An Exploration of News Meta-Search Across Multiple Languages

Sheila Patel
Computer Science Department
Stanford University
sheila@cs.stanford.edu

ABSTRACT
This paper presents aspects involved in executing a cross-language meta-search for news. A specific solution, here forth referred to as *Global Reporter* is described, as are users' feedback, and information along with fact coverage results from a controlled experiment with the Global Reporter. Where suitable, recommendations for modifying the system are presented as well as trade-offs to alternative implementations.

Keywords
News meta-search, cross-language, parallel query execution, machine translation.

1. INTRODUCTION
The Global Reporter project commenced as an effort to capture the differences in media coverage in different languages. While there are several meta-search engines for online news sites, such as Ithaki [7], MyHawker [9], and Google News [6] among others, none yet appear to incorporate cross-language articles into their corpus of news articles. MyHawker, for instance, allows a user to search for news articles in 10 different languages; however, the articles are only retrieved from news sources in the language of the query and not across others. CLEF (Cross-Language Evaluation Forum) workshop proceedings indicate the need for tools providing multi-lingual access and retrieval on the web. As a result, a cross-language meta-search engine for news appears to be valuable.

A first cut of the Global Reporter project attempts to provide just this, the ability to carry out a cross-language meta-search of news sites. A user may enter a query in English, and select to retrieve results from an *English-only Search* or a *Multi-language Search*. The final results, regardless of the language of the news source, are translated to English, ranked relative to the other articles, and then returned to the user in English. With the first cut of the system, an easily expandable basic architecture was built, and some retrieval properties regarding information coverage and distinct fact contributions from multi-language versus English-only meta-search results were studied. Traditional retrieval performance measures of recall, precision and response time were not evaluated for reasons mentioned later in the paper. The majority of the users, who experimented with the system in studying retrieval properties, found such a system to be potentially useful. However, their feedback for the base system indicates that the *Multi-language Search* feature of Global Reporter is best augmented in other ways, such as by attempting to expose actual reporting differences across different languages, before it can provide further utility to a user beyond just an English-only search.

1.1 Related Work
Although not specifically for meta-searching news sites on-demand, several projects have attempted to integrate cross-language searching with meta-searching. In 1999, a collaborative effort between NTT (Japan), KAIST (Korea), and KRDL (Singapore) lead to the construction of a system referred to as "Cross-language Information Retrieval Architecture" [5]. This system used natural language processing and distributed search engines in the Japanese, Chinese and Korean languages to search for results to a user's query.

Another interesting project, in the realm of cross-language meta-searching pertains to the UCLIR multi-language information retrieval tool [1]. The UCLIR system attempts to integrate interactive user feedback and modifications to translated queries prior to submitting the queries to search engines for result retrieval. Another system, called CLARIT, applies pseudo-relevance feedback to cross-language retrieval to improve retrieval performance [3].

Chin-Yew Lin's work at the University of Southern California on the MuST system also involves information retrieval and summarization across the language barrier [8]. The MuST system allows users to carry out cross-language searches with commercial web search engines such as the Altavista and Excite search engines in English along with several in Asian languages

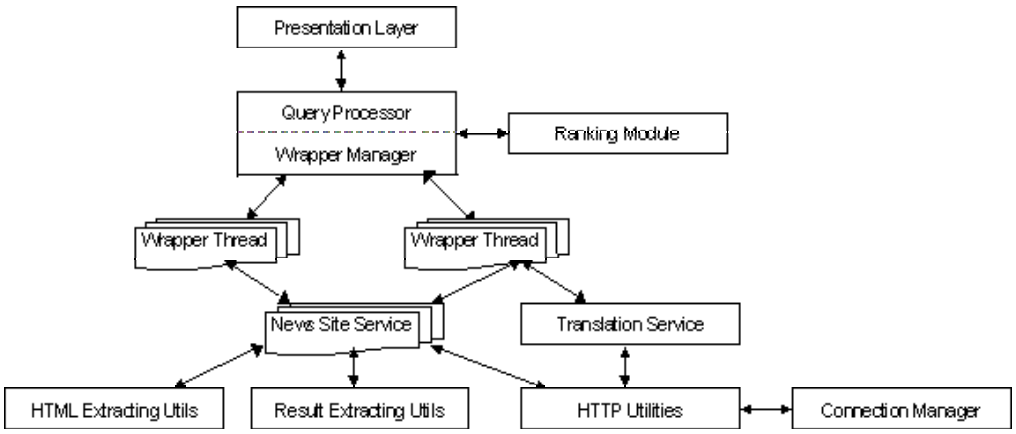

Figure 1. Global Reporter Architecture

like Taiwan's Yam search engine. Following retrieval, user biased summarization techniques are used in MuST to allow users to assess the relevance value of documents quickly.

As mentioned earlier, several engines like Ithaki [7] and MyHawker [9] currently provide meta-search capability for online news sites. However, the results are restricted to those sites in the same language as the original query that was submitted to the meta-search engine.

1.2 Outline of Paper

Section 2 of the paper provides an overview of the basic architecture for the Global Reporter system, and describes some alternatives and potential enhancements to the implementation. Section 3 discusses the decisions behind selecting certain news sources, and not selecting others. Experimental setup and results from evaluating the basic system are presented in Section 4. Lastly, Section 5 concludes with a summary of the work.

2. SYSTEM ARCHITECTURE

In developing a cross-language meta-search engine for news, the main components deemed necessary for such a system were: a Query Processor, a Translation Service, several News Site Services, as well as HTML extraction and HTTP connection utilities. The *News Site Services* serve as sources of news articles both in English and non-English languages. Global Reporter was built with these components in mind. The basic architecture of the Global Reporter system is illustrated in Figure 1. The system was implemented entirely in the Java[1] programming language (API Specification Version 1.4.0). The remainder of this section provides a brief introduction of the major components and modules. For detailed information regarding the implementation, please consult the javadoc documentation for the Global Reporter source.

The *Presentation Layer* in the diagram was implemented using the Java Servlet Technology, and will not be discussed much further. Java Swing or AWT may replace this layer taking the place of the servlets. Also, the *Presentation Layer* may be replaced altogether by a client application accessing the *Query Processor* API.

2.1 Query Processor

The *Query Processor* receives queries from the *Presentation Layer*, does the majority of processing involved in order to return result articles back to the user, and interfaces most of the other modules either directly or indirectly. While the first cut of the system, as mentioned previously, allows users to query and retrieve all results translated to English, the base language for the system may be easily configured to be any for which Java has a supporting language locale[2]. This base language locale is passed to and set in the *Query Processor* by the calling *Presentation Layer* or other client application.

The *Query Processor* has access to a *Wrapper Manager*, which generates a *Wrapper Thread* to retrieve translated news articles in the base language for each *News Site Service* used as a data source. In Figure 1, the set of *Wrapper Threads* on the right access *News Site Services* whose content is not in the base language. Therefore, those on the right require a *Translation Service* to facilitate translation of the original query and for retrieval of articles in the base language, whereas those on the left do not.

Once the *Wrapper Threads* retrieve news stories that may potentially answer the query in the base language of the system, the *Ranking Module* is invoked to rank these news stories to determine the relevance order in which to

[1] Refer to http://www.javasoft.com for further information on the Java programming language, Java Servlet Technology, Java AWT, or Java Swing.

[2] Refer to http://www.ics.uci.edu/pub/ietf/http/related/iso639.txt for a list of complete ISO Language Codes supported.

list the abstracts from the result stories. Currently, the Lucene[1] text-search engine is used to rank the news stories from the various *News Site Services* relative to one another after they have been translated to the base language.

2.2 Utilities for HTTP & HTML

The *HTTP Utilities* facilitate connecting to external HTTP servers such as news sources, for the *News Site Service* and *Translation Service* modules. These service modules may thus obtain responses from form submissions. The *HTTP Utilities* access the *Connection Manager* to acquire and release persistent connections to HTTP servers. Persistent connections are utilized in order to reduce the response time of the overall system, by reducing the network time since fewer TCP connections are opened and closed among other benefits. Further advantages and details of persistent connections are presented by W3C [4]. Connection requests are actually sent using the lesser GNU licensed HTTPClient[2] library.

Both the *HTML Extracting Utils* and the *Result Extracting Modules* comprise the HTML Utilities available in the system. The *Result Extracting Utils* are able to retrieve items from an HTML result list page. In the current implementation, XFetch Utilities that rely upon Republica Corporation's XFetch Wrapper® [10] specifically, provide extraction of result list items. The XFetch Wrapper® gives structure to HTML content by converting it to XML. XML data binding is used to extract the actual XML elements as Java objects, and is accomplished using the Castor XML[3]. The *HTML Extracting Utils* facilitate stripping out HTML tags from content pages such as those pages having a full news story.

2.3 Translation Service

The *Translation Service* facilitates translating a piece of text, regardless of word count or length, from one language to another. (In the project source code and earlier this was referred to as the *Translation Wrapper*). The present system relies upon the online Babel Fish[4] Translation service powered by SYSTRAN[5] to perform the actual translations. Several other online translation services such as IBM's Machine Translation alphaWorks project were evaluated as well; however, the translation quality and response time appeared to be best with Babel Fish. Although, the current system only needs the translation facility to/from English and the French, German, Spanish languages, news sources in any of the languages supported by the *Translation Service* (which are those supported by the underlying Babel Fish Translation service) may augment the current system.

2.4 News Site Services

A *News Site Service* provides a data source for news articles. These *News Site Services* act as wrappers around both English and non-English news sites such as http://www.cnn.com in English, and http://www.cnnenespanol.com in Spanish. Each news site has its own *News Site Service*, which may be also be referred to as *News Site Wrapper*. The *News Site Service* can accept a query in the language of the news site, and retrieve result articles from a search with the query. In doing so, the service relies upon the *HTTP Utilities*, *Result Extracting Utils*, and *HTML Extracting Utils* to handle intermediate steps.

2.5 Alternatives & Extensions

While an implementation of the architecture as described, does provide a basic cross-language news meta-search engine, some aspects of the system such as response time, robustness, and query expansion could be further improved. These aspects were not stressed for the first cut of the system, as studying the basic properties and utility of such as system was the emphasis along with building an extensible system.

As mentioned in the subsection about the *HTTP Utilities*, persistent connections were used to somewhat improve the response time of the system. Additionally, the *Wrapper Threads* allow for parallel query execution at the different *News Site Services*, whereby reducing the response time to the worse case time for the slowest site. Despite these attempts, query responses from Global Reporter have been found to take from 8 seconds to 150 seconds on average. In some queries, the time is even longer. These times reflect the Global Reporter under the condition when news site sources retrieve all articles from the past 30 days. An ideal system would cache articles, and possibly crawl news sites versus simply executing queries on the fly, unlike the current system, in order to deliver results within a shorter response time.

The existing system could also be made more resilient by having backup *Translation Service(s)*. In the present system, if the *Translation Service* is unavailable, the non-English news sources are unlikely to return all the relevant results from a query, since they do not expect to receive the query in English or another language unlike the site's.

---

[1] The Apache Jakarta Lucene Project documentation may be found at http://jakarta.apache.org/lucene/docs/index.html

[2] The official web site for the HTTPClient library is at http://www.innovation.ch/java/HTTPClient

[3] Information about the Castor XML, XML data binding framework is available online at http://castor.exolab.org/xml-framework.html

[4] Babel Fish online translation is available through http://babelfish.altavista.com

[5] The SYSTRAN Information and Translation Technologies web site is at http://www.systransoft.com

As for improving the result quality from the news sources not in the base language of the system, query expansion and refinement may be worth exploring and integrating in the system. Strictly machine-translated queries are sometimes ambiguous or irrelevant as opposed to translated queries that are further refined. For example, proper names like "Bush" may be un-intentionally translated and searched for incorrectly. The team working on the CLARIT system, which examined the application of pseudo-relevance feedback for cross-language information retrieval, also used the SYSTRAN system for machine translation in their project. They identified seven types of commonly occurring translation errors with the SYSTRAN system, among which are missing translations (causes requests to return untranslated), incorrect disambiguation by removal of capitalization, and incorrect phrase translation [3]. Their experiments provide evidence that pseudo-relevance feedback improves both recall and precision measures for a cross-language retrieval so that they are comparable with those for monolingual retrieval. Unlike CLARIT, as the Global Reporter project does not focus on the details of the translation component, and as the quality of the translation component is far from reliable, evaluating Global Reporter using retrieval performance measures such as recall and precision was ruled out. Nonetheless, based upon the results from CLARIT, extending Global Reporter to include pseudo-relevance feedback appears to be a promising future endeavor. Extending Global Reporter with interactive user feedback in modifying the query also seems to be worthwhile. The UCLIR system, for example, allows a user to modify a translated query prior to submitting a search request to a foreign source [1]. Initial experimental studies with UCLIR aimed at evaluating the usefulness of taking an interactive approach to query translation showed that results were better for interactively refined translated queries versus those submitted for searches without interactive refinement [1].

3. NEWS SOURCE SELECTION

News sources chosen to provide *News Site Services* for Global Reporter were carefully selected. The system offers the option to perform either an "English-only Search" or a "Multi-language Search", as visible from the Global Reporter UI in Figure 2. Consequently, the English news sources were queried for both the *English-only Search* and the *Multi-language Search*. For English news sources, a news syndication service such as Reuters was not used for two reasons. The first reason is since Reuters is a source of information for other online news sites such as CNN <http://www.cnn.com>, Yahoo! News <http://news.yahoo.com>, etc., whereby Reuters results would dominate and duplicate the results of other sites like

Figure 2. Global Reporter user interface

CNN and Yahoo! News. The second reason that Reuters was excluded, was because its non-English counterparts, such as the French AFP (Agence France-Presse) and the German DPA (Deutsche Presse-Agentur) do not provide search capability to individual consumers, as would be the case for Global Reporter. So, instead of AFP and DPA, client websites subscribing to AFP like the French Yahoo! Actualités, and subscribing to the DPA like the German N-TV were included as *News Site Services*. As for Reuters subscribers, CNN and Yahoo! News were included.

After observing the results retrieved for several queries from news sites that were wrapped as *News Site Services*, six were selected prior to conducting the experiments. Of these six, three were English news sites, namely CNN <http://www.cnn.com>, Yahoo! News <http://news.yahoo.com>, and ABC News <http://abcnews.go.com>. The French Yahoo! Actualités <http://fr.news.yahoo.com>, the Spanish CNN Español <http://cnnenespanol.com>, and the German N-TV <http://www.n-tv.de> were the remaining and non-English sites used in the system.

One may note, that online newspaper sites were excluded from the Global Reporter setup. Reasons for excluding these were since most newspaper sites, such as http://www.lemonde.fr and http://www.spiegel.de, do not provide complete access to their archives. Some news stories require an access fee. Also, in general, the number and quality of results retrieved by news station sites such as the German N-TV, appeared to be better in terms of answering a particular translated query than those from a newspaper site like Spiegel.

Some sites could not be included prior to the experiments with Global Reporter, but may be worthy of including in the future due to their coverage. One such site is the French Voila Actu <http://actu.voila.fr> that sources content extensively from the AFP. Another is the English BBC News Online <http://news.bbc.co.uk> as this site sources content from the Press Association, Associated Press, Reuters, and AFP. News sites having content from other press agencies like Switzerland's SDA (Schweizerische Depeschenagentur) are also worth

considering, as are those with content from Asian press agencies.

4. EXPERIMENTS

In order to assess the potential utility of a cross-language news meta-search engine like Global Reporter, several experiments were devised to study the retrieval properties of information coverage and distinct fact counts. The retrieval properties examined overall coverage of news events, along with distinct event (or fact) contribution from an *English-only Search* versus the overall coverage and distinct event contribution from a *Multi-language Search* for a set of queries. Additionally, users, who conducted and gathered data for the experiments regarding the aforementioned retrieval properties, were asked for a feedback score of the potential benefit of a system like Global Reporter. This score along with their comments and suggestions are presented later in Subsection 4.4.

4.1 Experimental Setup

As mentioned previously in Section 3, the Global Reporter system used three news sites as data sources for the *English-only Search* selection, namely CNN, Yahoo! News, and ABC News for the experiments. For the *Multi-language Search* selection, the French Yahoo! Actualités, the Spanish CNN Español and the German N-TV supplemented these English sites. Query results from each site were limited to those articles that were published in the past 30 days in order to have all possible retrievable documents from each site over the same time period, without adversely effecting the response time.

Six users with relatively different backgrounds conducted the experiments. Five users were graduate students from two academic institutions in the fields of Business, Computer Science, Dentistry, Mathematics, and Psychology. The final user was a software engineer from industry. On account of their varying backgrounds, the level of expertise in using and clarity in understanding the system varied somewhat.

All users remotely accessed the machine hosting the Global Reporter site within a two-day period. Users were instructed to execute nine fixed queries found in Table 1, once with the *English-Only Search* and once again with the *Multi-language Search* selection. From the results, they were asked to count the number of articles for each news site they found to be relevant answers to the particular query. The results of a Global Reporter search were presented as a list like most search engines, where each result included the title, date, excerpt, and news source. Users were also requested to note the specific event and count for the number of times each news site referred to the event, excluding near duplicate articles from their counts. Finally, they were asked for comments, suggestions and about the potential of cross-language news meta-search like Global Reporter.

Table 1. The nine fixed queries used in the experiments

Number	Query String
1	Portugal champion
2	Microsoft IE security
3	transportation strike
4	kidnapped reporter
5	airplane hijack attempt
6	flood town
7	China champion
8	contaminated water
9	Israel Palestine violence

4.2 Overall Information Coverage

Data gathered from users was used to compare the overall information coverage of news stories between results from the *English-only Search* and *Multi-language Search* search types for the fixed queries in Table 1.

In considering overall information coverage of news stories for particular queries, the total number of stories retrieved by each group of news sites queried in the *English-only Search*, and those queried in the *Multi-language Search* were considered. The details of how the total number of stories for each query and query type were determined will be discussed now.

News stories reported by any news site found in the user data were counted. If users found both CNN and Yahoo! News reported about the French air controllers' strike in one article each, the user would have recorded a count of one each under the two news sites for the results for the "transportation strike" query. Also, if CNN reported about the French air controllers' strike in say three articles; however, the article content varied so that there was an information gain, then the user would have counted three results under CNN for that particular event. Users did not count near duplicate articles. For each search type, each query and each news site, the six users' result counts were averaged. Then, for the *English-only Search* the averages for each query from CNN, ABC News, and Yahoo! News were added. Similarly, for the *Multi-language Search*, the averages for each query from CNN, ABC, Yahoo! News, CNN Español, French Yahoo! Actualités and the German N-TV were summed together.

Figure 2 presents the results for overall coverage provided by each search type for each query using these sums. As expected, the *Multi-language Search*, which supplements news sources from the *English-only Search* by news sources in Spanish, French and German, has at least

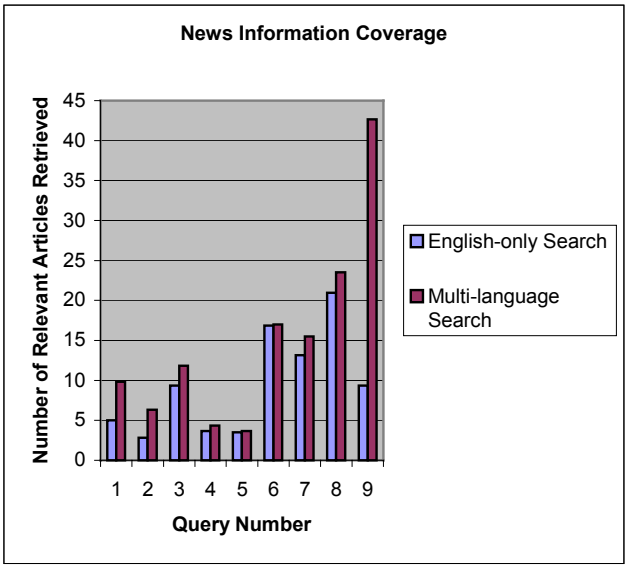

Figure 2. Comparison of coverage from the two search types

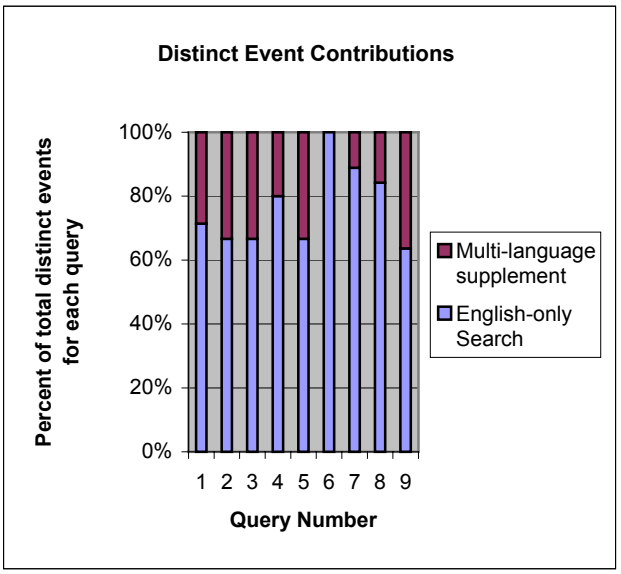

Figure 3. Distinct event contributions by the two search types

the same amount of information coverage as the *English-only Search*. Though particularly, for queries 1 (Portugal champion), 2 (Microsoft IE security) and 9 (Israel Palestine violence) the coverage was significantly greater with the *Multi-language Search* compared to the *English-only Search*. Queries 3, 4, 6 and 7 showed a moderate increase in coverage with the *Multi-language Search*, whereas queries 5 (airplane hijack attempt) and 6 (flood town) did not exhibit much coverage difference between the two search types. While the information coverage results show an improvement with the *Multi-language Search* for the majority of the nine fixed queries that were rather randomly chosen, further experimentation involving a larger sample of queries seems appropriate prior to drawing concrete conclusions.

4.3 Distinct Event Contributions

The data tabulated by users that was described in the previous two subsections was also used to study the difference in distinct event contribution between the two search types by measuring the number of facts found. More specifically, the English-only contribution was considered, as was the multi-language contribution that supplemented the English-only one for the *Multi-language Search*. In the context of this system, a unique fact is equivalent to a specific news event like the French air controllers strike. For each search type and each query, every uniquely identifiable event was counted once regardless of the number of users or sites that identified it. The sum of all the uniquely identifiable events for each search type and query was then considered to be the distinct event count for that search type for that particular query.

The results from determining these distinct event contributions are illustrated in Figure 3. As on the information coverage graph, the query numbers on the x-axis correspond to the fixed queries found in Table 1. The graph readings for query 4 (kidnapped reporter) mean that if for example there were 5 total events in the *Multi-language Search*, 80% or 4 of the unique events were contributed by the *English-only Search* news sources, and the non-English multi-language sites (supplementing the English ones) contributed the remaining 20% or 1 distinct event. Apart from query 6 (airplane hijack attempt), the multi-language news sources supplemented the *English-only Search* distinct events for the other eight queries by 11% to 36%. Consequently, using the *Multi-language Search* appears to provide better distinct event coverage of news stories for most queries than an *English-only Search*. Nonetheless, further experimentation with a greater number of queries should be performed to solidify the initial findings from this smaller sample. Also, experimenting with a non-English base language for the system, may lead to interesting results since most of the web is in English, and resources in other languages are often scarcer.

4.4 User Comments and Feedback

After concluding their Global Reporter runs with the fixed queries, users were encouraged to pose several queries of their choice to get a better feel for the system. Following this, they were asked to indicate, on a scale of 1 to 5, how useful they found a news meta-search engine that gathers results from multi-language sites (translating them before

returning them to the user) to be versus one that just gathers results from English sites only. A score of 3 indicated that a user found a multi-language news search engine as useful an English-only search engine.

Two of the six users rated a multi-language search engine with a score of 3, and indicated that the poor translation quality mostly influenced their rating decision. The remaining four users rated the potential of a multi-language news search engine with a score of 4. Some of these users stated that they liked the idea of such a service; however, would like to see the meta-search engine augmented to include a presentation of different perspectives based on language for the same event.

Suggestions from users included providing the relevance percentage or other relevance score from the ranking for each article whose excerpt was in the results' list, providing support for phrase queries within quotations, providing support for user manipulation of translated queries prior to searching, and improving the system's response time. Some users also suggested adding news sources in non-European languages, and allowing the user to select the specific news sites to fetch results from.

5. CONCLUSION

This paper has presented a flexible basic architectural framework for a cross-language news meta-search engine, as well as properties observed from results retrieved by an English-only meta-search of news sources, and by the English news sources supplemented with news sources in other languages from such a system. An initial investigation of retrieval properties for results from the Global Reporter system shows promising evidence of improved overall information coverage, and an increased contribution of distinct news events using the multi-language search feature of the system. Nonetheless, further experiments with a larger set of queries should be performed to reach more definite conclusions. Intervening with and refining the query translation process may also be worth considering to further improve the retrieval results instead of using machine translation alone, as was discussed in Section 2.

From user reactions, it appears a system like Global Reporter would provide more utility with additional features such as highlighting different versions of the same news story on account of the story source being in multiple languages. Adding additional news sources to the existing system also appears to be a potentially worthwhile investment.

6. ACKNOWLEDGMENTS

Gratitude goes out to Angel Patel, Ayesha Shajahan, I-Heng Mei, Nina Patel, Pranav Shajahan, and Thomas Devanneaux for their considerable time and effort in conducting the tests for the experiments. The author would also like to acknowledge Thomas Devanneaux for suggesting the use of persistent connections in the system, and Andreas Paepcke for his suggestions of German news sites and potential user interface features. The discussions, feedback, and suggestions from Christopher Manning, Hinrich Schütze, and Taher Haveliwala throughout the course of the project were also very helpful in the development of the project.

7. REFERENCES

[1] Abdelali, A., Cowie, J., Farwell, D., and Ogden, W. UCLIR: a Multilingual Information Retrieval Tool. In Proceedings of the Iberamia Workshop, Seville, Spain, November 2002.

[2] CLEF: Cross-Language Evaluation Forum.
http://clef.iei.pi.cnr.it:2002

[3] Eilerman, N., Evans, D., Jin, H., and Qu, Y. The Effect of Pseudo Relevance Feedback on MT-Based CLIR. CLARITECH Corporation, Pittsburgh, PA, 2000.

[4] Fielding, et al. Connections. W3C HTTP/1.1 RFC 2616
http://www.w3.org/Protocols/rfc2616/rfc2616-sec8.html

[5] For Overcoming the Language Barrier in the Use of Internet. NTT Press release. 1999.
http://www.nttamerica.com/news/1999/990224.html

[6] Google News
http://news.google.com

[7] Ithaki Metasearch Engine
http://www.ithaki.net

[8] Lin, C. Machine Translation for Information Access across the Language Barrier: the MuST system. In Machine Translation Summit VII, Singapore, September 1999.

[9] My Hawker Metasearch for News
http://www.myhawker.com

[10] X-Fetch Wrapper. Republica Corporation, Finland.
http://www.x-fetch.com/xhtml/wrapper.html