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?

World Internet Users
March 2008

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

Internet Users in the World
March 2008

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

World Internet Penetration Rates
March 2008

North America - 73.1%
Oceania / Australia - 57.0%
Europe - 47.7%
Latin America / Caribbean - 23.8%
Middle East - 21.3%
World, Average - 21.1%
Asia - 14.0%
Africa - 5.3%

Internet Users Growth in the World
Between 2000 and 2008

Middle East - 1176.8%
Africa - 1030.2%
Latin America / Caribbean - 659.9%
Asia - 363.4%
World Growth Avg. - 290.0%
Europe - 263.5%
Oceania / Australia - 154.0%
North America - 127.9%

Note: Penetration Rates are based on a world population of 6,676,120,238 for mid-year 2008.
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](#)
Blogs versus Wikis

Blogs
„Collective Thinking, individual writing“

Wikis
„Collective Thinking, collective writing“

Publishing

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 keywords
    - Ex: `<*, located_in, Europe>` instead of „located in Europe“
  - Inference finds implicit 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
<table>
<thead>
<tr>
<th></th>
<th>Web 2.0</th>
<th>Web 3.0</th>
</tr>
</thead>
</table>
| Tagging                | • Annotation with *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

➢ Problem: High Recall but Low Precision
Motivation:
Events in an IR query overlap with event types from IE (ACE)

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

- 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

  - 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

IE as core for semantic annotation
• identification
• discovery
• validation
• evaluation
of semantic relationships & as basis for the automatic creation of meta data

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

Seeds: a short list of known NE instances/type

- Location
  - New York
  - Rabat
  - Germany
  ...
- Person
  - Bon Jovi
  - Mr.
  ...

Preprocessing:
- Tokenization;
- Pos Tagging;
- Chunk parsing;
- Dependency Parsing;

Core ML engine:
- Annotate
- Extract patterns
- Instantiate patterns
- New NE candidates
- Evaluate

Copy

Un-annotated documents

Preprocessed documents

New found entries

Identification of NE boundaries (phrases)

Classification of NE cands. (spelling, context)

Few language specific feature function
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
- 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**
  آقای
  دکتر
  خانم
  جناب
  بانو
  مهندس

- **Lexicon OrgDesc**
  استانداری
  وزارت
  دولت
  رئیس
  شورای
  شهرداری
  شهر

- **Lexicon POSITION**
  رئیس
  بانو
  مهندس
  شهردار
  شهرداری

- **Descriptors for named entities**
  **Lexicon PerDesc**
  سابق
  پیشینه

  **Lexicon CityDesc**
  شرکت

  **Lexicon CountryDesc**
  کشور
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="r101" 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

Seeds: a short list of known Single relation instances

<table>
<thead>
<tr>
<th>Location</th>
<th>Person</th>
<th>Born_in</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>Bon Jovi</td>
<td>Is born in</td>
</tr>
<tr>
<td>Rabat</td>
<td>Mr.</td>
<td>, born in</td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Preprocessing:
- Tokenization;
- Pos Tagging;
- Chunk parsing;
- **Dependency Parsing**;

Core ML engine:
- Annotate
- Extract patterns
- Instantiate patterns
- New RE candidates
- Evaluate

Identification of NE/Rel structure (subj, obj, verb phrase, etc.)

Classification of Rel cands. (spelling, context)

Un-annotated documents

Preprocessed documents

Copy

New found entries

Few language specific feature function
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 dependency 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)

<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>CPOS</th>
<th>POS</th>
<th>FEATS</th>
<th>HEAD</th>
<th>DEPREL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>This</td>
<td>this</td>
<td>pronoun</td>
<td>demon</td>
<td>sg</td>
<td>2</td>
<td>subj</td>
</tr>
<tr>
<td>2</td>
<td>is</td>
<td>be</td>
<td>v</td>
<td>v-fin</td>
<td>3</td>
<td>sg</td>
<td>pres</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>a</td>
<td>art</td>
<td>art</td>
<td>indef</td>
<td>4</td>
<td>det</td>
</tr>
<tr>
<td>4</td>
<td>test</td>
<td>test</td>
<td>n</td>
<td>nc</td>
<td>sg</td>
<td>2</td>
<td>comp</td>
</tr>
<tr>
<td>5</td>
<td>.</td>
<td>.</td>
<td>punc</td>
<td>punc</td>
<td>_</td>
<td>2</td>
<td>punc</td>
</tr>
</tbody>
</table>
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
اّتِّفاق

بَيْنَ

لُبْنَانّ

وَ

سُورّيا

عَلَى

رَفّعّ

مُسْتَوَى

التَبادُلّ

التّجارّيِّ

إّلَى

مّلْيُونّ

dُولارٍ

مُلْيُون
### Results for CoNLL 2006

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

<table>
<thead>
<tr>
<th></th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target languages</strong></td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td><strong>Type of questions</strong></td>
<td>200 Factoid</td>
<td>+ temporal restrictions + Definitions</td>
<td>-Type of questions + Lists</td>
<td>+ Linked questions + Closed lists</td>
<td></td>
</tr>
<tr>
<td><strong>Supporting information</strong></td>
<td>Doc.</td>
<td>Doc.</td>
<td>Doc.</td>
<td>Snippet</td>
<td>Snippet</td>
</tr>
<tr>
<td><strong>Pilots and exercises</strong></td>
<td>-Temporal restrictions - Lists</td>
<td>-AVE - RealTime - WiQA</td>
<td>- AVE - QAST</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 decrease of performance)
Results: Best and Average scores

![Graph showing best and average scores for CLEF03 to CLEF07]
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

- NL query In language X
- Machine Translation
- OD-QA SYSTEM
- NL texts In language Y
- Machine Translation
- ?
- ?
Two main different approaches used in Cross-Language QA systems:

1. Translation of the question into the target language (i.e. in the language of the document collection).
   - 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

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

1. Question processing in the source language to retrieve information (such as keywords, question focus, expected answer type, etc.)
   - Translation of the question into the target language (i.e. in the language of the document collection)
   - Answer extraction

2. Translation and expansion of the retrieved data
   - Answer extraction

Approaches in CL QA

Before Method
- S-CMI
- ISI
- DFKI
- EN2DE
- Limerik

After Method
- DFKI
- DE2EN
- ITC-irst
- RALI
- ISI
- DFKI
- EN2DE
- Limerik
- S-CMI
DFKI’s Cross-lingual Approach to ODQA

Before Method

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

Source Question (DE/EN/ES/PT)

External MT services

Possibly Via English

German/English Questions Q1,Q2,Q3

Confidence Selection

QO1 QO2 QO3

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

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

German/English Wh-parser

Best QO

Answer Proc
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 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)
QALL-ME central QA planner

Local Information Sources

Semantics representation

English Answer Extractor

German Answer Extractor

Spanish Answer Extractor

Italian Answer Extractor

Question Type ontology

Answer Type ontology

Speech Recognizers

Dialog Models

Service Provider

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

“Who wrote the script for Saw III?”

complex linguistic & knowledge-based inference

"Who was the author of the script for the movie Saw III?"

SELECT DISTINCT ?writerName WHERE {
  ?movie name "Saw III"^^string .
  ?movie hasWriter ?writer .
  ?writer name ?writerName .}


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

• 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
  – 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

NL Question

Linguistic Analysis

Domain Ontology

One-to-one mapping between NL patterns and DB query patterns

DBMS

"SELECT ?cinema ... WHERE ?movie name Dreamgirls ..."

Answer: Facts

Xanadu

Crosslinguality through (manual) alignment of translated NL patterns.

<table>
<thead>
<tr>
<th>Frage Muster</th>
<th>DB-Anfrage Muster (Ausschnitte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In welchem Kino kann man [MOVIE] sehen?</td>
<td>SELECT ?cinema ...</td>
</tr>
<tr>
<td>Wo ist das Kino [CINEMA]?</td>
<td>SELECT ?location ...</td>
</tr>
<tr>
<td>Wer führte bei dem Film [MOVIE] die Regie?</td>
<td>SELECT ?director ...</td>
</tr>
</tbody>
</table>
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