Clarity

Abstract

This document describes the Clarity project: a 36-month EU 5th Framework RTD project. The project consortium consists of three academic groups and three companies. The project has run for just over two years and here we report on the goals of the project and the results gained so-far.

Introduction

Cross language information retrieval (CLIR) is the retrieval of documents written in one language by queries written in another. Such systems typically work as follows: queries written in one language (referred to as the *source language*) are automatically translated in some manner into the language of the document collection (the *target* language) and retrieval takes place[1]. CLIR has been studied for around ten years. Over the decade, researchers have identified the major problems of translating queries for retrieval; determined a series of approaches to building and utilising a variety of translation resources; and produced a number of retrieval algorithms that address the problems of CLIR ensuring a high quality ranking of documents results from a cross language search. It is a problem area where in a relatively short space of time, academia has made CLIR a workable, though not perfect, technology. Although a great deal of effort has gone into making the technology of CLIR work, particularly for languages for which a great many translation resources exist, very little research studying the users of CLIR systems has taken place, neither has there been much research examining CLIR when languages with few translations resources exist, so-called *low density languages*. The aim of Clarity is to explore such aspects of the subject.

In the rest of this article, usability testing and ways of dealing with low density languages are described followed by conclusions and details on the Clarity consortium.

Users and usability

When considering the users of a CLIR system, one must first tackle an important question. One of the first things that most people will ask when hearing about CLIR, is why would anyone want to retrieve documents they presumably cannot read? There are a number of answers:

- A searcher may not be able to read the target language, has access to a human translator but has too many documents to be translated. By retrieving documents for a particular query of interest and passing the top N only to the translator, may make more efficient use of that person.
- Some searchers may read a foreign language but struggle to query in it, if a CLIR system can search accurately, then it could be a useful solution for such searchers.
- It has also been said of media companies that deliver their output in languages spoken by few people, wish to use CLIR systems. Many of the potential audience of the media company's output will most likely be able to speak another more widely spoken language, often English. Such people are less likely to expect relevant information in their native language and when searching, only query in the more popular language. The media companies want to attract the audience back to their "minority language content" through CLIR.
- Even people who can read and write many languages (polyglots) may want CLIR system as they could enter a query in once and have documents returned written in all the languages they know. Such people constitute the user groups in the Clarity consortium.

It was the needs of the last group that informed the design of the Clarity system: journalists and translators, from two of the companies in the consortium, are polyglots who wish to search in multiple languages, but would prefer to enter their query in only once. A user-centred interface design methodology was adopted. By monitoring existing practices with monolingual search engines and discussing interface mock-ups at the two companies (BBC Monitoring and Alma Media), it was then possible to compile a list of user requirements. These requirements were subsequently used to create a

[1] It is possible to translate the whole document collection instead of each query, but this approach is not often pursued.

prototype user interface. At present, the prototype has been tested at the two consortium sites. In the final year of Clarity, the system will be extensively evaluated and user-tested on site with the user groups. ***Figure 1*** shows the interface in its current state.

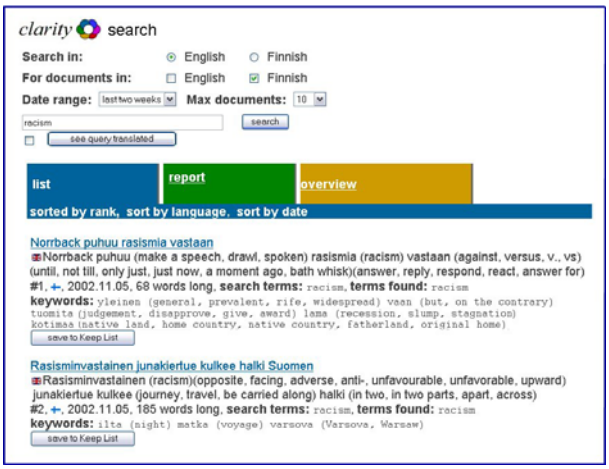

Figure 1 Prototype user interface displaying results of a query, sorted by rank. Each result contains a document summary and a number of keywords, displayed in both the target language and the source language translation.

In the final version of the system, two additional interface features will enable users to better interact with a CLIR system.

- The first is a document-organisation tool based on concept hierarchies that show a dynamically generated hierarchy of words extracted from the retrieved documents arranged with general terms at the top leading to related more specific terms below. The terms can be shown in which ever language the user prefers.
- The second feature is a retrieval report produced after each search. Here key terms from documents and the styles of the documents retrieved (analysed automatically) are described.

The system has been implemented using a novel distributed architecture that allows the separate development sites (University of Sheffield, University of Tampere, SICS, Tilde) to build their component parts locally and integrate them through an Internet based protocol. Communication between the components is through SOAP. This is proving to be a most valuable approach, as Clarity has been able to start building a prototype much faster and cheaper than would have been the case if a conventional approach of centralising integration had been taken.

Low density languages

Much existing CLIR research assumes the availability of translation resources that, for the majority of languages, do not exist. A second focus of Clarity is low density languages: languages for which there are few translation resources. Here two approaches to aiding CLIR are being investigated: first, using multiple bilingual dictionaries for transitive cross-language retrieval via one, possibly many, so-called pivot languages (see figure 2); and second, using advanced n-gram techniques for translating words not in dictionaries (e.g. proper names). The two methods are now described.

Triangulated translation

Most approaches to CLIR assume that resources providing a direct translation between the query and document languages exist. Such an assumption, however, is often false. In such cases, an intermediate (or *pivot*) language provides a means of transitive translation of the query language to the document via the pivot, at the cost, however, of introducing much error. Clarity has addressed a novel approach of translating in parallel across multiple intermediate languages and fusing the results (see ***Figure 2***). Such a technique removes the error, raising the effectiveness of the tested retrieval system, up to and possibly above the level expected, had a direct translation route existed. Across a number of retrieval situations and combinations of languages, the approach proves to be highly effective.

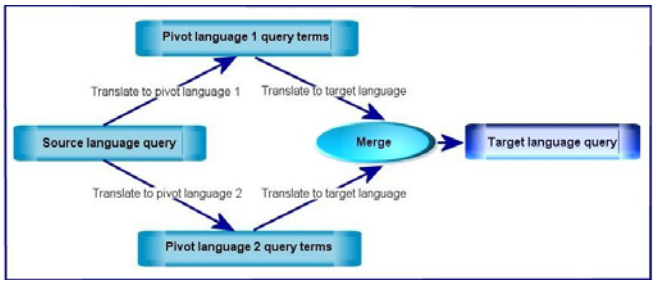

Figure 2 Lexical triangulation – by translating the original query into multiple pivot languages then into the target language, a set of query terms is produced. When these query terms are merged, this produces a more effective target language query than when a single pivot language is used.

One question that might reasonably be asked of the triangulation work, is do any CLIR users actually need or have the translation resources for triangulation via multiple pivot languages? It turns out that such examples exist. Good translation facilities from Latvian and Lithuanian into English exist, as do translation resources into Russian, however, no other translation resource for other languages currently exist online. For these languages triangulation is a necessity. At the half-way point in the Clarity project, it extended its scope and language coverage through the addition of a Latvian partner, Tilde: a Baltic language technology company. Now, the two Baltic languages will, via triangulation, be added to the Clarity system.

Untranslatable words translated

A common problem when translating a user's query via some form of translation resource, for example, a bilingual dictionary is finding that some or all of the query words do not occur in the dictionary. Such a problem is particularly common with Proper nouns: names of places, and in some languages, names of people are commonly spelled differently. Clarity is exploring the use of both n-gram and skip-gram techniques to locate likely candidate translations. A user enters a query, one of the words of which is not in a translation dictionary. Words in the document collection that appear to be close misspellings of the query word are chosen as possible candidate translations and added to the query. Even though the query becomes a combination of correct and incorrect translations, initial results testing on existing document test collections show the technique to be promising, indicating that Information Retrieval systems are tolerant of a level of error present in a query.

Conclusion

Cross language retrieval is a topic that has been long studied with clear progress made. However, certain key aspects of CLIR have not been studied – usability and low density languages – success in both is key if CLIR is to be adopted by those who wish to use it. Clarity is building a retrieval system based on such research. Results indicate the approaches taken are promising.

Clarity details

For further information about Clarity including publications resulting from its research work, see the Web site, clarity.shef.ac.uk. The consortium partners are:

- The University of Sheffield, UK, departments of Computer Science and Information Studies
  - Main investigators and joint project coordinators, Robert Gaizauskas & Mark Sanderson
- The University of Tampere, Finland department of Information Studies
  - Main investigator, Kalervo Järvelin
- SICS – Swedish Institute of Computer Science, Stockholm Sweden
  - Main investigator, Jussi Karlgren
- Alma Media, Helsinki, a large Baltic state media company
- BBC Monitoring, Reading, UK, a private company attached to the BBC that monitors foreign broadcast news.
- Tilde, Riga, Latvia, a Baltic language technology company.