A Scalable Cross-Language Metasearch Architecture* for Multilingual Information Access on the Web

Yoshihiko Hayashi, Genichiro Kikui, Toshiaki Iwadera
NTT Cyberspace Laboratories
1-1 Hikari-no-Oka, Yokosuka, 239-0847. Japan
E-mail: hayashi@nttnly.isl.ntt.co.jp

Abstract

This position paper for the special session on "Multilingual Information Access" comprises of three parts. The first part reviews possible demands for Multilingual Information Access (hereafter, MLIA) on the Web, and examines required technical elements. Among those, we, in the second part, focus on Cross-Language Information Retrieval (hereafter, CLIR), particularly a scalable architecture which enables CLIR in a number of language combinations. Such a distributed architecture developed around XIRCH project (an international joint experimental project currently involves NTT, KRDL, and KAIST) is then described in a certain detail. The final part discusses some NLP/MT related issues associated with such a CLIR architecture.

1 Introduction

A survey report titled "Web Languages Hit Parade" published in 1997 [Babel, 1997] estimated that more than 80% of the Web home pages are written in English, followed by German, Japanese, French, and so on. While the estimation method described was relatively rough, the figure seems to be reasonable.

What about the current and/or the future situation? To our knowledge, there is no comparable information source which gives the current estimation. An international online marketing company however makes another type of estimation, saying that the number of people on the net whose native language is other than English will overcome that of English-native people by Y2K [Global Reach, 1999]. As the Web is expected to grow continuously over the globe, this estimation would not be very misdirected. Naturally with this estimation, networked resources written in languages other than English are supposed to increase as well as demands for seeking those resources using a variety of languages.

As discussed in the next section. CLIR definitely plays an essential role in MLIA. Introduced below is the "Grand Challenge" statement[1] discussed in the wrap-up panel session of 1997 AAAI Spring Symposium on Cross-Language Text and Speech Retrieval[2] .

> *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.*

This is, as seen, a general and strong statement which enoughly covers many research issues on intelligent multimedia/multilingual information access technichs/systems. Goals of MLIA are naturally implied by this statement. In the next section, we try to figure out technical ingredients relevant to MLIA on the Web with this statement in mind.

2 Ingredients for MLIA on the Web

The term "Information Access" has been used in various ways. Hearst, for example, sees that the goal of information access is helping users to find documents that satisfy their information needs [Hearst, 1999]; using the term almost same as traditional "Information Retrieval." On the other hand, Schäuble defines information access more broadly as follows [Schäuble, 1998].

* The XIRCH project described in this paper is joint work with **MunKew Leong** (KRDL, Singapore), and **Key-Sun Choi** (KAIST, Korea).

[1] The wording was arranged by David Hull: accessible at http://www.ee.umd.edu/medlab/filter/sss/panel.txt.

[2] http://www.ee.umd.edu/medlab/filter/sss/

Information Access = Information Retrieval + Information Extraction + Hypermedia Browsing + Document Visualization

We, in this paper, take a somewhat middle road, and expand it to multilingual version. That is, we see "MLIA on the Web" casually as follows.

MLIA on the Web = Cross-Language Information Retrieval + Language-Oriented Navigation + Multilingual Document Browsing

The assumption here is users on the Web are fluent in their native languages possibly with few other languages which they have some knowledge.

Cross-Language Information Retrieval: The goal of CLIR is to allow a user to issue queries in one language of her choice, and retrieve documents written in other languages. Here we assume that the user can input her query to a search engine/service with appropriate input method[3], for example Microsoft Global IME. In order to achieve the goal, researches with substantial volume have been done as surveyed in [Oard, 1997] and compiled in [Grefenstette, 1998]. CLIR is naturally associated with MT (Machine Translation) and IR (Information Retrieval).

Language-Oriented Navigation: Usually search results are presented to a user as a ranked list of relevant documents (Web pages), sometimes with metadata-like information (data type/size, indexed date, etc.) and/or mechanically generated summary of the page. The user has to make a decision on which pages to actually visit with the limited information. If the presented information, particularly summaries which may be tremendously useful for the navigation process starting from the retrieved (and yet unvisited) page, is provided in the language of the user's choice, it would be of a great help for her. IE (Information Extraction) and/or Text Summarization coupled with some MT technique would be a new area to be explored for the purpose.

Multilingual Document Browsing: After some relevant pages are found, the user would then read through the pages. If the pages are properly translated and presented to her, it would be nice. However she still might want to read through the pages in the original languages, if she has some knowledge about the languages. In this sense, multilingual browser is also an important element for MLIA. Popular browsers, such as Netscape's and Microsoft's, are already capable of displaying pages in many languages provided proper font sets have been installed. Even 'language-mixture' pages could be properly displayed, if they are encoded with Unicode[4] . We however are forced to manually select appropriate font set frequently in the course of browsing. A browser which does not bother users with such annoying thing should be able to correctly select a font set according to the page viewed. Automatic LI (Language Identification) technique with character encoding system recognition is necessary for this purpose. Our research group has developed such a browser [Watanabe. 1999] based on LI technique described in [Kikui. 1996].

In addition to these, interactive interfaces are crucially important in order to integrate those technical elements properly. We however would like to focus on the first element, CLIR, in this paper, since we see search results as starting points from which the user's information exploration in the Web information space is initiated.

3 A Scalable Cross-Language Metasearch Architecture

3.1 Beyond CLIR between a Language Pair

CLIR has been more focused recently as typically shown in the situation where recent TREC conferences[5] have cross-language tracks. Major research issues in the field so far have flavor of "CL extensions" to traditional IR; disambiguation in query translation, query expansion with pseudo-relevance feedback, and utilization of linguistic resources such as bilingual corpora. As seen in these research topics, little emphasis has been on Cross-Languages IR. For example, architectures which enable CLIR in a number of language combinations have been examined only by a few groups [Powell, 1998], [Picchi, 1998].

3.2 Metasearching

Considering a comprehensive Web search service, aparting from Multilingual/Cross-Language issues, metasearching is considered as a smart approach, given the situation where even major search services, such as AltaVista or Lycos, cover only some portion of the entire web [Lawrence, 1998]. A metasearcher is free from gathering and indexing huge amount of Web pages by itself, rather it accesses to adequate search services having their own indexes. Scalability thus can be achieved ideally by this approach, given a situation where the metasearcher can have accesses to the search services/engines. An initial attempt of the metasearching on the Web dates to as early as 1995 as represented by MetaCrawler [Selberg, 1995].

[3] Input methods should also be itemized as a technical issue especially for languages that are not major.

[4] http://www.unicode.org/

[5] http://trec.nist.gov/

A naive metasearcher however may suffer from several problems when it tries to answer a user's query; three major problems [Gravano, 1997] are listed as follows.

1. How to choose the best search site/engine to evaluate the query?
2. How to make each site properly evaluate the given query?
3. How to merge the results from these sites?

To solve the first problem, the metasearcher should have knowledge about search sites to which it has accesses. As query syntax can differ from site to site, the metasearcher has to mediate the differences in order to accommodate the second problem. The third problem arose, because ranking algorithms employed by the sites are different and usually kept secret.

STARTS [Gravano, 1997] is a protocol proposal to address these problems with metasearching[6] .

While STARTS shares goals with Z39.50, it was intended to be simpler than the ANSI standard. Truly, STARTS is an ambitious proposal to these problems, and may succeed if major search services adopt this protocol. However, it completely lacks considerations for cross-language searches in multilingual environments, which should be crucially important on the current/future Web. As will be seen in the following paragraphs, our architecture is based on STARTS.

3.3 Overall Architecture

Figure 1 illustrates our overall architecture in general [Iwadera, 1998]. The following elements deployed on the Web are ingredients of the architecture. They communicate each other by utilizing HTTP.

CLMS: a CLMS (Cross-Language MetaSearcher) provides users with its own search services (not necessarily cross-language ones, but we intend them). It has knowledge about capabilities and characteristics of the search sites to which it accesses through the metadata about the sites. It answers user's information request by consulting appropriate search sites, given the query condition, using the *distributed cross-language information retrieval protocol* named XIRCH, which will be introduced in the next section.

Search Engine/Site: a search site/engine has its own text retrieval engine with an internal query/result syntax. The protocol handler, which adheres to the retrieval engine, mediates syntax gap between external ones (defined by XIRCH protocol) and internal ones.

NLP Server: a NLP (Natural Language Processing) server is responsible for language-oriented processing, such as term extraction for language-X, and/or translation between language-X and language-Y. It can be placed anywhere on the Web in principle. However it will be naturally realized close to a CLMS and/or a search site. Note that open protocols between each NLP server and other elements are still difficult to be defined, because language or language-pair dependent issues must be addressed. (This issue will be revisited in this paper.)

3.4 XIRCH Protocol Proposal

As in STARTS, the protocol defines possible interactions between a CLMS and the associated search sites. The interactions fall into two classes. Major interactions are naturally for querying: query and the results format as well as other requirements especially for language conversion are defined. Another class of interactions are associated with metadata, which describes capability and characteristics of a search site. Note that XIRCH protocol can be understood as slight extension of STARTS with the standard attribute set; means that the most of the extensions are realized by newly introduced attributes and modifiers.

Query and the Results

Let us start with an example of query shown in Figure 2, which is represented in SOIF[7] . As seen in the figure, we support both "Filter expression" and "Ranking expression" as originally in STARTS; the former designates Boolean condition, and the latter is used for ranking relevant documents. The query displayed in the figure says that a relevant document must satisfy the following.

1. The language of the document is English.
2. The document is located in the US domain.
3. English tokens derived from Japanese phrase " ワイン工場の歴史 " *(history of winery)* appear in the title.
4. English tokens derived from Japanese phrase (actually it is a term) "カリフォルニア" *(California)* appear in the title.

It also says that the relevant documents containing English tokens derived from Japanese phrase "ロゼ " *(rose)* should be ranked higher. As seen in this example, modifiers **Tokenize** and **Translate** are introduced to designate NLP functions to be performed.

[6] In [Gravano, 1997], these problems are named "the source selection" problem, "the query-language" problem. and "the rank-merging" problem respectively.

[7] Query data object like this is usually constructed by CLMS, which converts user's information requirement to the query form. Using SOIF (Semantic Object Interchange Format) is not the protocol specification.

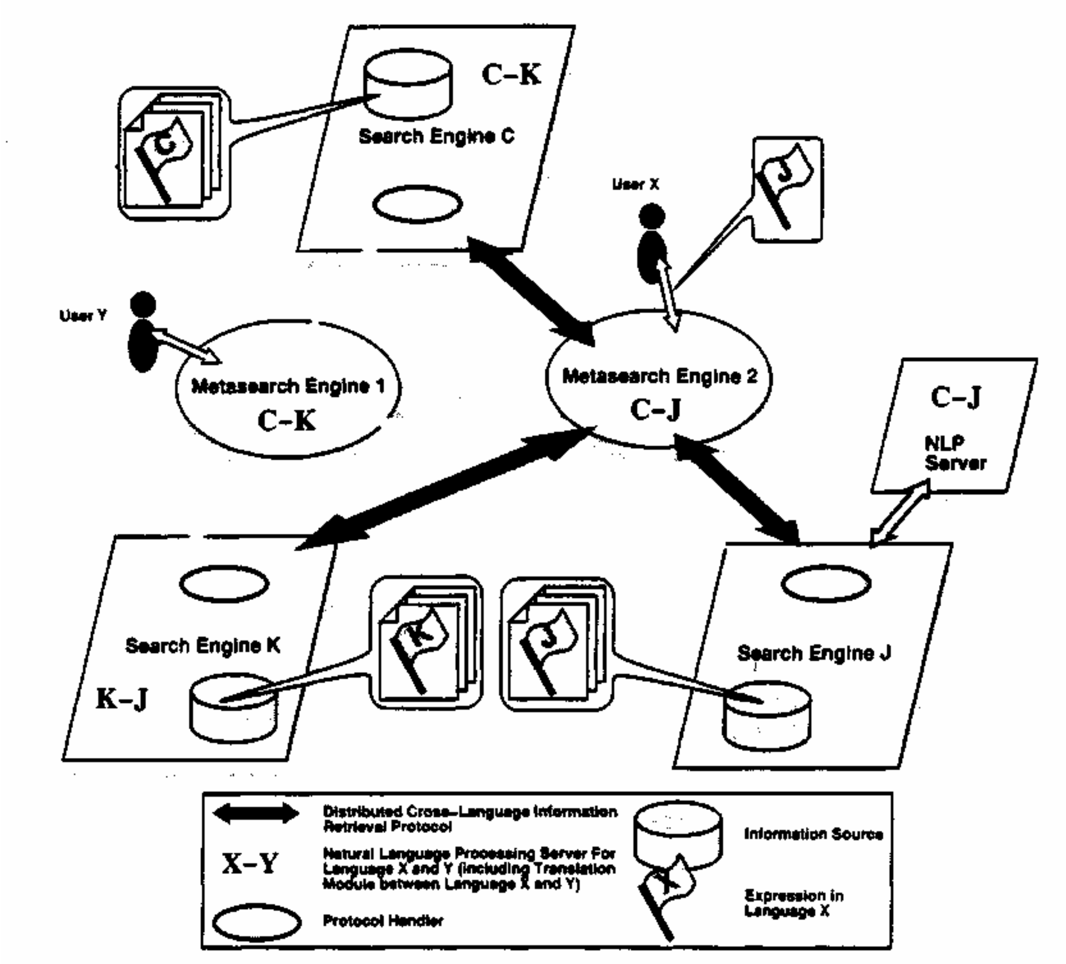


Figure 1: The overall architecture.

Furthermore, requirements for the result presentation, such as **DocumentTranslationTargetLanguage** is also incorporated as an attribute.

A part of the associated results slightly simplified is exemplified in Figure 3: the translated title is given as requested, along with the original title. Information about the original document such as **OriginalEncoding and TermStats** are included. Information like this may ease the user's browsing action which follows the search, when utilized nicely by the CLMS.

In implementation, these data objects are encoded using Unicode (UTF-8 encoded), while CLMSs and search sites can use any internal character encoding system. Details of the XIRCH protocol and the related information will be accessible at our project site[8] .

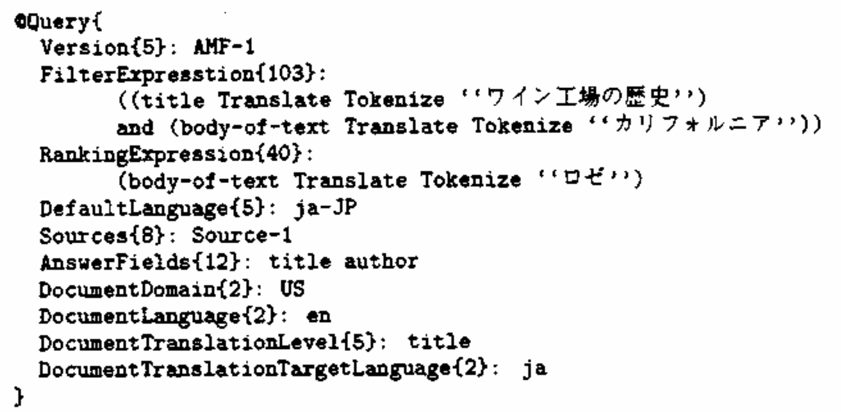
Metadata

In the architecture, each search site has to declare its characteristics; that is, summary of the content, supported search attributes/modifiers and NLP related capabilities should be acquired by the CLMSs in advance to the service time. Knowing such information, a CLMS manages to appropriately choose one or more search site against the given query, and merge results from these sites properly. In our protocol, especially "Source metadata attributes" in STARTS are extended as exemplified in Figure 4 (again in SOIF), As shown in the figure, NLP related attributes and modifiers, such as **TranslationTargetLanguageSupported** and **QueryExpansionMethodSupported** are incorporated in order to support cross-language and/or advanced searches.

3.5 Revisits three problems

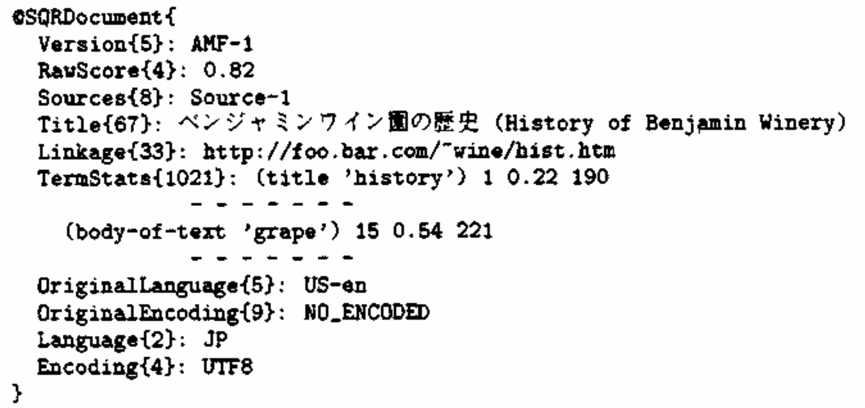
We quickly review how our architecture solves (or does not solve) the problems introduced by STARTS. The first problem "the source selection problem" cannot be solved without metadata provided by a search sites. STARTS provides necessary framework, but only in basic level. However we gave a clear way to utilize the framework; language(s) of documents indexed at each site is mainly used to characterize a search site. It is quite natural to assume language-specific search sites are available on the Web. The second problem

[8] http://titan.isl.ntt.co.jp/xirch/



```
@Query{
  Version{5}: AMF-1
  FilterExpresstion{103}:
        ((title Translate Tokenize ''ワイン工場の歴史'')
        and (body-of-text Translate Tokenize ''カリフォルニア''))
  RankingExpression{40}:
        (body-of-text Translate Tokenize ''ロゼ'')
  DefaultLanguage{5}: ja-JP
  Sources{8}: Source-1
  AnswerFields{12}: title author
  DocumentDomain{2}: US
  DocumentLanguage{2}: en
  DocumentTranslationLevel{5}: title
  DocumentTranslationTargetLanguage{2}: ja
}
```

Figure 2: An example of query in SOIF format.



```
@SQRDocument{
  Version{5}: AMF-1
  RawScore{4}: 0.82
  Sources{8}: Source-1
  Title{67}: ペンジャミンワイン園の歴史 (History of Benjamin Winery)
  Linkage{33}: http://foo.bar.com/~wine/hist.htm
  TermStats{1021}: (title 'history') 1 0.22 190
                   - - - - - - -
    (body-of-text 'grape') 15 0.54 221
                   - - - - - - -
  OriginalLanguage{5}: US-en
  OriginalEncoding{9}: NO_ENCODED
  Language{2}: JP
  Encoding{4}: UTF8
}
```

Figure 3: An example of the search results.

"the query-language problem" can be solved even for cross-language querying with our extensions, if we can utilize appropriate NLP servers with MT functionalities. The third problem "the rank merge problem" is still a further issue especially in cross-language settings, as described in several reports[9] from TREC-7 Cross-Language Track.

4 XIRCH as Joint Project

4.1 Project Overview

"XIRCH" is also used for designating an international joint project, which currently involves NTT[10] , KRDL[11] and KAIST[12] . The group has jointly reviewed the first version protocol initiated by NTT. Each organization is now individually implementing its own CLMS and search site; means that each of us can develop necessary elements as they like and/or constrained by usable linguistic resources/tools. For example, NTT is developing a search site which primarily gathers pages in Japanese, and a CLMS; both are able to translate queries and page titles in Japanese to English, and vice versa. Single NLP server, deployed on the Web, is employed by both of them. Almost same story may apply to KRDL and KAIST sites. Therefore, as long as the coordinate protocol is properly implemented by the sites, cross-language searches between English, Japanese, Chinese, and Korean will be realized.

4.2 CLMS/NTT: NTT's Metasearcher

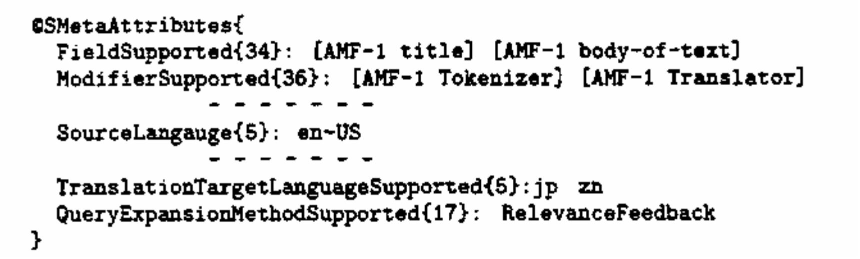
As an example of the being implemented CLMS, NTT's CLMS (CLMS/NTT) is introduced here. Our CLMS is based on TITAN [Hayashi, 1997], which is one of the pioneering Cross-Language search engines on the Web. TITAN accepts queries in Japanese and English, and provides query translation between these

[9] http://trec.nist.gov/pubs/trec7/papers/index.track.html

[10] Nippon Telegraph and Telephone Corporation, Japan

[11] Kent Ridge Digital Laboratories, Singapore; Yes! it is the venue of MT Summit VII.

[12] Korea Advanced Institute of Science and Technology



```
@SMetaAttributes{
  FieldSupported{34}: [AMF-1 title] [AMF-1 body-of-text]
  ModifierSupported{36}: [AMF-1 Tokenizer} [AMF-1 Translator]
                    - - - - - - -
  SourceLangauge{5}: en-US
                    - - - - - - -
  TranslationTargetLanguageSupported{5}:jp zn
  QueryExpansionMethodSupported{17}: RelevanceFeedback
}
```

Figure 4: An example of the metadata for attributes.

two languages. TITAN also provides several search options, such as restrictions by URL domain and language of the document[13] . Functions being implemented in CLMS/NTT are naturally inherited from TITAN.

A preliminary input form of the CLMS/NTT is shown in Figure 5, and the associated results screen is shown in Figure 6. Note that the input query is interpreted as "Filter expression" in this input form. In the current test implementation, we utilize free WAIS-sf[14] as the internal text retrieval engine, still ranked results can be obtained even with the filter query. As shown in the Figure 5, optional search conditions using the URL domain and/or the languages of the page can be specified. Query expansion with thesauri is also planned.

As this input form is written using Unicode (UTF-8). the input query is transferred to the CLMS as UTF-8 encoded. Therefore, in general, we need to know the language of the query string. We can specify the language manually, or leave the decision to the CLMS as shown in this example. Currently, search sites are manually chosen: in this example, the screen snapshot was captured during the internal debugging process, NTT and NTTKR. which is a simulated Korean site by NTT. are chosen. While query translation condition can be manually specified. "Automatic Selection" is also provided. This automatic mode implies something like "best-effort" mode as seen in the results screen. Note that the input query phrase is "ソウルのホテルの予約 " in Japanese, whose equivalent in English would be "reservation of a hotel in Seoul."

The results screen consists of four frames; of which the top one is just displays our logo. In the left frame, query form sent to the sites are shown. The one sent to the NTT's site is in Japanese; three tokens ("ソウル", "ホテル", and "予約") are extracted from the query phrase. On the other hand, the one sent to the NTTKR's site is in English. This is because that the CLMS has translating capability between Japanese and English, but not between Japanese and Korean,

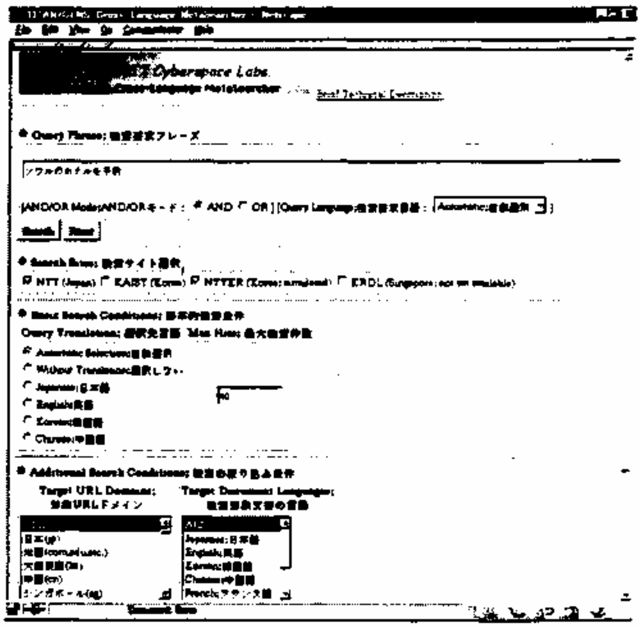


Figure 5: CLMS/NTT Input Form (preliminary

it just tokenizes the query in Japanese and translates the extracted tokens to English ("Seoul", "hotel", and "reservation").

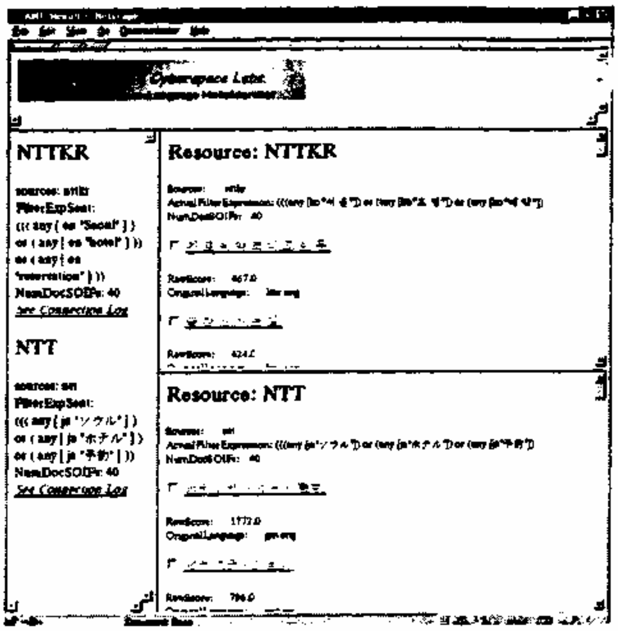
How these query forms are processed in the selected sites are displayed in the two frames right, as well as the search results. At the NTT site (the lower frame), the sent query form is used as is and the search results are displayed with check boxes for the relevance feedback. On the other hand, at the NTTKR site (the upper frame), the English tokens contained in the received query form are translated into Korean. The actually used query form is represented with the attribute **ActualFilterExpression.**

5 Discussions

Query translation is not the only approach to CLIR; documents database could be translated in advance, and matched against queries. We. however in this section, limit our discussions to query-translation based

[13] The Internet robot utilized by TITAN implements language identification algorithm [Kikui, 1996], and adds language tags to gathered pages.

[14] http://ls6-www.cs.uni-dortmund.de/



NTTKR

sources: nttkr
FilterExpSent:
(((any { en "Seoul" })
or (any { en "hotel" }))
or (any { en "reservation" }))
NumDocSOIFs: 40
See Connection Log

NTT

sources: ntt
FilterExpSent:
(((any { ja "ソウル" })
or (any { ja "ホテル" })
or (any { ja "予約" }))
NumDocSOIFs: 40
See Connection Log

Resource: NTTKR

Source: nttkr
Actual Filter Expression: [illegible]
NumDocSOIFs: 40

RawScore: 467.0
OriginalLanguage: [illegible]

RawScore: 424.[illegible]

Resource: NTT

Source: ntt
Actual Filter Expression: (((any {ja "ソウル"}) or (any {ja "ホテル"}) or (any {ja "予約"})
NumDocSOIFs: 40

RawScore: 1772.0
OriginalLanguage: [illegible]

RawScore: 796.0

Figure 6: CLMS/NTT Results Screen (preliminary).

systems with particular attention on distributed environments as presented in our architecture. These discussions may include further issues to be considered in the next step of our project.

"Term" as a Basic Unit: We consider "term" as a basic unit in indexing/querying. Terms are usually extracted from query phrases and documents by language-dependent processes, such as tokenization, morphological analysis, stop-words deletion, stemming, and so on. In a environment where many sites participate with their own NLP tools, it is rather difficult to define stable definition for terms, even single language is considered. In our case, as multiple language are considered, it would be more chaotic. In principle, the protocol should be defined to allow a processing site (CLMS or Search site) to remedy mismatches as it likes. Particularly, proper recognition of compounds are important to improve retrieval effectiveness.

The CLIR Issues: As in the usual CLIR, ambiguities in translation and existence of untranslatable terms are big problems to be addressed. From the protocol viewpoint, how to convey useful contextual information for disambiguation is a problem to be addressed.

Multiple Translation Steps: We do not intend to consider English as the pivot language, while currently it is so. For example, Japanese query entered to NTT's CLMS may be first translated into English, and sent to KAIST search site: where the English query would be translated into Korean. EuroSearch[15] explicitly considers English as the pivot [Picchi,1998] in order to welcome a new language to the federation, while in our architecture any new language pair can be incorporated, if such an element joins the environment. Some kind of interlingua is of course desirable, but might be possible only for limited domain. In that case, something like "domain-indicator" should be included into common vocabulary of the protocol.

NLP Protocol: Currently. NLP servers are not disclosed to even other participants of the project. That is, for example, NTT's NLP server is only utilized by NTT's CLMS and the search site, even it is implemented as a possibly accessible Web server. To make such NLP servers open to other parties and usable, "standard NLP protocol" should be developed while considering some standard APIs necessary for CLIR. Such a NLP server is not necessarily open to the project participants: it might open a door to NLP/MT server business on the Web.

The Rank Merge Problem: The monolingual rank merge problem can be solved with STARTS protocol, because most information necessary for conventional document ranking can be obtained through the data under the protocol. However for the CLIR, we will have to solve *another* rank merge problem, as in the CL track of TREC. We somehow must adjust scores from search sites with different languages, or recompute the document scores with some unique major. Revision of the protocol may be necessary for efficient merging. It is however still unclear whether merging is necessary in actual CLIR applications, not for the evaluations like TREC.

All the items listed above are possible research issues toward the next version of the architecture. Topics such as retrieval of more structured documents in XML, and use of metadata description, if significantly popularized, to improve the query planning by the metasearchers are also relevant as for the monolingual case.

Along with these technical issues, we, in this arena, should demonstrate some useful applications, not necessarily be the AltaVista-like Web search, in order to facilitate sound progress of the field.

6 Concluding Remarks

We in this paper focused on our project toward MLIA on the Web; a scalable CL metasearch architecture. Here the scalability primarily means the number of supported languages can be increased relatively easily. This nature comes from our architecture: a site

[15] http://eurosearch.iol.it/

dedicated to a new language can be incorporated to the federation, if it implements the protocol. Or. an existing site can support new languages, if appropriate language-dependent NLP servers deployed somewhere on the Web are accessible. The features necessary for cross-language searching are realized by slight extensions to STARTS: means that it preserves benefits of STARTS, such as simplicity. The internal joint project XIRCH is now under way, and the first version of the prototype service will be launched very soon.

Several NLP/MT related issues with the project were also introduced. These technical issues should be better solved by a group of people with different language background. In this sense, we will welcome offers for joining the project from the world, as well as comments and suggestions.

Acknowledgments

The authors wish to thank to MunKew Long (KRDL) and Key-Sun Choi (KAIST) for their continuous cooperation in the joint project.

References

[Babel. 1997] Babel Project. Web Language Hit Parade. http://www.isoc.org:8080/palmares.html

[Global Reach. 1999] Global Reach. Global Internet Statistics (by Language). http://www.euromktg.com/globstats/

[Grefenstette. 1998] Grefenstette, G. (editor). Cross-Language Information Retrieval. *Kluwer Academic Publishers,* 1998.

[Gravano. 1997] Gravano, E., Chang, K., Garcia-Molina, H., Lagoze, C., and Paepcke, A. STARTS: Stanford Protocol Proposal for Internet Retrieval and Search. Technical Report, Stanford University, 1997. [16]

[Hayashi. 1997] Hayashi, Y. Kikui, G., and Susaki, S. TITAN: A Cross-Linguistic Search Engine for the WWW. in *Cross-Language Speech and Text Retrieval,* AAAI Technical Report SS-97-05, 1997.

[Hearst, 1999] Hearst. M. Untangling Text Data Mining. *Proc. of the 37th Annual Meeting of the ACL.* pp.3-10. 1999.

[Iwadera. 1998] Iwadera, T., Hayashi, Y., Kikui, G., Obashi, Y., Leong, M., and Choi, K. An Distributed Cross-Language Information Retrieval Architecture, (in Japanese) *IPSJ SIG Notes NLP.* Vol.98, No.81, pp.63-70, 1998.

[Kikui, 1996] Kikui, G. Identifying the Coding System and Language of On-line Documents on the Internet *Proc. of the 16th COLING,* pp.652-658, 1996.

[Lawrence. 1998] Lawrence, S. and Giles. C.L Searching the World Wide Web. *Science,* Vol.280, No. 3, pp.98-100, 1998.

[Oard, 1997] Oard, D. Serving Users in Many Languages, Cross-Language Information Retrieval for Digital Libraries. *D-Lib Magazine,* December 1997. [17]

[Picchi, 1998] Picchi, E. and Peters, C Exploiting Lexical Resources and Linguistic Tools in Cross-Language Information Retrieval: the EuroSearch approach. In *First International Conference on Language Resource and Evaluation,* 1998.[18]

[Powell, 1998] Powell, J. and Fox, E.A. Multilingual Federated Searching Across Heterogeneous Collections. *D-Lib Magazine,* September 1998. [19]

[Schäuble, 1998] Schäuble, P. Multilingual Information Access (MLIA). http://www.eurospider.ch/MLIA98/index.htm.

[Selberg, 1995] Selberg, E. and Etzioni, O. Multi-service search and comparison using the MetaCrawler. *Proc. of the 4th International World Wide Web Conference,* 1995. [20]

[Watanabe, 1999] Watanabe, K. and Kikui, G. Smooth Surfing Beyond the Language Boundary, A multilingual browser that identifies character-coding systems. *Proc. of the 5th NLPRS,* to appear, 1999.

[16] http://www-db.stanford.edu/ gravano/starts.html

[17] http://www.dlib.org/dlib/december97/oard/12oard.html

[18] http://eurosearch.iol.it/papers.html

[19] http://www.dlib.org/dlib/september98/powell/09powell.html

[20] http://www.w3.org/Conferences/WWW4/Papers/169/