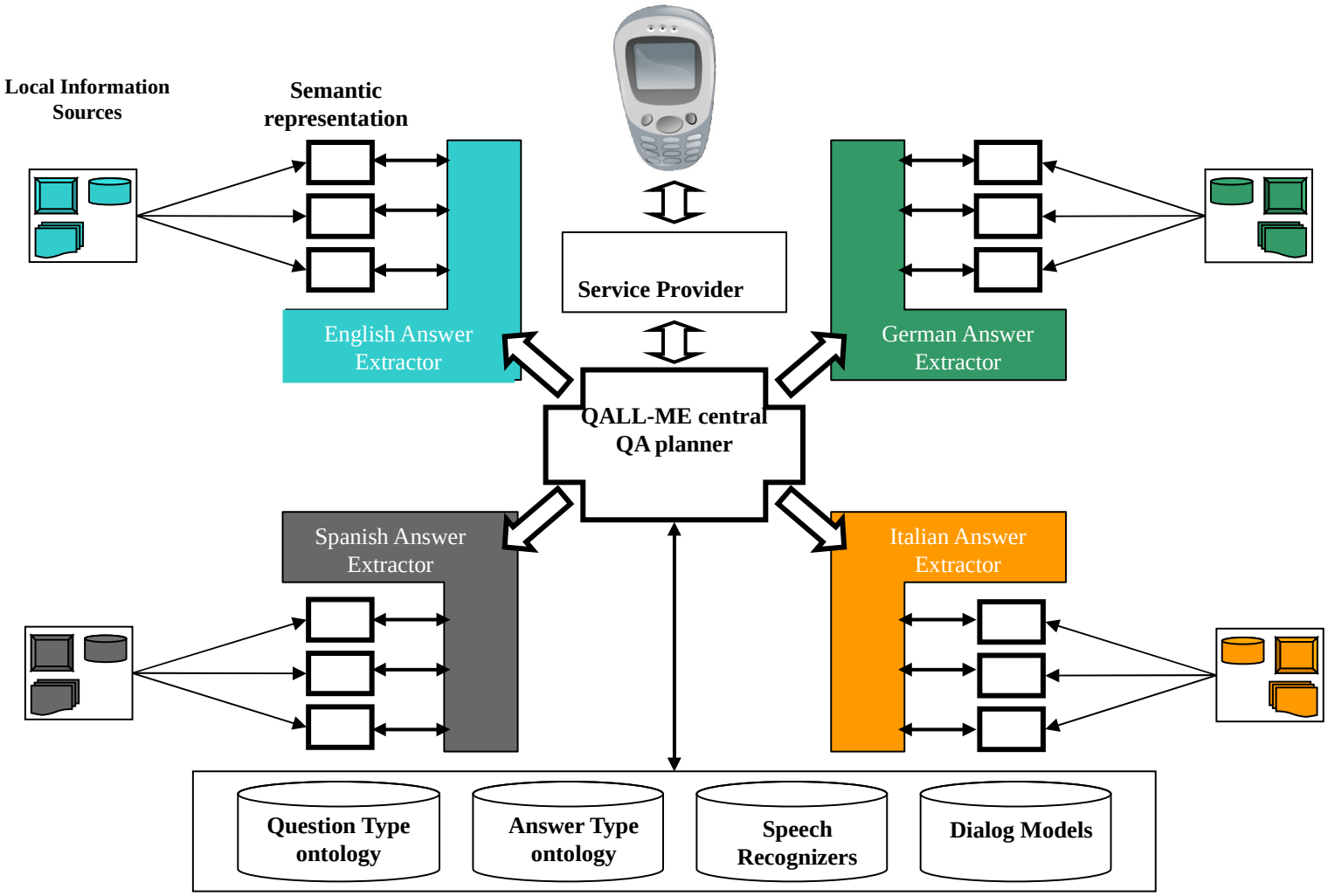
Dove posso mangiare paella questa sera?

QALL-ME
QA service provider

ZAPATA
offers paella
today.

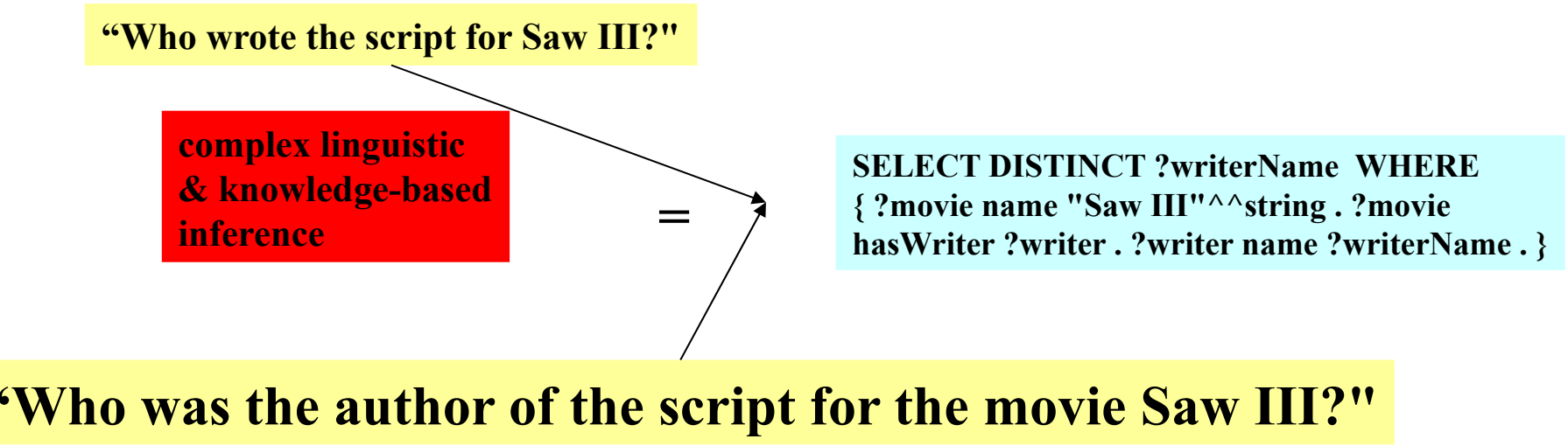
QALL-ME offers:
- Semantic access to tourism specific regional information
- NL query understanding in several languages entered via mobile devices (e.g., speech, SMS)
- Correct, complete and concise answers with different output presentation formats (e.g., texts, maps, images)
- spatial & temporal context (e.g., via GPS, time of call)

Architecture



The QA Bootleneck

- Hybrid QA:
  - Increase of semantic structure (Semantic Web, Web 2.0) ⇒ conflation of ontology-based data bases and information extraction from texts
  - Dynamic and openness of the web requires additional **new** complexity of the NL interfaces



Solutions

- Complete computation (inference)
  - AI complete; in particular, if incomplete/wrong queries are allowed
- Controlled sub-language
  - The user is only allowed to express questions in a particular form and with unique semantics
  - cognitive overhead is not acceptable
- Controlled mapping
  - One-to-one mapping between NL patterns and DB query patterns
  - NL degree of freedom realized through “textual inference”

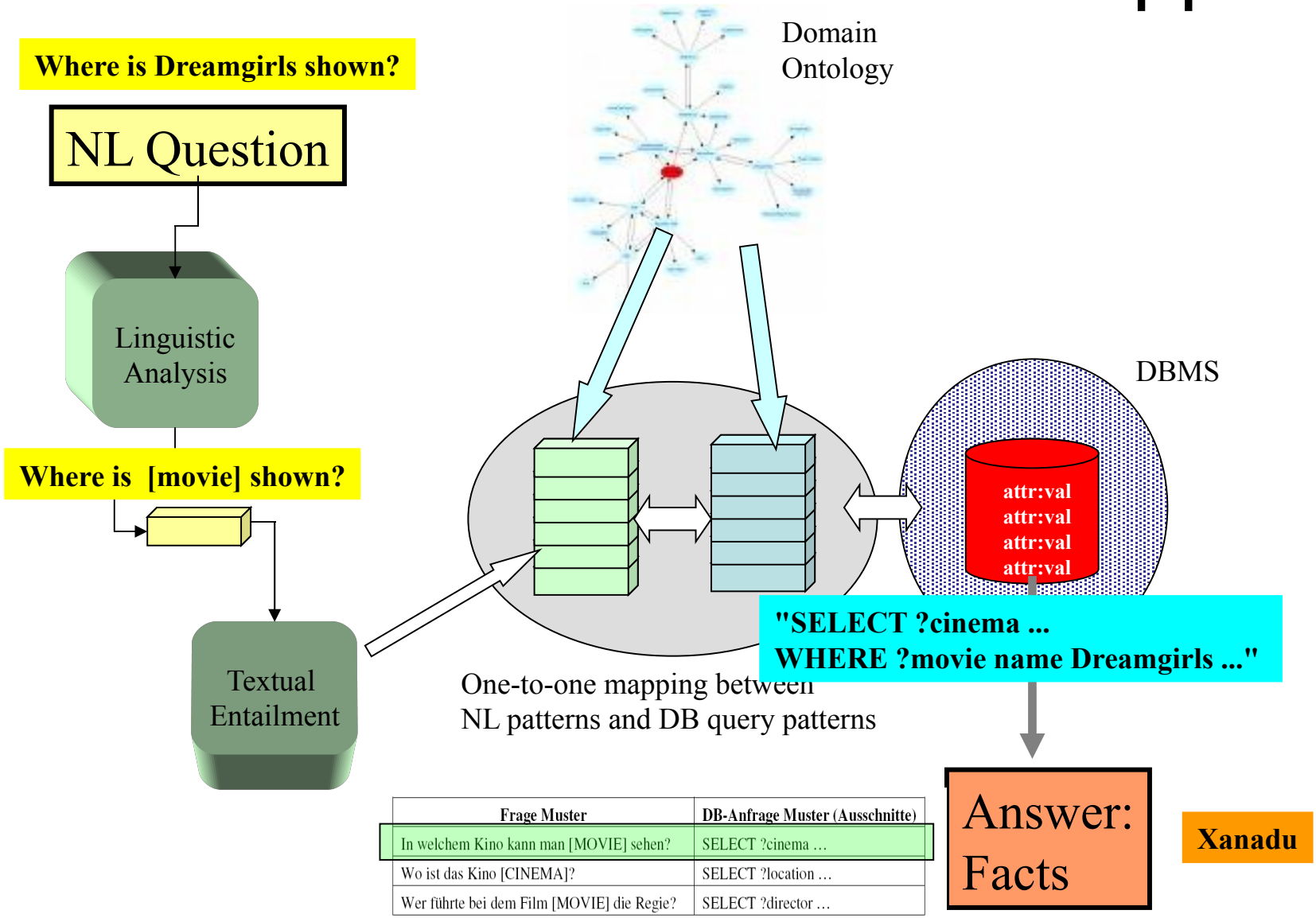
Textual Inference

Prof. Smart, who owns a chair at University the Best, has published a new paper.

Prof. Smart works for University the Best

- Motivation: textual variability of semantic expressions
- Idea: given are two text expressions T & H:
  - Does text T support an inference to hypothesis H?
  - Is H semantically entailed in T?
- PASCAL Recognising Textual Entailment (RTE) Challenge
  - since 2005, cf. Dagan et al.
  - 2007: 3te RTE challenge, 25 teams
- RTE is establishing itself as a core technology for text understanding applications:
  - QA, IE, semantic search, summarization, …

Entailment-based QA: A new approach



Frage Muster	DB-Anfrage Muster (Ausschnitte)
In welchem Kino kann man [MOVIE] sehen?	SELECT ?cinema …
Wo ist das Kino [CINEMA]?	SELECT ?location …
Wer führte bei dem Film [MOVIE] die Regie?	SELECT ?director …

Crosslinguality through
(manual) alignment of translated NL patterns.

Advantages

- Inferences is applied on the NL level
- RTE methods are by definition robust → supports processing of incomplete/ill-formed NL questions
- Opens up the possibility of automatically acquire mappings on basis of ontology-based and multilingual IE → hot research topic

Summary

- More and more Internet users with different languages
- Web2.0 allows NL based interaction through Web pages
- Cross-linguality and multi-lingual is the next natural step in the evolution of the Web
- High demands on multilingual HLT core technologies and applications, especially in the area of:
  - MT and multilingual (dependency) parsing
  - Integrated data-driven and symbolic strategies
  - Multilingual and cross-lingual corpora