Multilingual Information Access

Carol Peters
IEI-CNR, Pisa
carol@iei.pi.cnr.it

Páraic Sheridan
MNIS-TextWise Labs
paraic@textwise.com

The 24th Annual International Conference on Research and Development in Information Retrieval (SIGIR2001)
Outline of Tutorial

- Introduction
- Multilingual Text Processing
- Cross-Language Text Retrieval
- Cross-Language Multimedia Retrieval
- System Evaluation
- Some Working Systems
- Library Applications
- Future Directions

SIGIR Tutorial, 9 September 2001
1. Introduction

- What is MLIA?
- Who is involved?
- Why is MLIA Important?
- Terminology / Definitions
- History of Research in MLIA
- Challenges
MLIA - The Problem

- Increasing pressure for access to information without language or cultural barriers means there is a strong demand to be able to:
  - find information written in foreign languages
  - read and interpret that information
  - merge it with information in other languages

- Need for Multilingual Information Access

SIGIR Tutorial, 9 September 2001
What is MLIA?

- MLIA related research regards the storage, access, retrieval and presentation of information in any of the world's languages.

- Two main areas of interest:
  - multiple language access, browsing, display
  - cross-language information discovery and retrieval
Multi-Language Access, Browsing, Display

The enabling technology:

- character encoding
- specific requirements of particular languages and scripts
- localization and presentation

SIGIR Tutorial, 9 September 2001
Cross-Language Information Retrieval

Crossing the language barrier...

- querying of multilingual collection in one language against documents in many other languages...
- filtering, selecting, ranking retrieved documents
**Multilingual Information Access**

**Terminology**

- Multilingual Information Access (MLIA) issues that involve overall management of multilingual information
- Cross-Language Information Retrieval (CLIR) querying of multilingual collection in one language to retrieve relevant documents in other languages
- Multilingual/Translingual Information Retrieval
MLIA is multidisciplinary

MLIA involves researchers from the following fields:

- information retrieval, natural language processing, machine translation and summarization, speech processing, document image understanding, human-computer interaction

SIGIR Tutorial, 9 September 2001
Areas of R&D involved

- **Data exchange** issues such as character encoding, font displays, browsing, etc.
- **Information processing** issues including NLP technologies, information discovery in multiple languages, cross-language access, speech processing, and summarization.
- **Language resources**, e.g. dictionaries and thesauri, corpora and test collections
Why is MLIA Important?

- Internationalisation
  - Multilingual countries (Switzerland, Canada)
  - Economic co-operative areas (EU, EFTA, NAFTA)
- Globalization of the economy
  - multinational companies
  - employees speak multiple languages
  - customers speak multiple languages
  - documents requiring access in multiple languages
Global Information Society

Information Society has enormous implications:

- wide range of applications where information should be accessible to users regardless of language:
  - e-commerce
  - education
  - entertainment
Global Information Society

- WWW as platform for knowledge dissemination
  - Distance Learning.....
  - Digital Libraries.....

- information providers and seekers should have equal opportunities
- preservation of national languages

SIGIR Tutorial, 9 September 2001
WWW and Internet

- Internet is no longer monolingual and non-English content is growing rapidly

- User profile has changed radically
  - From primarily academic use to widespread commercial, leisure, educational, entertainment etc. uses
How the Internet is Changing

- By 2005, 78% of internet users will speak a language other than English
- Total users will rise from 171 million to 345 million by 2005
- Therefore… 270 million non-English speaking Internet users in 2005 (up from 83M today)
83M Non-English Users Today

- Japanese: 17.4%
- Spanish: 17.1%
- German: 16.8%
- French: 10%
- Chinese: 7.7%
- Dutch: 5.3%
- Swedish: 4.3%
- Korean: 4.4%
- Italian: 4%
- Portuguese: 2.2%

SIGIR Tutorial, 9 September 2001
78% of Internet Users will be Non-English Speaking by 2005

2000

- English: 52%
- Other: 2%
- Spanish: 5%
- Japanese: 9%
- Chinese: 4%
- Italian: 6%
- Korean: 3%
- Dutch: 3%
- French: 3%
- Portuguese: 4%
- Scandanavian: 2%

2005

- English: 21%
- Other: 3%
- Spanish: 3%
- Japanese: 32%
- Chinese: 5%
- Italian: 5%
- Korean: 5%
- Dutch: 5%
- French: 5%
- Portuguese: 5%
- Scandanavian: 2%
Regional Web User Growth (‘99 to ’04)

W. Europe: 1.8X
Japan: 2.5X
Asia Pacific: 4.8X

Relative to US growth for same period

Source: IDC

SIGIR Tutorial, 9 September 2001
Widely Spoken Languages

Source: http://www.g11n.com/faq.html
# The Economics of Language Populations

<table>
<thead>
<tr>
<th>Language Name</th>
<th>World Population</th>
<th>Web Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese (various)</td>
<td>1,240,902,512</td>
<td>20.7%</td>
</tr>
<tr>
<td>Spanish</td>
<td>332,000,000</td>
<td>5.5%</td>
</tr>
<tr>
<td>English</td>
<td>322,000,000</td>
<td>5.4%</td>
</tr>
<tr>
<td>English ESL</td>
<td>420,000,000</td>
<td>7.0%</td>
</tr>
<tr>
<td>Bengali</td>
<td>189,000,000</td>
<td>3.2%</td>
</tr>
<tr>
<td>Hindi</td>
<td>182,000,000</td>
<td>3.0%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>170,000,000</td>
<td>2.8%</td>
</tr>
<tr>
<td>Russian</td>
<td>170,000,000</td>
<td>2.8%</td>
</tr>
<tr>
<td>Japanese</td>
<td>125,000,000</td>
<td>2.1%</td>
</tr>
<tr>
<td>German</td>
<td>98,000,000</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

Source: Summer Institute for Linguistics and GlobalReach

SIGIR Tutorial, 9 September 2001
People Prefer Their Native Language

Source: IDC Project Atlas, 1999

SIGIR Tutorial, 9 September 2001
The Language Gap

Year 2001

85% Language Gap

% Web Pages in English

45% % English Web Users

Source: IDC
Mono vs Multilingual Websites

Source: IDC eWorld Survey, 2001, IT Spending Weighted
### Website Language Support

<table>
<thead>
<tr>
<th>Language</th>
<th>Worldwide</th>
<th>Asia-Pacific</th>
<th>Latin America</th>
<th>North America</th>
<th>South Africa</th>
<th>Western Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted Count</td>
<td>9711</td>
<td>1548</td>
<td>1100</td>
<td>1683</td>
<td>235</td>
<td>5145</td>
</tr>
<tr>
<td>Chinese (Big 5)</td>
<td>1.6%</td>
<td>2.4%</td>
<td>0.1%</td>
<td>1.8%</td>
<td>2.0%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Chinese Simplified</td>
<td>2.8%</td>
<td>9.8%</td>
<td>0.9%</td>
<td>1.6%</td>
<td>3.0%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Danish</td>
<td>1.9%</td>
<td>0.2%</td>
<td>1.3%</td>
<td>1.3%</td>
<td>2.9%</td>
<td>4.1%</td>
</tr>
<tr>
<td>Dutch</td>
<td>3.9%</td>
<td>0.7%</td>
<td>0.5%</td>
<td>1.5%</td>
<td>5.8%</td>
<td>10.1%</td>
</tr>
<tr>
<td>English</td>
<td>75.7%</td>
<td>37.3%</td>
<td>40.0%</td>
<td>99.7%</td>
<td>100.0%</td>
<td>61.1%</td>
</tr>
<tr>
<td>Finnish</td>
<td>1.6%</td>
<td>0.1%</td>
<td>1.2%</td>
<td>1.4%</td>
<td>4.5%</td>
<td>2.9%</td>
</tr>
<tr>
<td>French</td>
<td>12.0%</td>
<td>1.0%</td>
<td>2.8%</td>
<td>9.1%</td>
<td>5.7%</td>
<td>24.8%</td>
</tr>
<tr>
<td>German</td>
<td>13.6%</td>
<td>1.1%</td>
<td>3.6%</td>
<td>5.7%</td>
<td>6.3%</td>
<td>35.7%</td>
</tr>
<tr>
<td>Italian</td>
<td>4.7%</td>
<td>0.5%</td>
<td>2.0%</td>
<td>2.2%</td>
<td>5.3%</td>
<td>11.7%</td>
</tr>
<tr>
<td>Japanese</td>
<td>15.8%</td>
<td>75.2%</td>
<td>1.7%</td>
<td>3.5%</td>
<td>5.7%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Korean</td>
<td>2.4%</td>
<td>9.3%</td>
<td>0.0%</td>
<td>1.1%</td>
<td>2.2%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Norwegian</td>
<td>1.4%</td>
<td>0.1%</td>
<td>1.1%</td>
<td>0.9%</td>
<td>3.9%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Portuguese - Brazilian</td>
<td>2.5%</td>
<td>0.4%</td>
<td>58.7%</td>
<td>1.3%</td>
<td>4.9%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0.8%</td>
<td>0.4%</td>
<td>0.7%</td>
<td>0.5%</td>
<td>3.9%</td>
<td>1.5%</td>
</tr>
<tr>
<td>Spanish</td>
<td>10.5%</td>
<td>0.8%</td>
<td>55.7%</td>
<td>12.6%</td>
<td>6.0%</td>
<td>9.1%</td>
</tr>
<tr>
<td>Swedish</td>
<td>2.5%</td>
<td>0.1%</td>
<td>1.5%</td>
<td>1.1%</td>
<td>2.7%</td>
<td>6.3%</td>
</tr>
<tr>
<td>Other</td>
<td>1.8%</td>
<td>0.3%</td>
<td>1.4%</td>
<td>1.4%</td>
<td>6.5%</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

History – Technology Milestones

- 1993: ISO/IEC 10646 is released as "Universal Multiple-Octet Coded Character Set" (UCS).
History – Technology Milestones


- Netscape/Explorer now offer support for UNICODE and foreign language fonts.
History – First Approaches

- 1970: English-German thesaurus (Salton)
- 1991-94: EMIR project - 1st EC CLIR project - free test retrieval for English, French and German
- 1994: 1st PhD thesis on CLIR by Khaled Radwan (France)
History – First Approaches

- 1996: Dictionary based retrieval
  - Umass & XEROX Grenoble

- 1996-7: Corpus-based approaches
  - ETH Zurich (Similarity Thesaurus)
  - CMU (Generalized Vector Space Model)
History – Community Milestones

- 1996: 1st Workshop on “Cross-Lingual Information Retrieval” at SIGIR ’96. Research community begins to be identified around this area.

- 1997: AAAI Spring Symposium on Cross-Language Text and Speech Retrieval
AAAII – The Grand Challenge

- Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified.

[AAAII Stanford Symposium 1997]
History – Community Milestones

- 1997: EU-NSF Working Group on Multilingual Information Access
- Tutorials/Workshops on MLIA/CLIR common at Information Retrieval, Computational Linguistics and Digital Libraries conferences around the world.

SIGIR Tutorial, 9 September 2001
History – Evaluation Milestones

- 1997: 1\textsuperscript{st} Cross-Language IR track within TREC (Text REtrieval Conferences)
- 1999: 1\textsuperscript{st} Japanese NTCIR Workshop on Text Retrieval System Evaluation includes track for Cross-Lingual Information Retrieval
History – Evaluation Milestones

- 1999: TIDES (Translingual Information Detection, Extraction, and Summarization) starts in U.S.

- 2000: CLEF – Cross-Language Evaluation Forum for European Languages is launched
Goals

- Support access to multilingual information in multiple media (text, speech, video)
- Process foreign-language information for indexing
- Retrieve relevant information in multiple languages from a single query
- Present retrieved information in form that is comprehensible to user
2. Multilingual Text Processing

- Writing Systems
- Character Encoding
- Language Recognition
- Tokenization
- Stop word removal
- Part-of-speech tagging
- Phrase identification

SIGIR Tutorial, 9 September 2001
## Writing Systems

<table>
<thead>
<tr>
<th>System</th>
<th>Level of Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Alphabet, single byte encoding</td>
</tr>
<tr>
<td>Most common W. European languages</td>
<td>Similar alphabets, some diacritics, differing punctuation, single encoding method available</td>
</tr>
<tr>
<td>Arabic</td>
<td>Bidirectional script, all cursive, alphabet, characters can vary with position, multiple diacritics,</td>
</tr>
<tr>
<td>Japanese</td>
<td>Possible mix of four writing systems, three alphabets plus Kanji characters, few diacritics, can be written in multiple directions. One word can be multiple characters. No spaces between words. A character can be one or more bytes</td>
</tr>
<tr>
<td>Chinese</td>
<td>Characters, some quite complex. Two systems, traditional and simplified. One word can be multiple characters. No spaces between words</td>
</tr>
</tbody>
</table>
Formatting

- Date & time, time zone
- Numbers
- Currency, percentage
- Text order

- Different order, symbols, length, name and even calendar
- Different shapes, symbols, negation
- Different separators, symbols, placement
- Critical to dynamic text

SIGIR Tutorial, 9 September 2001
### Numbers

<table>
<thead>
<tr>
<th>European</th>
<th>Arabic</th>
<th>Eastern Arabic</th>
<th>Indic-Devanagari</th>
<th>Indic-Bengali</th>
<th>Thai</th>
<th>Ideographic</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>०</td>
<td>०</td>
<td>০</td>
<td>๐</td>
<td>零</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>١</td>
<td>১</td>
<td>১</td>
<td>๑</td>
<td>一一</td>
</tr>
<tr>
<td>2</td>
<td>٢</td>
<td>٢</td>
<td>২</td>
<td>২</td>
<td>๒</td>
<td>二四</td>
</tr>
<tr>
<td>3</td>
<td>٣</td>
<td>٣</td>
<td>৩</td>
<td>৩</td>
<td>๓</td>
<td>三五</td>
</tr>
<tr>
<td>4</td>
<td>٤</td>
<td>٤</td>
<td>৪</td>
<td>৪</td>
<td>๔</td>
<td>四六</td>
</tr>
<tr>
<td>5</td>
<td>٥</td>
<td>५</td>
<td>৫</td>
<td>৫</td>
<td>๕</td>
<td>五七</td>
</tr>
<tr>
<td>6</td>
<td>٦</td>
<td>৬</td>
<td>৬</td>
<td>৬</td>
<td>๖</td>
<td>六八</td>
</tr>
<tr>
<td>7</td>
<td>٧</td>
<td>७</td>
<td>৭</td>
<td>৭</td>
<td>๗</td>
<td>七八</td>
</tr>
<tr>
<td>8</td>
<td>٨</td>
<td>৮</td>
<td>৮</td>
<td>৮</td>
<td>๘</td>
<td>八九</td>
</tr>
<tr>
<td>9</td>
<td>٩</td>
<td>৯</td>
<td>৯</td>
<td>৯</td>
<td>๙</td>
<td>九</td>
</tr>
</tbody>
</table>
## Currency Formats

<table>
<thead>
<tr>
<th>Country</th>
<th>Format</th>
<th>Currency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>$1,234,56</td>
<td>Peso</td>
</tr>
<tr>
<td>China</td>
<td>1,234.56</td>
<td>Yuan</td>
</tr>
<tr>
<td>France</td>
<td>1 234, 56 F</td>
<td>Franc</td>
</tr>
<tr>
<td>France</td>
<td>1 234,56€</td>
<td>Euro</td>
</tr>
<tr>
<td>Germany</td>
<td>1.234,56 DM</td>
<td>Deutsche Mark</td>
</tr>
<tr>
<td>Italy</td>
<td>L. 1.235</td>
<td>Lira</td>
</tr>
<tr>
<td>Italy</td>
<td>€1.234,56</td>
<td>Euro</td>
</tr>
<tr>
<td>Portugal</td>
<td>1.234$56 Esc.</td>
<td>Escudo</td>
</tr>
<tr>
<td>Russia</td>
<td>1 234,56p</td>
<td>Ruble</td>
</tr>
<tr>
<td>Switzerland</td>
<td>SFr. 1,234.56</td>
<td>Swiss Franc</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>£1,234.56</td>
<td>Pound</td>
</tr>
</tbody>
</table>

Adapted from “Java Internationalization,” O’Reilly, Cambridge MA, 2001

**SIGIR Tutorial, 9 September 2001**
# Punctuation Differences

<table>
<thead>
<tr>
<th>Quotes</th>
<th>Question Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>“English”</td>
<td>English?</td>
</tr>
<tr>
<td>&lt;&lt;French&gt;&gt;</td>
<td>¿Spanish?</td>
</tr>
<tr>
<td>„German”</td>
<td>Greek;</td>
</tr>
<tr>
<td>”Swedish”</td>
<td></td>
</tr>
<tr>
<td>&gt;&gt;Slovenian&lt;&lt;</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from “Java Internationalization,” O’Reilly, Cambridge MA, 2001
Character Encoding

- Binary representation of language’s alphabet.
- Text usually encoded in language-native form.
- Some languages require double-byte encoding (e.g. Japanese, Chinese).
- UNICODE standard for representation of all world’s languages.
- Support native codes or transform to UNICODE for processing/retrieval?
Character Encoding

- Language (alphabet) specific native encoding:
  - Chinese GB, Big5,
  - Western European ISO-8859-1 (Latin1)
  - Russian KOI-8, ISO-8859-5, CP-1251

- UNICODE (ISO/IEC 10646)
  - UTF-8 variable-byte length
  - UTF-16, UCS-2 fixed double-byte
UNICODE / ISO 10646

- Single 16-bit (2-byte) encoding designed to encompass all world’s written languages
- 16 bits allow for over 65,000 characters
- UNICODE currently specifies 38,887 characters
- Covers languages from Americas, Europe, Middle East, Africa, India, Asia
- Space for new characters or application-specific characters

SIGIR Tutorial, 9 September 2001
The Multilingual WorldWide Web

- Character encoding specified in HTTP Content-Type header field
  - “Content-type: text/html; charset=iso-2022-JP”
- HTML content – numeric character references interpreted according to ISO 10646/UNICODE
- HTML “Lang” attribute can be included in most HTML elements
  - <TEXT Lang=es>
The Multilingual WorldWide Web

- URL encoding in UTF-8 allowing local-language addresses for Web sites (e.g. Asian languages)

- Other issues
  - Bidirectional text – where left-to-right and right-to-left scripts are mixed
  - Format and units used in display of times, dates, weights etc.
Language Identification

- **Problem definition I**: Given a monolingual document taken from a multilingual collection, determine its language.

- **Problem definition II**: Given a multilingual document, determine the language of the paragraphs/sentences.

*SIGIR Tutorial, 9 September 2001*
Language Identification

- Based on native character encoding.
- Use statistical model of N-grams or words.
- Recognize language-specific characters.
- Use stopwords from IR
- Using language of previous paragraph or a default language
Multi-Language Indexing

- Simple language processing improves retrieval effectiveness
  - Stop word removal - 30% - 50% text size reduction
  - word normalization (word form → base form)
    - “Some form of stemming is almost always beneficial” [Hull & Grefenstette, 1996]
Tokenization

● Punctuation separated from words – incl. Word separation characters.

● String split into lexical units - incl. Segmentation (Chinese) and compound-splitting (German)

● “The train stopped.” →

  “The”, “train”, “stopped”, “.”
Chinese Segmentation Example

SIGIR Tutorial, 9 September 2001
Segmentation Strategies

- Choose a model
  - Unique, plausible strings, plausible interpretations
- Assemble evidence
  - Lexicons, corpora, algorithms, user knowledge
- Choose a preference criterion
  - Longest substring, proper name detection, etc.
- Choose a search strategy
  - Greedy, exhaustive, dynamic programming
Segmentation Models

- Unique segmentation
  - Decide whether to put a boundary at each point
- Plausible strings
  - Produce all substrings that might be useful
- Plausible interpretation
  - Produce all terms that might plausibly be implied
    - Contractions, alternate spellings, etc.
Segmentation Evidence

- Lexical knowledge
  - Dictionaries, term lists, name lists, gazetteers
- Corpus statistics
  - Within-document, within-collection, balanced
- Algorithmic knowledge
  - Transliteration rules, name cues, parsers, …
- User knowledge
  - Forced join, forced split

SIGIR Tutorial, 9 September 2001
German Segmentation

- Unrestricted compounding in German
  - *Abendnachrichtensendungsblock*

- Use compound analysis together with CELEX German dictionary (360,000 words)
  - *Treuhandanstalt* → \{ treuhand, anstalt \}
  - *Washington* → \{ * was, hing, ton \}

- Dictionary maintenance is crucial
Stop Word Removal

- Frequent stop words (e.g. “the”, “an”, …)
- Frequent non-stop words (e.g. “Mexico” in the TREC El Norte collection)
- Infrequent stop words (e.g. “nichtsdestotrotz”)
- Domain dependent vs. independent stop words (e.g. “SDA” in SDA news)
Normalizing Indexing Features

- Reduce words from their surface (inflected) form to some ‘base’ (root) form to broaden matching:
  - Rule-based suffix stripping (e.g. Porter)
  - Finite-state morphology (e.g. InXight)
  - Dictionary-Based normalization

SIGIR Tutorial, 9 September 2001
Porter Stemming

- Rule-based suffix stripping
- 65 rules applied in up to 5 iterations
- Linguistic correctness of resulting stems not necessary
- 36% reduction in indexing vocabulary (English)
- Versions written for many languages
<table>
<thead>
<tr>
<th>Rule</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(True)IES → I</td>
<td>sensitivities</td>
</tr>
<tr>
<td>(m &gt; 0)IVITI → IVE</td>
<td>sensitiviti</td>
</tr>
<tr>
<td>(m &gt; 1)IVE → ε</td>
<td>sensitive</td>
</tr>
<tr>
<td></td>
<td>sensit</td>
</tr>
</tbody>
</table>
# Stemming French (ETH)

- 84 suffix rules arranged in 8 groups

<table>
<thead>
<tr>
<th>Suffix group</th>
<th>Stemming Rule</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>er(s), ère(s)</td>
<td>(*)ère → er</td>
<td>dernière → dernier</td>
</tr>
<tr>
<td>teur(s), trice(s)</td>
<td>(*)trice → teur</td>
<td>éducatrice → éducateur</td>
</tr>
</tbody>
</table>
**Stemming Italian (ETH)**

- 220 rules (plural → singular, gender, verb tenses and moods)

<table>
<thead>
<tr>
<th>Stemming Rules</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>(*c)e → ia</td>
<td>province → provinciа</td>
</tr>
<tr>
<td>(*l)cissim[aeio] → ce</td>
<td>dolcissima → dolce</td>
</tr>
<tr>
<td>(*e)sse → re</td>
<td>volesse → volere</td>
</tr>
</tbody>
</table>

*SIGIR Tutorial, 9 September 2001*
Morphological Analysis

- Inxight LinguistX analyzer produces lexemes instead of non-word stems (as opposed to Porter stemming algorithm)
- Inflectional versus derivational morphology
- See [Hull & Grefenstette 1996] for a comparison of morphological analysis and stemming.
- Likely to be language-dependent choice (e.g. English vs Finnish vs Arabic)
Part-of-speech Tagging

- Assign part-of-speech tags from a standard ‘tagset’ – trade off tagset size versus complexity
- English - typical tag set is about 50
- French - 264 large set or 56 reduced set
  - AFS   adjective feminine singular
  - NFS   noun feminine singular
  - V1SPI verb first person singular present indicative
- Reduce set by truncating from right to left
POS Tagging Examples

- English:
  present (noun)
  present (verb)

- French:
  marine - AFS    marine - V1SPS
  marine - NFS    marine - V2SPM
  marine - NMS    marine - V3SPI
  marine - V1SPI  marine - V3SPS

SIGIR Tutorial, 9 September 2001
POS-Tagging Techniques

- Statistical - Church 1988
- Rule-based
  - “Learned” Rules (Brill 1992)
- Combinations - Bell Labs, Xerox- Grenoble

For multilingual information access, robust tagging reduces ambiguity.
Phrase identification

- Non-compositional phrases are critical in cross-language retrieval
  - “Fast food” ≠ “nourriture rapide”
- Retrieval accuracy is improved through effective phrase identification
- Translation resource must include correct translations for non-compositional phrases
Phrase Identification

- Statistical methods
  - eliminate stop words
  - cluster by context and frequency
  - E.g. All word pairs which co-occur >25 times

- Symbolic methods
  - tag text for part of speech
  - rules for phrase identification
Named entity recognition

- A special case of phrase recognition
- Named entity terms are productive and rarely used
- Accurate processing techniques cannot depend exclusively on a stable lexicon of terms
- Recognition based on grammar rules and cues
Named Entity Classification

[Diagram showing a hierarchy of entity types and subtypes, including Geographic, Affiliation, Organization, Human, Document, Equipment, Scientific, Temporal, and Misc. categories with further subcategories like City, Port, Religion, Nationality, Company, Person, Document, Software, Disease, Date, and Misc.]
3. Cross-language Text Retrieval

- Approaches
- Difficulties
- Resources

SIGIR Tutorial, 9 September 2001
The Problem

- How can concepts represented by a query in one language be matched against information contained in documents in other languages?
The Problem
CLIR - Approaches

- Multilingual Thesauri
- Machine Translation
- Bilingual Dictionaries
- Parallel/ Comparable Corpora
- Conceptual Interlingua

SIGIR Tutorial, 9 September 2001
Using Thesauri

Controlled Vocabulary Searching

- defined set of **concepts** for indexing and searching represented by sets of terms in each language
- ambiguity eliminated
- (some) results guaranteed
Using Thesauri

Problems

- Thesauri are costly to build and maintain
- Manually assigning terms to documents is expensive
- Mapping between thesauri in different languages is difficult
- Users have problems using thesauri in searching

SIGIR Tutorial, 9 September 2001
CLIR – Machine Translation

- Translate all **documents** into each of the possible query languages
  - Not viable on large collections
  - Not viable if many possible query languages

- Document translation using MT system is expensive and redundant for CLIR
CLIR – Machine Translation

- Translate the **query** into languages of the content being searched
  - no context for accurate translation
  - system selects preferred target term
- Query translation using MT system is inadequate for CLIR
CLIR – Using Dictionaries

- bilingual machine-readable dictionaries (in-house or commercial)
- look-up query terms in dictionary and replace with translations in document languages
- automatic query translation gives ca 50% precision wrt monolingual retrieval
- query expansion techniques reduce ambiguity / improve recall
CLIR – Using Dictionaries

Problems

- ambiguity
- many terms are out-of-vocabulary
- lack of multiword terms
- phrase identification
- bilingual dictionary needed for every query-document language pair of interest

SIGIR Tutorial, 9 September 2001
**MT versus Dictionary Lookup**

- **Search request**
  - montrer tous les messages traitent des sources d’intimité en raison des caméras vidéo toujours actives

- **English translation**
  - show all messages about privacy concerns due to always active video cameras.

- **Porter stemming**
  - show messag privaci concern due alwai activ video camera
Machine Translation

(montrer tous les messages traitent des sources d’intimité en raison des caméras vidéo toujours actives)

to show all the messages always treating the worries the intimacies camcorders active

- limited vocabulary ≠ search term mismatch
- erroneous translation ≠ retrieve non-relevant documents
Dictionary Term Lookup

(montrer tous les messages traitent des sources d’intimité en raison des caméras vidéo toujours actives)

dict point out show reveal appear - message get message across - treat negotiate do or make deal with process -

problem marigold - intimacy cosines depths private life

privacy - reason ratio - movie camera video - always anyway still - vigorous working person buoyant

stimulate stir up stoke speed up active activate

excessive vocabulary □ search term mismatch
eroneous translation □ retrieve non-relevant docs
CLIR – Using Corpora

Provide lexical equivalences over languages

- Parallel Corpora
  - translation equivalent
  - e.g. UN corpus in French, Spanish & English

- Comparable Corpora
  - similar for topic, style, time etc.
  - e.g. Swiss news agency reports in German, French, Italian

SIGIR Tutorial, 9 September 2001
Using Corpora

- Query translation using Parallel Corpora
  - align texts using statistical information or bilingual dictionaries
  - find correspondence between words in source language and words in target language
  - create pseudo-translation of query for retrieval in target language
Using Corpora

- Query Translation using Comparable Corpora
  - align related documents via descriptors
    - dates, keywords, proper nouns
  - build co-occurrence lexicon
  - terms in different languages referring to same topic will cooccur in each document
  - use correspondence to pseudo-translate queries
Corpus-based Approaches

- Generalized Vector Space Model (GVSM)
  - uses a bilingual training corpus to build matrices of documents & term weights in each language
  - uses parallel corpora for each language pair

- Latent Semantic Indexing
  - reduces the GVSM further
  - requires parallel or comparable corpora
  - computationally expensive
Corpus-based Approaches

- Similarity thesauri
  - a generalization of the Vector Space Model
  - extracts translationally equivalent terms from aligned multilingual corpora
  - records equivalencies in an external thesauri
  - reliant on high quality corpora

SIGIR Tutorial, 9 September 2001
Similarity Thesaurus

French Documents

German Documents

Align Comparable Documents

Bilingual French /German Documents

Build Multilingual Similarity Thesaurus

Nordirland

irland
irlandais
ulster
protestant
Using Corpora

- Problems
  - suitable corpora are hard to obtain
  - training corpora must be very large
  - corpora tend to be application/domain dependent
synonyms in all languages

Enables matching to equivalent terms and concepts.

Terms and phrases from multiple languages

CLIR – Conceptual Interlingua
Multilingual Information Access

Conceptual Interlingua

Concept Space

General Language Concepts

Vocabulary 1
French

Vocabulary 2
Spanish

Vocabulary 3
Japanese

SIGIR Tutorial, 9 September 2001
Advantages

- CLIR in any language combinations – not only bi-directional (e.g. French-Japanese)
- Language-independent retrieval based on natural language concepts
- Populate terminology within a well understood conceptual framework
- Well-understood framework for word sense disambiguation
Disadvantages

- expensive to construct
- problem of vocabulary coverage
- language-dependent concept incompatibilities
State of the Art

- Machine Translation
  ~80% monolingual effectiveness in general domain

- Dictionary Techniques
  ~ 80% of monolingual effectiveness in general domain

- Parallel and Comparable Corpus Techniques
  ~ 80% monolingual effectiveness in general domain
  ~ 90% monolingual in special domain

SIGIR Tutorial, 9 September 2001
Main CLIR Difficulties (I)

- Translation
  - ambiguity
  - erroneous translations
  - phrase/multiword identification
  - synonym identification
Reduce Ambiguity

- syntactic preprocessing
- statistical analysis
  - term cooccurrence
    - related terms tend to occur together
    - disambiguate translations using co-occurrence statistics
- user feedback
- pseudo-relevance relevance feedback
Pseudo-Relevance Feedback

- Enter query in source language
- Perform retrieval on parallel corpus or on comparable corpus to retrieve documents in the source language
- Use mate-documents of retrieved documents and derive a query in target language (e.g. by Rocchio)
- Use query to retrieve documents in target language
Main CLIR Difficulties (II)

- MLIA systems need well-developed resources
  - Language Processing Tools
  - Language Resources

- Resources are expensive to acquire, maintain, update
Resources - Main problems

- availability
- extensibility
- coverage
- quality
- standards
Language Processing Tools

- language identification tools
- character set conversion
- tokenisation/word segmentation
- morphologies/stemming tools

See ACL Natural language Software Registry
http://registry/dfki/de/
Language Resources

- dictionaries
- corpora
- lexicons and terminologies
- thesauri and ontologies


and LDC - Linguistic Data Consortium - http://www.ldc.upenn.edu

SIGIR Tutorial, 9 September 2001
4. CLIR for Multimedia

- Technology is required to automate access to information contained in multilingual, multimedia digital archives:
  - text
  - speech
  - images
  - video

- Need for complex integration of multiple technologies
CMU Informedia Project

The results set shows the best 12 of 628 matches on any of "tell me about the evolution of species."

Click on a word to focus on it. Press the shift key while clicking to have multi-word focus.

Darwin observed Galapagos and developed theory of evolution, 0:00:42, 1988

When Charles Darwin came here 150 years ago he was puzzled by the differences between similar animals from
CLIR for Multimedia

- Retrieval from a mixed media collection is non-trivial problem
- Different media processed in different ways and suffer from different kinds of indexing errors:
  - spoken documents indexed using speech recognition
  - handwritten documents indexed using OCR
  - images indexed using significant features
CLIR for Spoken Documents

- monolingual speech retrieval active area of research within IR community
- current research primarily in English, with some work in Chinese, Japanese and some European languages (German, French, Italian, Dutch)
- very few studies into cross-language spoken document retrieval

SIGIR Tutorial, 9 September 2001
Speech Retrieval Overview

- Analysis typically consists of a speech recognition phase followed by indexing feature extraction.
- Speech recognition viewed as a black box with either phonemic or lexical outputs
- Output of speech recognition used as input to indexing – focus is to find best indexing features from speech output
Multilingual Speech Processing

- Speech recognizers typically trained over many hours of tagged spoken corpora
- Not readily available in many languages
- Current research primarily in English, with some work in Chinese, Japanese and German
- After digital signal processing, speech processing much like lexical analysis of Chinese text...!

SIGIR Tutorial, 9 September 2001
Cross-Language Speech Retrieval

- Retrieve spoken German documents from queries written in French (ETH)
- Build French-German similarity thesaurus over comparable corpus of French and German Swiss (SDA) news
- Index German (Swiss) radio news as phoneme tri-grams from output of speech recognition
Cross-Language Speech Retrieval

- Accept written French query from user
- Use similarity thesaurus to “pseudo-translate” query into German text
- Use phoneme dictionary to translate German text into associated phonemes
- Match German query phonemes in index of German spoken news
- Return spoken German documents
Cross-Language Speech Retrieval

- Evaluated over a collection of 30 hours of spoken German news
- As baseline (German queries) 44% of queries had at least one relevant document in top 10
- Cross-language, 20% of queries had at least one relevant document in top 10.
- Cross-language retrieval ~50% of monolingual
- Difficult evaluation environment – resources

SIGIR Tutorial, 9 September 2001
Current Directions

- Recent TREC evaluation (track) focused on Spoken Document Retrieval
- Topic Detection and Tracking (TDT) evaluation focused on spoken content (including Chinese)
- Summer 2000 Johns Hopkins workshop on Cross-Language Speech Retrieval (Oard et al)
- Automatic Content Extraction (ACE) program involving multilingual speech

SIGIR Tutorial, 9 September 2001
Current Directions

- EU-NSF Working Group on Spoken Document Archives in Multiple Languages
- DELOS WG to study cross-language spoken document system evaluation
5. Evaluation for MLIA Systems

- Why System Evaluation is Important
- What is implied by Evaluation
- Evaluation programs
- An Example

SIGIR Tutorial, 9 September 2001
**Why we need Evaluation**

- evaluation permits hypotheses to be validated and progress assessed
- evaluation leads to contrastive analysis of approaches/technologies
- evaluation saves developers time and money

CLIR systems are still in experimental stage

*Evaluation is particularly important!*

*SIGIR Tutorial, 9 September 2001*
CLIR System Evaluation is Complex

CLIR systems consist of integration of components and technologies

- need to evaluate single components
- need to evaluate overall system performance

SIGIR Tutorial, 9 September 2001
CLIR System Evaluation is Complex

- need to distinguish methodological aspects from linguistic knowledge
- ideally, an evaluation protocol would require distinction between architecture, program and linguistic data
Technology vs. Usage Evaluation

Usage Evaluation:

- shows value of a technology for user
- determines the technology thresholds that are indispensable for specific usage
- provides directions for choice of criteria for technology evaluation

Influence of language and culture on usability of technology needs to be understood

SIGIR Tutorial, 9 September 2001
Organising an Evaluation Activity

- select control task(s)
- provide data to test and tune systems
- define protocol and metrics to be used in results assessment

Aim is an objective comparison between systems and approaches
Advantages of Evaluation

- Participants are motivated to improve their systems to ensure good results.
- Test-suites are developed and available for research community.
- Areas where more R&D is needed are identified.
Main CLIR Evaluation Programs

- **TIDES**: sponsors TREC (Text REtrieval Conferences) and TDT (Topic Detection and Tracking) - Chinese-English tracks in 2000; TREC focussing on English/French - Arabic in 2001
- **NTCIR**: Nat.Inst. for Informatics, Tokyo. Chinese-English; Japanese-English C-L tracks
- **AMARYLLIS**: focused on French; 2nd campaign 98-99 included C-L track; 3rd campaign begins Sept.01
- **CLEF**: Cross Language Evaluation Forum - C-L evaluation for European languages
Cross-Language Evaluation Forum - CLEF

- Funded by DELOS Network of Excellence for Digital libraries and US National Institute for Standards and Technology (NIST)
- Extension of CLIR track at TREC (1997-1999)
- Coordination is distributed - national sites for each language in multilingual collection
CLEF - Main Goals

Promote research by providing an appropriate infrastructure for:

- CLIR system evaluation, testing and tuning
- comparison and discussion of results
- building of test-suites for system developers

SIGIR Tutorial, 9 September 2001
CLEF 2001 - Task Description

Four main evaluation tracks in CLEF 2001:

- multilingual information retrieval
- bilingual information retrieval
- monolingual (non-English) information retrieval
- domain-specific IR

plus

- experimental track for interactive C-L systems
CLEF 2001 - Data Collection

- Multilingual comparable corpus of news agencies and newspaper documents for six languages (DE, EN, FR, IT, NL, SP). Nearly 1 million documents

- Common set of 50 topics (from which queries are extracted) created in 9 European languages (DE, EN, FR, IT, NL, SP+FI, RU, SV) and 3 Asian languages (JP, TH, ZH)
CLEF 2001 - Multilingual IR

Topics either DE, EN, FR, IT, FI, NL, SP, SV, RU, ZH, JP, TH

Documents
- English
- German
- French
- Italian
- Spanish

Participant’s Cross-Language Information Retrieval System

One result list of DE, EN, FR, IT and SP documents ranked in decreasing order of estimated relevance

SIGIR Tutorial, 9 September 2001
Task: query English or Dutch target document collections

Goal: retrieve documents for target language, listing results in ranked list

Easier task for beginners!

SIGIR Tutorial, 9 September 2001
Task: querying document collections in FR|DE|IT|NL|SP
Goal: acquire better understanding of language-dependent retrieval problems

- different languages present different retrieval problems
- issues involved include word order, morphology, diacritic characters, language variants
Goal: understand implications of querying in monolingual or cross-language task translations of document titles

German/English/Russian thesauri and English vertical domain (social sciences) in German

Task: querying a structured database from a domain-specific context
Multilingual Information Access

CLEF 2001 - Interactive C-L

Task: interactive document selection in an “unknown” target language

Goal: evaluation of results presentation rather than system performance

SIGIR Tutorial, 9 September 2001
CLEF2001 - Approaches

All traditional approaches used:

- commercial MT systems (Systran, Babelfish, Globalink Power Translator, )
  ✓ both query and document translation tried
- bilingual dictionary look-up (on-line and in-house tools)
- aligned parallel corpora (web-derived)
- comparable corpora (similarity thesaurus)
- conceptual networks (Eurowordnet, ZH-EN wordnet)
- multilingual thesaurus (domain-specific task)
Text processing for multiple languages:
- Porter stemmer, Inxight commercial stemmer, on-site tools
  - simple generic “quick&dirty” stemming
  - language independent stemming
- separate stopword lists vs single list
- morphological analysis
- n-gram indexing, word segmentation, decompounding (e.g. Chinese, German)
- use of NLP methods, e.g. phrase identification, morphosyntactic analysis
CLEF2000 - Techniques Tested

Cross-language strategies included:

- integration of methods (MT, corpora and MRDs)
- pivot language to translate from L1 -> L2 (DE -> FR, SP, IT via EN)
- N-gram based technique to match untranslatable words
- prior and post-translation pseudo-relevance feedback (query expanded by associating frequent cooccurrences)
- vector-based semantic analysis (query expanded by associating semantically similar terms)
CLEF2000 - Techniques Tested

- Different strategies experimented for results merging
- This remains still an unsolved problem
Evaluation - Summing up

- system evaluation is not a competition to find the best
- evaluation provides opportunity to test, tune, and compare approaches in order to improve system performance
- an evaluation campaign creates a community interested in examining the same issues and comparing ideas and experiences

SIGIR Tutorial, 9 September 2001
6. Working Systems

- Experimental Systems
- Commercial Systems
- Examples (Screenshots)
Working Systems

- Systems which have been built for IR experimentation (TREC, CLEF, NACSIS)
  - Support large scale document indexing
  - Support batch evaluation of long queries
- Systems which have been built for commercial use
  - Provide rapid responses to user queries
  - Support dynamic content update

SIGIR Tutorial, 9 September 2001
• SMART: the retrieval engine developed at Cornell. Used widely for IR experimentation. Used by several groups with different approaches to CLIR

• TwentyOne: system developed in EU funded project. Used to explore several approaches to CLIR, including dictionaries, MT, language models
Multilingual Information Access

Working Experimental Systems

- U.C. Berkeley: Experiments using Machine Translation
- IBM Watson: Statistical Machine Translation
- Montreal U.: Bilingual dictionaries extracted from parallel/comparable web corpora

SIGIR Tutorial, 9 September 2001
Working Experimental Systems

- New Mexico State University
- URSA – Unicode Retrieval System Architecture
- ARCTOS/J24 – Document Thumbnail images
- Keizei – Multilingual Retrieval & Summarization
- MUNDIAL – Multilingual access through standard web search engines
industrial applications of cyanide

FT921–9508

AN–CBLA2AF3.PT

920212

FT 12 FEB 92 / Commodities and Agriculture: US sued over Chilean fruit ban

By LESLIE CRAWFORD

SANTIAGO

CHILEAN FRUIT growers and exporters are suing the US government for $212m in compensation for the losses they suffered during a three-week trade embargo against Chilean fruit in March 1989.

Mr Ronald Bowen, president of the Chilean Exporters Association, said the lawsuit would be lodged today in Philadelphia, the main US port of entry for Chilean fruit.
The US Food and Drug Administration halted Chilean fruit imports three years ago after discovering two grapes laced with cyanide in a shipment.
bank interest rates

bank interest rates

bank interest rates
Keizai --- Cross Language Document Retrieval from CRL, the Computing Research Lab

Translate Query
Submit Query

Search News Source
Xinhua News 64-85 (Chinese)

View entire document

阿尔及利亚各省救援地震灾民

新华社8月13日电据阿尔及利亚通讯社报道，阿尔及利亚全国各省已动员起来救援穆斯科特省的地震灾民。

阿尔及利亚西部的穆斯科特省18日凌晨2时发生里氏5.6级地震，造成大约140人死亡，289人受伤，500至600户受灾。

报道说，第一批救援人员已于18日清晨抵达受灾最严重的地区哈西纳。

内政部长谢里夫也于地震的当天前往灾区了解灾情，并同当地官员一起监督救援工作。

穆斯科特省动员了1000多名民防队员参加救援工作，同时，穆斯科特省军警、宪兵、丘鲁拉、阿布利卡和哈西纳等省已动员起来向灾区运送帐篷、毛毯、食品、救护车和救援医务人员。在重灾区，已经搭起一些供灾民居住的帐篷。医疗队正赶来医治伤员，一些伤势最严重的伤员已送往医院接受紧急治疗。
La ricerca è stata così tradotta:

- cure
  - generale
  - accurata, precisione
  - cure, care, carefulness, accuracy, attention
  - medicina
  - treatment, cure, care, nursing

- cancro
  - medicina
  - cure, cancer, tumour
  - Astronomia, Astrologia
  - Caner

- polmonare
  - trad. pulmonary, lung

Eurosearch
Commercial Systems

- APORT: Online English/Russian search engine
- STRATEGIS: Canadian government portal for business information (user term selection)
- EUROSPIDER: Similarity Thesaurus CLIR
- TITAN: Japanese search portal with English MT
- AltaVista Babelfish: Systran MT of queries
- Document Management Partners: SCOUT
- CINDOR: Conceptual Interlingua Retrieval

SIGIR Tutorial, 9 September 2001
Gerichtsverfahren bei Globusübernahme durch Migros

<table>
<thead>
<tr>
<th>Relevanz</th>
<th>Dokument</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>überrelevante</td>
<td>BILANZ ONLINE</td>
<td>Gerade mal zehn Jahre Verschiedenfarbige Erstausgabe, aber wahr: Dieses Mänv...</td>
</tr>
<tr>
<td>överrelevante</td>
<td>NZZ</td>
<td>Transparenz ist die beste Medizin. Erkenntnisse einer Tagung der Schweizer Börse SWX Zunahme Normen? Information ist aller...</td>
</tr>
<tr>
<td>överrelevante</td>
<td>SDA</td>
<td>Globus: Übernahme durch Migros. Insider-Strafverfahren: denkbare Betrugsanwaltschaft...</td>
</tr>
<tr>
<td>überrelevante</td>
<td>WWW</td>
<td>Yahoo! Finanzen - Rahmenbedingungen: STIS-Weekly No. 13/99...</td>
</tr>
</tbody>
</table>

Suche neueste Unternehmensergebnisse
Wir über uns
Wie berauschend ich den Spiderview Wirtschaftsinformationsdienst?
Welche Dokumente sind auf der Suche nach?
Globus-Übernahme durch Migros: Insider-Strafverfahren denkbar Bezirksanwaltschaft
denkt über Erweiterung der Vorabklärung nach


Insider nicht gleich Insider

Entscheidend für die Einleitung eines Strafverfahrens sei, ob die beteiligten Unternehmen - also die Migros als Gesellschaft oder die Globus-Besitzerfamilien - selbst Wertpapiere gekauft haben, erläuterte Tewin weiter. Und diese Abklärung sei noch nicht ganz abgeschlossen. "Sobald ich die Gewissheit habe, dass die Unternehmen nicht als eigene Insider aufgetreten sind, muss ich davon ausgehen, dass jemand anders die Transaktionen getätigt hat. In diesem Fall wird je nach Marktanalyse ein Strafverfahren eröffnet", sagte der zuständige Bezirksanwalt. Denn Insider sind nicht gleich Insider. Falls nämlich die beteiligten Gesellschaften selbst ihr insiderwissen nutzten, kommt die
Rachat de Globus par Migros: Procédure pénale pour délit d'initié ouverte

Zurich, 22 oct (ATS). Le procureur de district zurichois Daniel Tewlin a ouvert une procédure pénale pour délit d'initié dans le cadre de la reprise de Globus par Migros. "Il ne s'agit pas d'une piste chaude, mais d'une formalité", a-t-il précisé à l'ATS.

L'ouverture d'une enquête est nécessaire pour avoir accès aux documents protégés par le secret bancaire, a ajouté M. Tewlin, confirmant une nouvelle publiée mercredi dans "Finanz und Wirtschaft". Seuls ces documents bancaires indiqueront si les conditions, sévères, du délit d'initié sont remplies. Celles-ci exigent notamment l'établissement d'un lien clair entre donneur et receveur de conseil. Le rachat en juillet de Globus par Migros avait soulevé des vagues à la bourse. L'opération de rachat n'avait visiblement pas surpris tous les opérateurs au même moment, avaient alors estimé les experts. A la veille de son annonce, près de 100 000 options sur le bon de participation Globus ont été traitées. Le volume quotidien des transactions avait été de 13 600 à 21 000 pièces dans les trois journées qui ont précédé.

22.10.1997

© Agence Télégraphique Suisse ATS 1997/1998

Ces dépêches sont uniquement destinées à un usage personnel. La reproduction, la publication, l'archivage des données dans une banque de données, l'utilisation à des fins commerciales ainsi que la transmission à des tiers sont interdits.
**CINDOR**

1. Accepts the user’s fully expressive natural language query in their native language
2. Maps the query into a ‘Conceptual Interlingua’
3. Provides a relevance-ranked list of returned documents, foldered by language
4. Optionally translates the retrieved foreign language documents into the user’s language

*SIGIR Tutorial, 9 September 2001*
Demo... Source Selection

Ask a question:

Collections: (Use ctrl and/or shift to select multiple corpora)

- Schweizerischen Depeschenagentur (1/88 - 12/90)--fr
- Associated Press (2/88 - 12/90)--en
- El Norte (1/92 - 12/92)--es
- Agence France Presse Data (11/19/99 to present)--es
- Nihon Kezai Shimbun News data (12/93 to 11/94)--ja
- Agence France Presse Data (11/19/99 to present)--fr
- Agence France Presse (5/94 - 1/96)--es
- EPO Data (1995)--fr
- EPO Data (1995)--en
- KYODO News Service/Dow Jones Telerate (11/94 to 6/95)--ja
- Agence France Presse Data (11/19/99 to present)--en

Language codes:
- en = English
- fr = French
- es = Spanish
- ja = Japanese
### Demo... Enter Query

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
</table>
| 2    | **Date Range:**  
|      | From: Jan 1, 1960  
|      | To: Dec 9, 1999  
| 3    | Ask Question in: English  
| 4    | Documents retrieved per language: 100  
| 5    | Automatically translate all results:  
| 6    | Enter a question in the form of a sentence, capitalizing all proper nouns.  
|      | workplace productivity  
| 7    | Perform Search  |
**Demo... View Results**

Your query: workplace productivity

View results in: Japanese  English  French  Spanish

Translate headlines to: English

<table>
<thead>
<tr>
<th>Rank</th>
<th>Result</th>
<th>Date:</th>
<th>Source: Nihon Kezai Shimbun News data (12/93 to 11/94)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>労働生産性9・2年の伸び、先進国中最低に——競争力の弱体化浮き彫り。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>中部生産性本部会長小川進氏——モノづくり以外でも（ロビィ）</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>データで見る「常識のウソ」——常識は生産性高く効率的（麺れ製造業）</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>「生産性向上」への注文（社説）</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Your query:** workplace productivity

**View results in:** Japanese  English  French  Spanish

<table>
<thead>
<tr>
<th>Rank</th>
<th>Result</th>
</tr>
</thead>
</table>
| 1    | Extension of labor productivity 92 year, lowest in advanced nation -- weakening relief of competitiveness.  
*Date:* September 15 of 1994  
*Source:* Nihon Keizai Shimbun News data (12/93 to 11/94) |
| 2    | Shin Ogawa main section chief of Chubu productivity -- mono it makes, at in addition to (lobby)  
*Date:* June 16 of 1994  
*Source:* Nihon Keizai Shimbun News data (12/93 to 11/94) |
| 3    | You see with the data, " lie of common sense " -- sensible 3 productivity it is high efficient (àS, è manufacturing industry)  
*Date:* January 11 of 1994  
*Source:* Nihon Keizai Shimbun News data (12/93 to 11/94) |
| 4    | " Productivity improvement " to order (editorial)  
*Date:* February 2 of 1994  
*Source:* Nihon Keizai Shimbun News data (12/93 to 11/94) |
| 5    | Compatibility with the protection of the environment in theme, Kansai productivity headquarters Mori new chairman interview  
*Date:* March 18 of 1994  
*Source:* Nihon Keizai Shimbun News data (12/93 to 11/94) |
| 6    | Japan productivity center and social economic national meeting, April integration decision -- " reorganization necessity of Kamei financial world ".  
*Date:* February 2 of 1994  
*Source:* Nihon Keizai Shimbun News data (12/93 to 11/94) |
労働生産性92年の伸び、先進国中最低に——労働力の弱体化浮き彫り。

Date: 1994年9月15日
Source: Nihon Kezai Shimbun

民間調査研究機関である社会経済生産性本部が十四日発表した「一九九二年の労働生産性の国際比較」によると、日本の生産性の伸び（前年比増加率）は○・一％で、先進国（十二カ国）中最低水準となった。景気後退に加え、バブル期の大量採用などが生産性的伸びを抑えた。日本の生産性の順位は前年の七位から八位に下げ、日本の競争力の弱体化が浮き彫りになった。

生産性本部は「生産性を高めるにはリストラ（事業の再構築）だけではなく、競争制限的な保護政策の真価など競争力の推進が不可欠」と説明している。

生産性本部が算出している労働生産性は、国内総生産（GDP）など各国の「付加価値」に賃賃力平価を掛けてものを、就労者数・時間で割ったもので経済成長が
7. Issues we haven’t covered

- Presentation of results
  - merging and ranking over different collections/different languages
  - multilingual summarization
  - document translation
- Users increasingly want results in language of query
Issues we haven’t covered

- Interface Issues
  - Human-computer interaction in a multilingual environment not well understood
  - Need user requirement studies for multilingual information access
  - Special UI issues for multilingual multimedia retrieval
Issues we haven’t covered

- Metadata Issues
  - Internationalized metadata schema; e.g. Dublin Core
  - Multilingual indexing languages, i.e. controlled vocabularies, thesauri, lexicons...
  - Different vocabulary sets for different fields, e.g. dates, people and institutions, places, subject categories, etc.
Remaining Challenges – The Future

- Understand user needs – use cases
- Need to accommodate variances in users
  - Language skills
  - Task at hand
- Need to involve user at many stages in process
  - providing feedback
Remaining Challenges – The Future

- Multilingual querying – controlled query expansion in a multilingual environment
- Document clustering across languages
- Document summarization, including summaries generated from multiple documents in different languages
- Visualization of retrieved sets (or general sets) of documents

SIGIR Tutorial, 9 September 2001
Remaining Challenges – The Future

- Resources – understand the relationship between linguistic resources and MLIA effectiveness
- Multilingual Multimedia – Expand to speech and video retrieval across languages
- Cross-disciplinary communication & education!
Some Useful URLs

- W3C - WINTER - http://www.w3.org/International/
- Multilingual Metadata - http://purl.org/DC/groups/languages.htm
- EC - Multilingual Information Society - http://www2.echo.lu/mlis/

SIGIR Tutorial, 9 September 2001
Our appreciation to...

- Judith Klavans  Columbia University
- Peter Schäuble  Eurospider
- Doug Oard  University of Maryland
- Steve McClure  IDC
- …for contributions of material

- DELOS and IFLA for the invitation!

SIGIR Tutorial, 9 September 2001