Panel on CLIR

Noriko Kando
National Institute of Informatics (NII), Japan
(formerly NACSIS)
http://www.rd.nacsis.ac.jp/~ntcadm/
CLEF 2000
September 21, 2000

Strong Aspects of CLEF

Collaboration of multiple site: a good model of distributed organization of CLIR evaluation

Comparable Corpus: more natural than parallel corpus,

Multilingual Runs with Native Speakers Judgments

Strong Needs for CLIR

Strong Aspects of CLEF(2)

Tradition of Working Together

Do monolingual IR with each language

Organization & Funding

Strong Language Abilities: direct translation from source language is available

Common (Similar) Character set

The layers of CLIR technologies

pragmatic layer: cultural & social aspects, convention

semantic layer: concept mapping

lexical layer: language identify, indexing

symbol layer: character codes

physical layer: network

For the next CLEF

Keep the strong points! CLEF is one of the ideal environments of real CLIR research

if we can think about a new task, **"post-retreival processing"** "We will need techniques to organize and present the information in the form that is at the same time understandable and immediately usable" (p.xvii, Natural Language Information Retrieval, Kluwer, 1999) A ranked list of target language documents is not understandable and immediately usable for source language speaking users.

More realistic: I suppose among the European Language speaking countries, sometimes including "compornential multilingual text", i.e., a document containing more than one language, is more natural and realistic. Trying such kind of documents would be another possible direction.

"post-retreival processing"

-techniques to support to make retrieved documents immediately usable

For example,
translation of the retrieved docs, CL skimming, CL automatic summarization, CL extracting answer passages, clustering the retrieved docs, visualization, text mining, comparison of the contents in the docs etc.

-Such processing needs combination of IR + NLP. In the CLEF community, many groups have strong NLP background. It would be suitable community to try such challenging new tasks.

Taxonomy of ML corpus

Parallel:

Comparable :

Non-comparable :

Componential : A component in a document is represented in different language. Ex. Paired document, English abstract in other language documents (We can easily find in non-English speaking countries)

Lexical : A word from different language involved in the texts. -- in Asian countries, Queries can be also lexically multilingual.

Open Questions

What we can learn from the evaluation? Can we learn what we should change to improve the R/P of whole IR process? How the system or algorithms we should modify?

Can we evaluate the component (module) as well as IR process whole? How can we provide for baseline for each component (module)?

How can we construct the reusable test collection for new technologies related to post-retrieval processing?