Machine Translation for Information Access across the Language Barrier: the MuST System

Chin-Yew Lin
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292-6695, USA
Tel: 1-310-822-1511
E-mail: cyl@isi.edu

Abstract

In this paper we describe the design and implementation of MuST, a multilingual information retrieval, summarization, and translation system. MuST integrates machine translation and other text processing services to enable users to perform cross-language information retrieval using available search services such as commercial Internet search engines. To handle non-standard languages, a new Internet indexing agent can be deployed, specialized local search services can be built, and shallow MT can be added to provide useful functionality. A case study of augmenting MuST with Indonesian is included. MuST adopts ubiquitous web browsers as its primary user interface, and provides tightly integrated automated shallow translation and user biased summarization to help users quickly judge the relevance of documents.

1 Introduction

In the past, machine translation systems were most often used in standalone mode, coupled at most with text processing systems. But in the Internet world we live in, there is an increasing need to couple MT engines to other text processing services such as text summarization, information retrieval, and web access. In particular, with the increasing amount of online information and the rapid growth of the number of non-English speaking Internet hosts, it is becoming increasingly important to offer users universal access to valuable information resources in difference languages. The European Multilingual Information Retrieval (EMIR) project [5], the MULINEX project [4], the TwentyOne Project [9], and the cross-language retrieval track in the TREC conference [8] all reflect people's interest in providing interoperability among different language processing environments and multilingual information retrieval.

What needs to be done to link MT and IR to create multilingual information retrieval (MLIR)? The problem of language encoding and display used to be an issue, but is now less daunting with the advent of Unicode and web browsers such as Microsoft Internet Explorer and Netscape Navigator. This allows us to focus on the two difficult problems of query translation and result translation.

One way to tackle the multilingual information retrieval problem is to translate all the target language text into source language text and then perform monolingual search on the translated text. Oard [20] reports that machine translation (MT) based document translation outperforms MT based query translation. However, translation of 251.840 documents from German to English takes about 10 machine-months on a mix of SPARC 20, SPARC 5. and Ultra SPARC 1 using the Logos translation engine[1]. Without better machines and high speed/quality MT. we can rule out the practical application of this approach for the web. The document translation approach is even more impractical when fully multilingual information systems were considered, because document translation has to be conducted on each language pair in such a system.

We therefore adopt the query translation approach. To translate user queries from source languages to target languages, we need multilingual/bilingual transfer dictionaries or corpora (parallel or non-parallel) [18]. This task includes the challenges of disambiguating senses of the translated queries and distributing the weighting for each translation candidate in a vector space model or a probabilistic retrieval model [7]. Our system MuST currently uses all the possible translations for each content word and performs no weight adjustment. Research on these specific issues will be the primary focus in the second phase of the MuST project. We currently concentrate on system

[1] Logos Corporation, 111 Howard Boulevard, Suite 214, Mount Arlington. NJ 07856, USA.

design and integration, which we describe in this paper.

With respect to the retrieved documents, the issue is whether they can be presented in the original language. Probably not: most users do not master many foreign languages. Oard [20] has argued for other applications, such as image retrieval of foreign sources where translation may not necessary. Here the retrieved image automatically explains itself. However, this will not always be the case: a caption for an image has a very good reason for existing. A rough translation probably suits the case better. Still, the question remains: of what quality? High quality is not always possible, and translation speed is also a concern. Therefore, shallow translation of browsing quality seems a more practical alternative.

To help overcome the problem of speed, one can consider producing only a translated summary of the foreign original text. Tombros & Sanderson [26] and Mani et al. [17] have separately reported that user biased summaries can improve monolingual retrieval performance. We believe translated summaries can also help users in a similar way. But what is the cost of developing a robust and portable multilingual text summarizer? Is this possible?

MuST is a prototype multilingual information retrieval, summarization, and translation system, in which we have tried to identify reasonable solutions for the questions mentioned above. Although the main focus of this paper is on system design and implementation, we believe that understanding of these issues helps explain many design decisions we have made. We describe the architecture and modules of MuST in the next section. We then discuss the issues involved in the implementation of MuST and provide a case study for Bahasa Indonesia, Finally, we conclude with remaining issues and future directions.

2 MuST

The goal of the MuST project is to develop a prototype system to facilitate not only retrieving documents from multilingual collections, but also to summarize and translate the retrieved document into the user's preferred language. We focus on the integration of state-of-the-art technologies, try to identify the critical path of enabling multilingual information access, and propose possible solutions. As far as possible, the system employs existing resources and products, such as the search technologies from MG [25]. America Online (AOL)/Personal Library System (PLS), and online Internet search engines. It incorporates web spider technology enabling users to target their areas and languages of interest. It provides multilingual summarization technology developed at ISI [11] enabling users to quickly judge the relevance of the retrieved documents. It also integrates deep [17] and shallow translation engines for online browsing of foreign language texts. We use the World Wide Web as our multilingual document sources and assume English is the source language. MuST can handle the languages English, Arabic, Japanese, Spanish, and Bahasa Indonesia. We plan to add more languages in the near future.

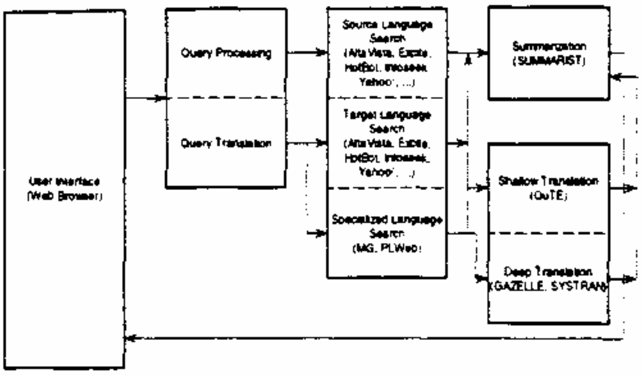


Figure 1. The architecture of MuST.

Figure 1 shows the architecture of MuST. The system consists of five major components: (1) an information retrieval module, (2) a query processing and translation module, (3) a machine translation module, (4) a text summarization module, and (5) a user interface module. We describe these modules in the following sections.

2.1 Information Retrieval Module

The retrieval module is a combination of several monolingual retrieval engines. Each monolingual retrieval engine connects to the query translation module through a wrapper, which converts a standard MuST query into the query language of the specific monolingual retrieval engine. Existing web search engines such AltaVista, Excite, Infoseek, Lycos, and Yahoo, to which MuST is linked, are good examples of monolingual retrieval engines for English, major European languages, and some Asian languages; while Yam of Taiwan [26] is a good example of a localized search engine (Taiwanese Mandarin) that provides Yahoo-like search service in Taiwan.

We expect that most countries will establish their own Yahoo-like search services in the near future. However, some users may want to have their personal search services for special topics such as world country history, gourmet recipes, and art museums. The proliferation of online news groups and virtual communities manifests the needs of specialized search services. The provision of a personal index agent is therefore necessary. MuST uses a free commercial index engine and spider, PLWeb from America Online Inc., to provide this capability.

2.2 Query Processing and Translation Module

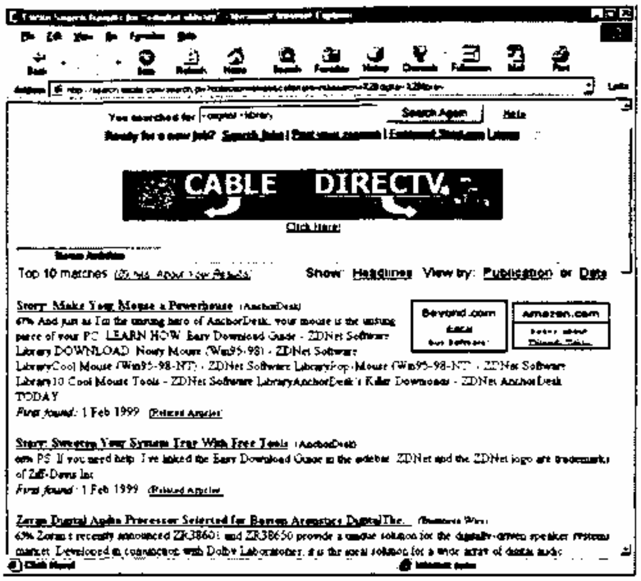
The query processing module determines if the user's query needs to be translated or not. If translation is necessary, it then passes the query to the query translation module. Query expansion is also carried out in this stage, although MuST currently only has limited capability for expansion. We plan to use topic signatures [16] to improve query expansion.

The query translation module translates a user query into the language of the target monolingual information source. A bilingual or multilingual transfer dictionary is required for this step. Using the machine translation module to carry out query translation seems another straightforward and economical solution. However, Oard [19] shows that a sophisticated machine translation system can outperform dictionary-based query translation methods on long queries, but not on short ones; while Ballesteros and Croft [1] demonstrate that combining dictionary-based translation with local feedback before and after translation can boost short and long query performance [1]. Since a high performance machine translation system is not always available and dictionary-based method with sophisticated expansion can perform well, MuST adopts the dictionary-based query translation approach. This also enables greater coverage of more languages.

2.3 Machine Translation Module

Machine translation is usually not considered an integral part of a cross-language information retrieval system [7]. It is assumed that users of such a system who are not fluent in a foreign language can read a retrieved document of the foreign language well enough to judge the document's relevance [1]. This assumption greatly reduces the usability of such systems, since users with little knowledge for the foreign language are denied access to possibly valuable information written in that language. Furthermore, users with the capability of judging the relevance of foreign documents should also be able to issue queries in the foreign language! In this case, the function of a cross-language information retrieval system is simply to offer users convenience.

To fully explore the potential of a cross-language information retrieval system. MuST includes several machine translation engines. A glossing engine, QuTE, was built to provide rapid but shallow translation of foreign language documents into the source language (currently English). QuTE enables users quickly to judge a document's relevance, even if they are not familiar with the foreign language. QuTE is also used as MuST's query translation module. If users decide they want to learn more details about one particular document, they can obtain higher quality translation using a full-fledged machine translation system such as GAZELLE [13] or SYSTRAN [6], or submit the document to a human translator.



You searched for: +digital +library

Top 10 matches [illegible] Show: Headlines View by: Publication or Date

Story: Make Your Mouse a Powerhouse (AnchorDesk)
67% And just as I'm the unsung hero of AnchorDesk, your mouse is the unsung piece of your PC. LEARN HOW Easy Download Guide - ZDNet Software Library DOWNLOAD Noisy Mouse (Win95-98) - ZDNet Software LibraryCool Mouse (Win95-98-NT) - ZDNet Software LibraryPop-Mouse (Win95-98-NT) - ZDNet Software Library10 Cool Mouse Tools - ZDNet Software LibraryAnchorDesk's Killer Downloads - ZDNet AnchorDesk TODAY
First found: 1 Feb 1999 (Related Articles)

Story: Sweeping Your System Tray With Free Tools (AnchorDesk)
68% PS: If you need help, I've linked the Easy Download Guide in the sidebar. ZDNet and the ZDNet logo are trademarks of Ziff-Davis Inc.
First found: 1 Feb 1999 (Related Articles)

Zoran Digital Audio Processor Selected for Boston Acoustics DigitalThe... (Business Wire)
63% Zoran's recently announced ZR38601 and ZR38650 provide a unique solution for the digitally-driven speaker systems market. Developed in connection with Dolby Laboratories, it is the ideal solution for a wide array of digital audio

Figure 2. An Excite search session with query: "+digital +library".

GAZELLE, which is linked into MuST. is a knowledge-based machine translation system for Arabic-English, Japanese-English, and Spanish-English developed at USC/ISI with support from the Department of Defense. The system operates over unrestricted newspaper text. It uses large-scale semantic representations and reasoning to improve accuracy, and automated linguistic knowledge induction from large corpora to increase coverage.

2.4 Text Summarization Module

Most commercial Internet search engines return search results with short summaries. The summaries are intended to give a quick overview of the search result and help users select the relevant web pages. Figure 2 shows an Excite search result page. The query is "+digital +library". We can guess roughly what the returned pages are about by reading their summaries, but the summaries often do not explain why these pages are returned. The reason for this deficiency is that these summaries are constructed at indexing time and do not take users' queries into account. Tombros and Sanderson [22] have shown that the use of query biased summaries significantly improves both the accuracy and speed of user relevance judgements.

MuST includes a multilingual text summarization engine, SUMMARIST [11,14]. The goal of SUMMARIST is to create summaries of arbitrary text in English and other languages. Like many summarization systems, SUMMARIST has the 3-stage architecture: Summarization = topic identification +

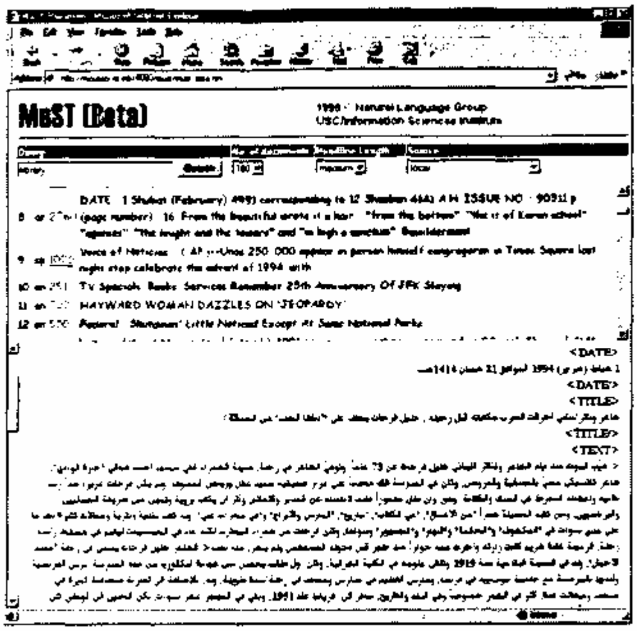

MuST (Beta)
1998 Natural Language Group
USC/Information Sciences Institute
<DATE>
<DATE>
<TITLE>
<TITLE>
<TEXT>

Figure 3. MuST user interface shows an Arabic text retrieved by query "library".

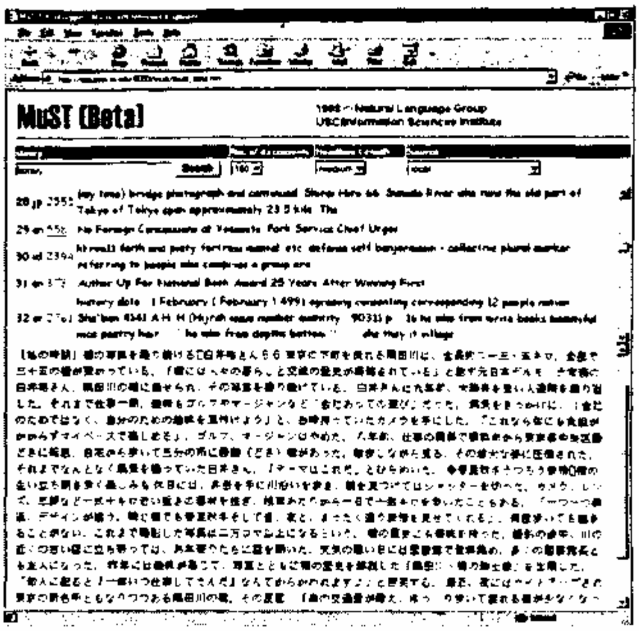

MuST (Beta)
1998 Natural Language Group
USC/Information Sciences Institute

Figure 4. MuST user interface shows a Japanese text retrieved by query "library".

interpretation + generation. The topic identification stage is by far the most developed; in fact, the production version of SUMMARIST presently produces extracts only. SUMMARIST is capable of generating query biased summaries that reflect both users' concerns and the main ideas of the respective documents through automated training [15]. Uses can also specify length of summaries.

The inclusion of a summarization engine not only boosts the performance of user relevance judgements. but also eliminates the cost of translating unnecessary information. As shown in Figure 1, users can choose to submit only summaries instead of full texts to QuTE or other deep translation engines. Based on the review of the translated summaries, they can discard the irrelevant documents and send the relevant ones for further processing.

The MULINEX project [4] reported that users valued summaries as helpful and time saving. Users also praised the availability of summary translations, although the quality of summaries and translations were not good enough and the lengths of the summaries were not always optimal. MuST addresses these factors by recognizing the raw quality of shallow translation and only using it as a rapid glossing/viewing aid to the users and SUMMARIST allows users to decide the optimal length of summary.

2.5 User Interface Module

MuST chooses web browsers as its primary user interface. The ubiquity of web browsers provides a natural way for users to interact with information access systems. The availability of Unicode and modern web browsers such as Netscape Navigator 4.0 and Microsoft Internet Explorer 4.0 greatly reduces the effort needed to enable multilingual access. Figure 3 shows how an Arabic enabled version of Internet Explorer can render Arabic text from right to left and take care of the ligature between Arabic characters. Figure 4 shows that the same browser can also display Japanese text by adding freely available Japanese language support. However, input for languages such as Arabic is still a problem[2]. We plan to develop a Java applet to handle languages without native support.

In the next section, we use Indonesian as an example to demonstrate the design decision we made for this particular language and how new language capability can be added to MuST in greater details.

3 Bahasa Indonesia: A Case Study

We choose to work on Bahasa Indonesia for the following reasons:

- Indonesia is the fourth most populous country after China. India, and the United States, with a population of about 200 million. Bahasa Indonesia is the official language of Indonesia.
- The Indonesian Internet resources are rich, and major Indonesian newspapers such as Kompas, Pikiran Rakyat, and Suara Merdeka are online.
- No major US commercial search engines provide exclusive search services for Indonesia.

[2] The latest version of Internet Explorer already provides the capability of inputting Chinese, Japanese, and Korean from the browser interface.

- Machine-readable dictionaries of Indonesian to English and English to Indonesian are available.
- No special treatment is required for input and display of Indonesian.
- The author would like to know what happens in Indonesia every day and what Indonesians say about the world events.

In the following, we discuss each step of the implementation.

3.1 Identifying Resources

We first performed a manual search on information regarding Indonesia through several Internet search engines, as seed pointers to potential points of interest. We then evaluated each potential site of interest based on its quality and accessibility. We ended up with ten news sites that include national and local general newspapers and business newspapers[3]. This pool reflects our interest in knowing day-to-day events reported in Indonesia.

3.2 Indexing Agent

In order to keep tracking the selected sites, we schedule PLSpider, a web robot from America Online Inc., to visit these sites every day and create a local searchable database based on PLWeb, a search tool also from America Online Inc. As shown in Figure 1, this setup is a sub-module of the MuST information retrieval component aiming at offering specialized information access needs.

Notice that we do not keep a local copy of the documents indexed by PLSpider. We only build a local index database that contains the URL links pointing to the actual documents on the web. However, pages on news sites tend to change every day and the same URL link may point to pages of the latest content instead of the material seen by the PLSpider at indexing time. URLs are simply placeholders in this case. To remedy this difficulty, creating a local cache of the indexed document seems a reasonable solution. How to resolve the copyright issue then becomes the major problem.

3.3 Bilingual Transfer Dictionary

When we first searched for relevant Indonesian resources on the Web, we also looked for online bilingual Indonesian-to-English dictionaries (IED) and English-to-Indonesian dictionaries (EID). We used the IED to build a shallow translator as described in the next section and the EID to aid query translation. Three IEDs and two EIDs were found. The quality of these dictionaries is not optimal and most of them required manual cleanup. We also manually added Indonesian-to-English phrase translations into the final merged EID. At the end, we had a 22,797 entry EID and 17,010 entry IED.

Although many free online dictionaries are available, they usually require normalization and consistency checking. Many of them only contain the most frequently used words, therefore coverage is a problem. However, the main advantage is that free online dictionaries usually are encoded with word for word mappings instead of implicit encoded translations found in commercial machine readable bilingual dictionaries [7]. We plan to use the corpus-based approach to remedy the lack of coverage problem as suggested by Sheridan & Ballerini [21].

3.4 Shallow Translator

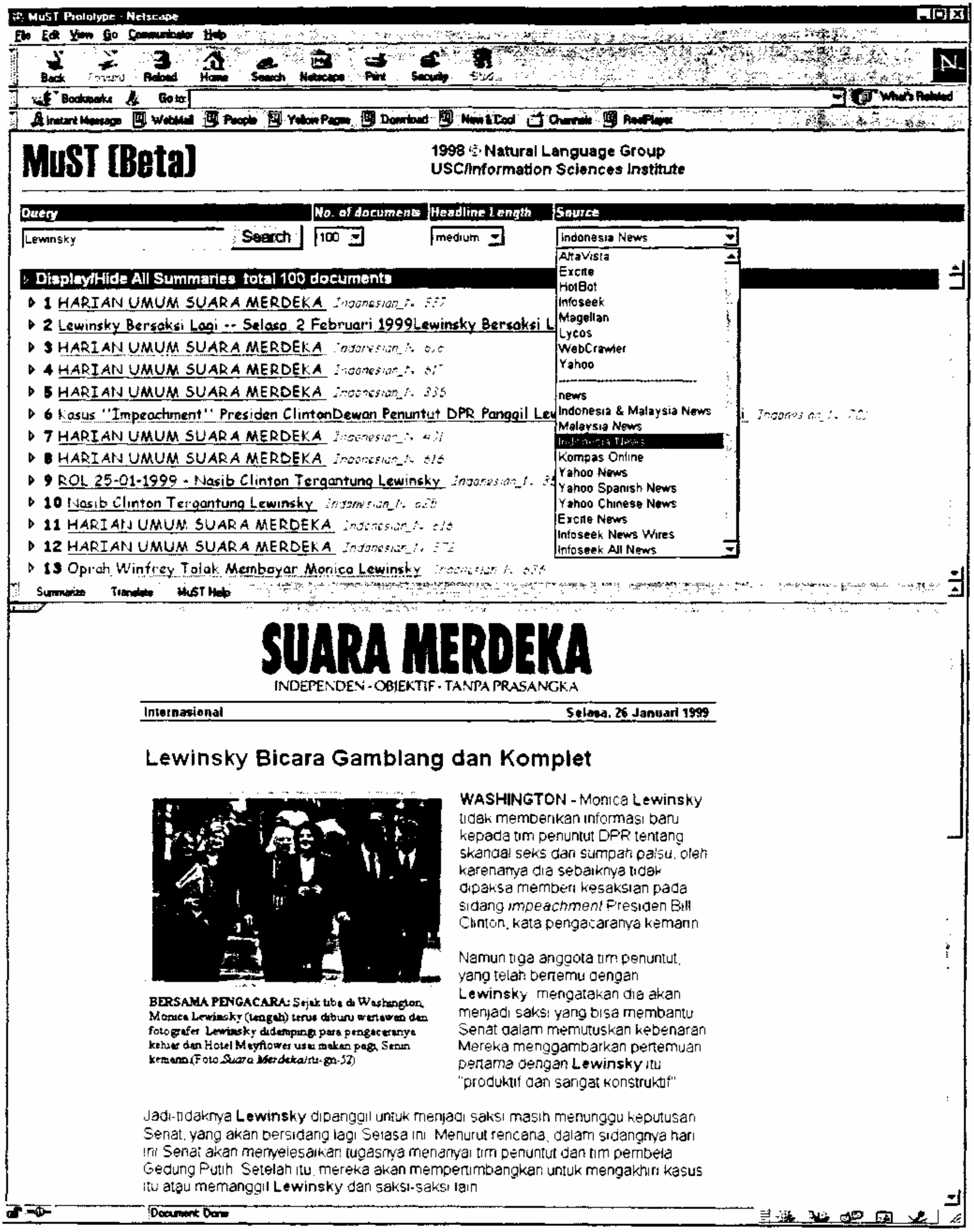
Building a word-for-word shallow translator for Indonesian-to-English is easier than for Korean-to-English since the word order of Indonesian is subject + verb + object, similar to our source language English. However, Indonesian is an inflected language and dictionary entries contain only root words. Morphological analysis is necessary to properly select the translation candidates. We built a simple morphological analysis engine for Indonesia that recognizes common Indonesian affixes, converts Indonesian inflections into their root forms, and attaches basic syntactic marker such as plural, passive, and so on.

Users of the shallow Indonesian-to-English translator have reported positive comments when using it as a browsing aid. However, in a separate experiment at ISI, users were not satisfied with an early prototype of a Korean-to-English shallow translator that was created in a similar manner. This indicates that for different language-pairs various amount of development time should be expected. How to normalize the performance of each language-pair and present it to users in a uniform quality is a subject of future research.

3.5 Indonesian Summarizer

The SUMMARIST [11,14] design makes augmenting it with Indonesian very easy. We implemented an Indonesian text normalization module that converts Indonesian plain texts or HTML pages into SUMMARIST normal form (SNF). The normalization module consists of a tokenizer and the morphological analyzer built for the shallow translator. Different topic identification modules are then run through the SNF. At the end, a sentence selector combines all the scores reported by various modules and the number of desirable summary sentences preset by the user is output as the summary.

[3] The ten sites are Bernas Online, Bisnis Indonesia, Jawa Post, Kompas Online, Pikiran Rakya, Repulika Online, Suara Merdeka, Suara Pembarun, Surabaya Post, and Tempo Interaktif.


MuST Prototype - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security
Bookmarks Go to: What's Related
Instant Message WebMail People Yellow Pages Download New&Cool Channels RealPlayer
Summarize Translate MuST Help
Document: Done

MuST [Beta]

1998 © Natural Language Group
USC/Information Sciences Institute

Query	No. of documents	Headline Length	Source
Lewinsky	100	medium	Indonesia News

Source: AltaVista, Excite, HotBot, Infoseek, Magellan, Lycos, WebCrawler, Yahoo, news, Indonesia & Malaysia News, Malaysia News, Indonesia News, Kompas Online, Yahoo News, Yahoo Spanish News, Yahoo Chinese News, Excite News, Infoseek News Wires, Infoseek All News

Display/Hide All Summaries total 100 documents

1 HARIAN UMUM SUARA MERDEKA *Indonesian_f. 557*
2 Lewinsky Bersaksi Lagi -- Selasa, 2 Februari 1999Lewinsky Bersaksi L
3 HARIAN UMUM SUARA MERDEKA *Indonesian_f. 6..*
4 HARIAN UMUM SUARA MERDEKA *Indonesian_f. 617*
5 HARIAN UMUM SUARA MERDEKA *Indonesian_f. 335*
6 Kasus ''Impeachment'' Presiden ClintonDewan Penuntut DPR Panggil Le *Indonesian_f. 70.*
7 HARIAN UMUM SUARA MERDEKA *Indonesian_f. 421*
8 HARIAN UMUM SUARA MERDEKA *Indonesian_f. 616*
9 ROL 25-01-1999 - Nasib Clinton Tergantung Lewinsky *Indonesian_f. 3.*
10 Nasib Clinton Tergantung Lewinsky *Indonesian_f. 625*
11 HARIAN UMUM SUARA MERDEKA *Indonesian_f. 616*
12 HARIAN UMUM SUARA MERDEKA *Indonesian_f. 572*
13 Oprah Winfrey Tolak Membayar Monica Lewinsky *Indonesian_f. 674*

SUARA MERDEKA
INDEPENDEN - OBJEKTIF - TANPA PRASANGKA

Internasional Selasa, 26 Januari 1999

Lewinsky Bicara Gamblang dan Komplet

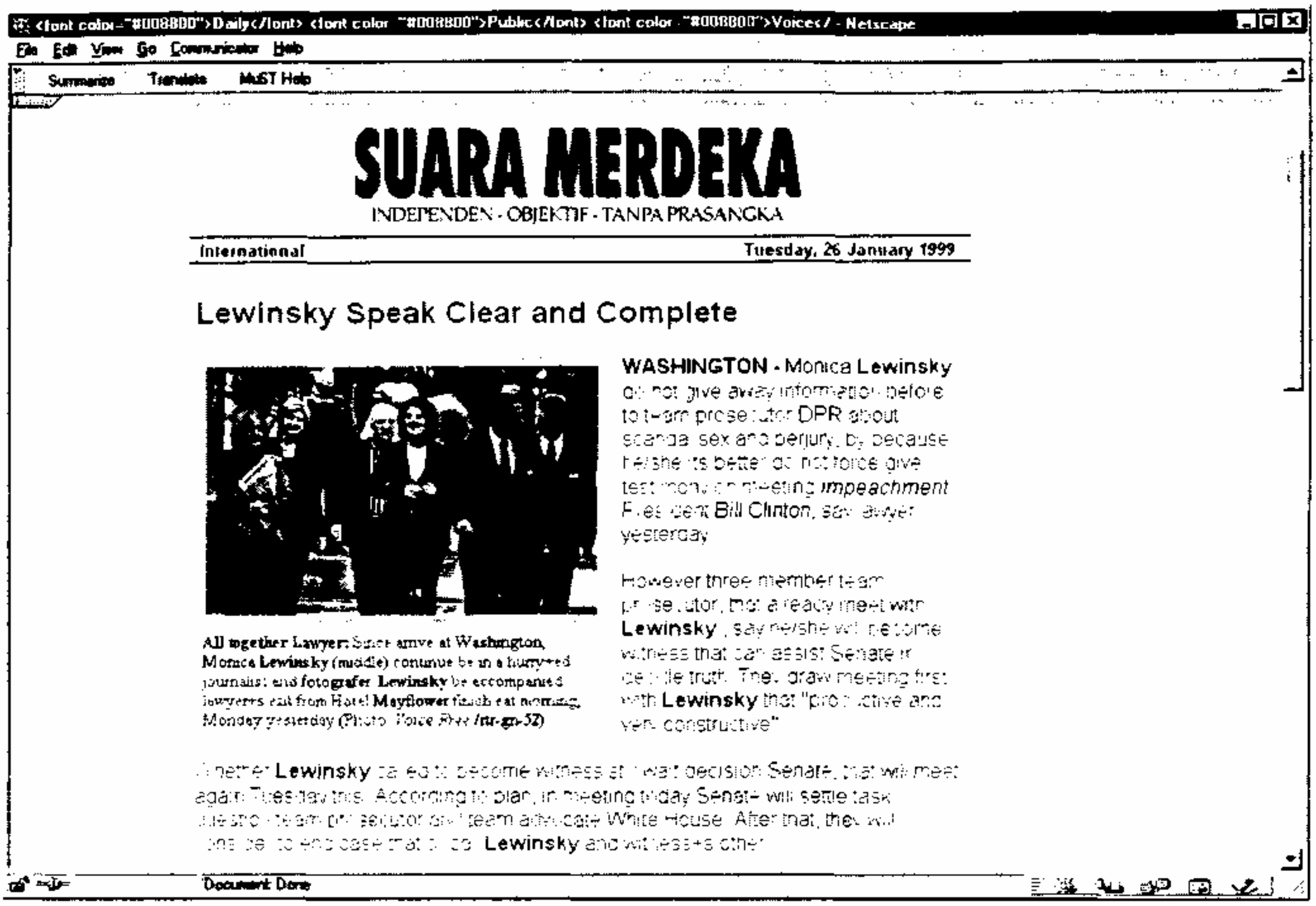
BERSAMA PENGACARA: Sejak tiba di Washington, Monica Lewinsky (tengah) terus diburu wartawan dan fotografer. Lewinsky didampingi para pengacaranya keluar dari Hotel Mayflower usai makan pagi, Senin kemarin.(Foto *Suara Merdeka*/rtr-gn-52)

WASHINGTON - Monica **Lewinsky** tidak memberikan informasi baru kepada tim penuntut DPR tentang skandal seks dan sumpah palsu, oleh karenanya dia sebaiknya tidak dipaksa memberi kesaksian pada sidang *impeachment* Presiden Bill Clinton, kata pengacaranya kemarin

Namun tiga anggota tim penuntut, yang telah bertemu dengan **Lewinsky** mengatakan dia akan menjadi saksi yang bisa membantu Senat dalam memutuskan kebenaran Mereka menggambarkan pertemuan pertama dengan **Lewinsky** itu "produktif dan sangat konstruktif"

Jadi-tidaknya **Lewinsky** dipanggil untuk menjadi saksi masih menunggu keputusan Senat, yang akan bersidang lagi Selasa ini Menurut rencana, dalam sidangnya hari ini Senat akan menyelesaikan tugasnya menanyai tim penuntut dan tim pembela Gedung Putih Setelah itu, mereka akan mempertimbangkan untuk mengakhiri kasus itu atau memanggil **Lewinsky** dan saksi-saksi lain

Figure 5. Screen shot of MuST in a retrieval session with query "Lewinsky". The top panel allows users to submit queries, set the number of returned documents, selects the maximum length of headline list, and choose the source of databases (web or local search engines). The middle panel shows the returned document list. The list names are taken directly from the `title` section of documents. In this case, Suara Merdeka put its company name in every HTML page. The bottom panel shows the full Indonesian text selected by the user.


<font color="#008800">Daily</font> <font color "#008800">Public</font> <font color "#008800">Voice</ - Netscape
File Edit View Go Communicator Help
Summarize
Translate
MuST Help
SUARA MERDEKA
INDEPENDEN - OBJEKTIF - TANPA PRASANGKA
Document Done

International Tuesday, 26 January 1999

Lewinsky Speak Clear and Complete

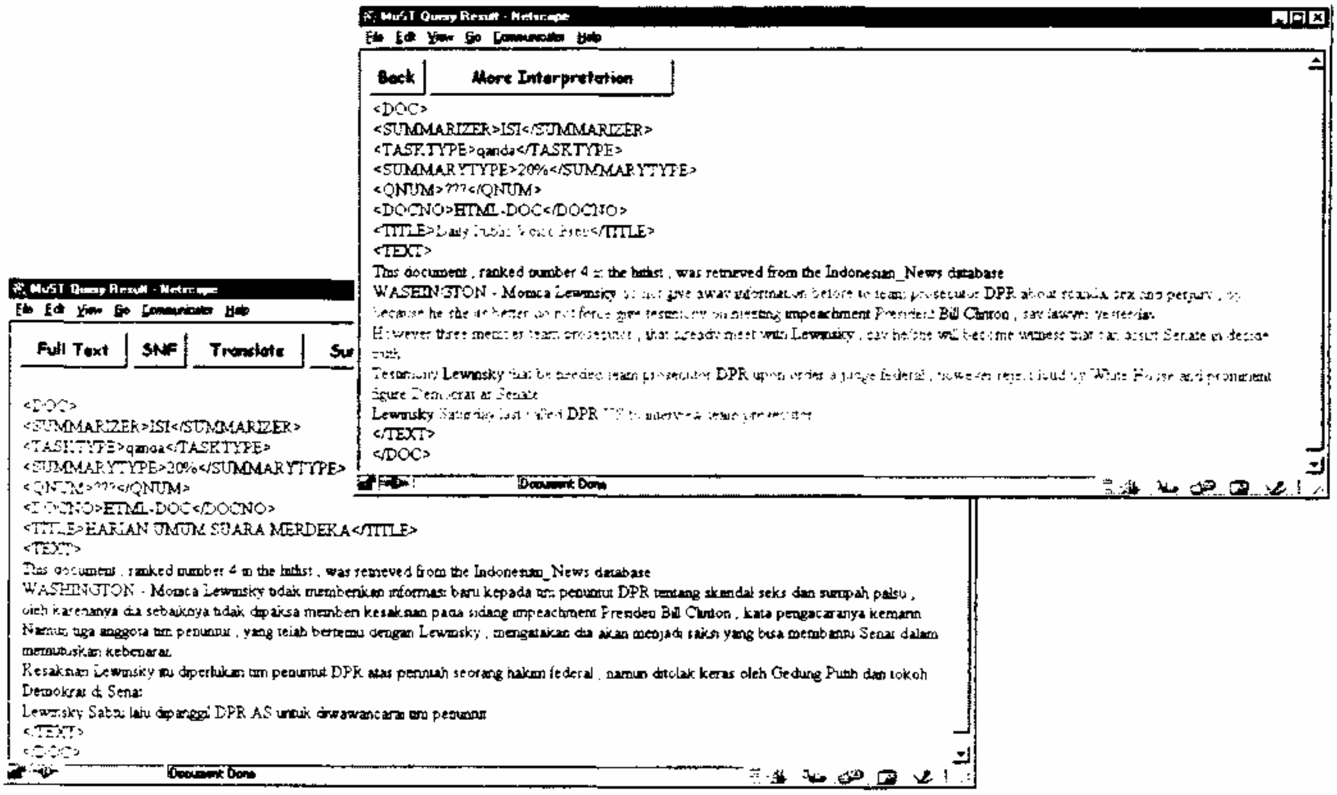
WASHINGTON - Monica Lewinsky do not give away information before to team prosecutor DPR about scandal sex and perjury, by because he/she its better do not force give testimony in meeting *impeachment* President Bill Clinton, say lawyer yesterday

However three member team prosecutor, that already meet with **Lewinsky**, say he/she will become witness that can assist Senate in decide truth. They draw meeting first with **Lewinsky** that "productive and very constructive"

All together Lawyer: Since arrive at **Washington**, Monica **Lewinsky (middle)** continue be in a hurry-ed journalist and **fotografer**. **Lewinsky** be accompanied lawyers exit from Hotel **Mayflower** finish eat morning, Monday yesterday (Photo: *Voice Free* **/rtr-gn-52)**

Whether **Lewinsky** called to become witness still wait decision Senate, that will meet again Tuesday this. According to plan, in meeting today Senate will settle task question team prosecutor and team advocate White House. After that, they will continue to end case that involve **Lewinsky** and witnesses other

Figure 7. Screen shot of the translated HTML page shown in Figure 5. The terms such as "DPR" (the House) and "impeachment" are not translated because they are not in the transfer dictionary.


MuST Query Result - Netscape
File Edit View Go Communicator Help
Back
More Interpretation
Document Done
MuST Query Result - Netscape
File Edit View Go Communicator Help
Full Text
SNF
Translate
Su
Document Done

<DOC>
<SUMMARIZER>ISI</SUMMARIZER>
<TASKTYPE>qanda</TASKTYPE>
<SUMMARYTYPE>20%</SUMMARYTYPE>
<QNUM>???</QNUM>
<DOCNO>HTML-DOC</DOCNO>
<TITLE>Daily Public Voice Free</TITLE>
<TEXT>
This document , ranked number 4 in the hitlist , was retrieved from the Indonesian_News database
WASHINGTON - Monica Lewinsky do not give away information before to team prosecutor DPR about scandal sex and perjury , by because he she its better do not force give testimony in meeting impeachment President Bill Clinton , say lawyer yesterday.
However three member team prosecutor , that already meet with Lewinsky , say he/she will become witness that can assist Senate in decide truth.
Testimony Lewinsky that be needed team prosecutor DPR upon order a judge federal , however reject loud by White House and prominent figure Democrat at Senate
Lewinsky Saturday last called DPR ??? to interview team prosecutor
</TEXT>
</DOC>

<DOC>
<SUMMARIZER>ISI</SUMMARIZER>
<TASKTYPE>qanda</TASKTYPE>
<SUMMARYTYPE>20%</SUMMARYTYPE>
<QNUM>???</QNUM>
<DOCNO>HTML-DOC</DOCNO>
<TITLE>HARIAN UMUM SUARA MERDEKA</TITLE>
<TEXT>
This document , ranked number 4 in the hitlist , was retrieved from the Indonesian_News database
WASHINGTON - Monica Lewinsky tidak memberikan informasi baru kepada tim penuntut DPR tentang skandal seks dan sumpah palsu , oleh karenanya dia sebaiknya tidak dipaksa memberi kesaksian pada sidang impeachment Presiden Bill Clinton , kata pengacaranya kemarin
Namun tiga anggota tim penuntut , yang telah bertemu dengan Lewinsky , mengatakan dia akan menjadi saksi yang bisa membantu Senat dalam memutuskan kebenaran
Kesaksian Lewinsky itu diperlukan tim penuntut DPR atas perintah seorang hakim federal , namun ditolak keras oleh Gedung Putih dan tokoh Demokrat di Senat
Lewinsky Sabtu lalu dipanggil DPR AS untuk diwawancarai tim penuntut
</TEXT>
</DOC>

Figure 6. Screen shot of 20% summary and its translation (top) of the Indonesian HTML page shown in Figure 5.

3.6 The Integrated System

Figure 5 shows a screen shot of MuST in a retrieval session with query "Lewinsky". The top panel allows users to submit queries, set the number of returned documents, select the maximum length of headline list, and choose a information source to search. Information sources can be remote, such as Yahoo, Excite, or Infoseek, or specialized local information sources such as Malaysia News and Indonesia News. Indonesia News is the active database in this search session.

The middle panel shows the returned document list. The list names are taking directly from the title section of documents. In this case, Suara Merdeka (Free Voice) places its company name in the title tag position in every HTML page in its site. Users can click the right arrow head bullet at the beginning of each title to display its headline.

The bottom panel shows the Indonesian web page selected by the users. Users can click the Translate button on the menu bar to translate the selected page into English. Figure 6 shows the translated page. Terms such as DPR (the House), impeachment, and proper names are not translated because they are not in the transfer dictionary.

If users would like to read a summary instead of the full text, they can click the Summarize button and a summary window is displayed as shown in Figure 7. The length of the summary can be adjusted through the Summary size pulldown menu. A translation of the summary can be obtained by clicking the Translate button on the Indonesian summary page. Figure 7 also shows the translated summary. If the translation is not indicative enough, users can select the More Interpretation button to see more translation lexical alternatives.

We have deployed a beta version of MuST at ISI and demonstrated the system in several conferences. Although the initial feedback from our users is positive and detailed evaluations of some modules such as the summarization engine are available [15], we plan to perform more user studies in the future.

4 Conclusions and Future Directions

In this paper we describe the design and implementation of MuST, a multilingual information retrieval, summarization, and translation system. MuST emphasizes enabling users to perform cross-language information access, reusing available search services whenever it is possible, building specialized local search services when special needs are present, adopting ubiquitous web browsers as its primary user interface, and tightly integrating automated shallow translation and summarization. The ideal deployment environment of MuST is the landscape where the user community wants to have a unified interface to general and specialized search services and the capability to access multilingual information.

The main differences between MuST and a related Commission of the European Communities project, MULINEX [4], are: (1) a shallow translation module, QuTE, enabling quick browsing, (2) a robust user biased text summarizer based on tested SUMMARIST technology that can be ported to many languages quickly, and (3) a streamlined methodology of adding new language capability as demonstrated in the Indonesian case study. MULINEX also includes a translation component. However, it totally relies on an external source to achieve uniform translation quality; ensuring the availability of any interested translator can be a problem. The summarizer component in MULINEX does not perform query biased summarization. Nevertheless, the document classification, information extraction, and user profile servers described in MULINEX Synthesis Report are missing in the current MuST architecture. We plan to integrate these capabilities into MuST in the future, extending some of the clustering and analysis techniques built in the C*ST*RD project at ISI.

MuST accepts only English as its source language at the present time. However, it can search any target language if a bilingual transfer dictionary and the target language monolingual search service are available. We plan to add several more source languages later.

According to Campbell [2], there are at least 90 languages in the world spoken by at least 5 to 10 million people. Based on our experience with Indonesian, it takes about 2 months for a full time researcher to develop the information processing and accessing capability as presented in the previous section. The resource requirement of including Indonesian into MuST is far less and easier to acquired than many other languages such as Cambodian, Thai, or Tibetan. The amount of time required to accommodate these languages is probably much longer than for Indonesian. Nevertheless, recent calls for developing machine translation techniques for languages of low diffusion reflect the high interest of research in this area. We plan to follow our experience with Indonesian and gradually add language support for as many new languages as possible.

To tackle the query translation problem as reported in much cross-language information retrieval research [1,3,12,21], we plan to create a large semantic knowledge base by ontology alignment, dictionary parsing, and web mining to overcome the meaning fanout problem. Initial results are reported in Lin & Hovy [16] and Hovy [10].

Voorhees & Tong [23] report that fusing retrieval results from multiple collections could achieve better performance than from a single collection. MuST currently directs users' queries to a single database. Allowing MuST users to a submit single query and search all the available collections would be a good addition to MuST.

Acknowledgements

The author would like to thank Eduard Hovy for valuable discussions, supporting this research, and reviewing of an early version of this paper. Ulrich Germann, Ulf Hermjakob, and Richard Whitney provided useful suggestions for the integration of the GAZELLE machine translation system. Thanks also go to AOL/PLS and MG group for making their search software freely available. This work is supported in part by C*ST*RD project under DARPA contract N66001 -97-9538.

References

1. Ballesteros, L. and Croft, W. B. Statistical Methods for Cross-language Information Retrieval. In G. Grefenstette, editor, Cross-Language Information Retrieval, chapter 3. Kluwer Academic Publishers, Boston, pp. 23-40, 1998.
2. Campbell, G. L. Concise Compendium of the World's Languages. Routledge, New York, 1995.
3. Davis. M. On the Effective Use of Large Parallel Corpora in Cross-language Text Retrieval. In G. Grefenstette, editor. Cross-Language Information Retrieval, chapter 2. Kluwer Academic Publishers, Boston, pp.12-22, 1998.
4. Erbach, G., Neumann, G., and Uszkoreit, H. MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World Wide Web. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech. Stanford, California, 1997.
5. Fluhr, C., and Radwan, K. 1993. Full Text Databases as Lexical Semantic Knowledge for Multilingual Interrogation and Machine Translation. In EWAIC'93, 1993.
6. Gachot, D. A., Lange, A., Yang, J. The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Cross-Language Information Retrieval. In G. Grefenstette, editor, Cross-Language Information Retrieval, pp.105-118, 1998.
7. Grefenstette, G. The Problem of Cross-Language Information Retrieval. In G. Grefenstette, eds., Cross-Language Information Retrieval, pp.1-9, 1998.
8. Harman, D. K. The 4th Text Retrieval Conference (TREC-4), Gaithersburg, Md., Nov. 1-3, 1995.
9. Hiemstra, D. A. Linguistically Motivated Probabilistic Model of Information Retrieval, In: C. Nicolaou and C. Stephanidis (eds.), Proc. of the second European Conf. on Research and Advanced Technology for Digital Libraries, pp. 569-584, 1998.
10. Hovy, E. H. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. In Proc. of the 1st International Conf. on Language Resources and Evaluation. 1998.
11. Hovy, E. H. and Lin, C. Y. 1999. Automating Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. Cambridge: MIT Press.
12. Hull, D. and Grefenstette, G. Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval. In Proc. of the 19th ACM/SIGIR Conf., pp. 49-57, 1996.
13. Knight, K. GAZELLE: Machine Translation of Natural Languages. Information Sciences Inst., Univ. of Southern California, Marina del Rey, Ca.
14. Lin, C. Y. 1998. Assembly of Topic Extraction Module in SUMMARIST. In Working Notes of AAAI Spring Symposium on Intelligent Text Summarization. Stanford, Ca, March 1998.
15. Lin, C. Y. Training a Selection Function for Extraction. Submitted.
16. Lin, C. Y. and E.H. Hovy. 1999. The Automated Acquisition of Topic Signatures for Text Summarization. Submitted.
17. Mani, I., Firmin, T., House, D., Klein, G., Sundheim, B., Hirschman, L. The TIPSTER SUMMAC Text Summarization Evaluation, Proc. of EACL'99, Bergen, Norway, June 8-12, 1999.
18. Oard, D. W. and Dorr, B. J. A Survey of Multilingual Text Retrieval. Technical Report UMIACS-TR-96-19, Univ. of Maryland, Inst. for Advanced Computer Studies, April 1996.
19. Oard, D. W. Serving Users in Many Languages: Cross-Language Retrieval for Digital Libraries. In D-Lib Magazine. December, 1997.
20. Oard, D. W. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In the 3rd Conf. of the Association for Machine Translation in the Americas (AMTA), Philadelphia, PA, October 1998.
21. Sheridan, P. and Ballerini, J. P. Experiments in Multilingual Information Retrieval Using the SPIDER System. In Proceedings of the 19th ACM/SIGIR Conference, pp. 58-65, 1996.
22. Tombros, A. and Sanderson M. Advantages of Query Biased Summaries in Information Retrieval. In Proc. of the 21st ACM/SIGIR Conf., pp. 2-10, 1998.
23. Voorhees, E. M., Tong, R. M. Multiple Search Engines in Database Merging. In Proc. of the 2nd ACM Int'l Conf. on Digital Libraries, pp. 93-102, 1997.
24. Witten, I. H., Moffat, A., and Bell, T. C. Managing Gigabytes: Compressing and Indexing Documents and Images. Van Nostrand Reinhold, New York. 1994.
25. Yam Digital Co. Taiwan. http://www.yam.com.tw.