LTC Communicator

Philippe NOZAY and Adriane RINSCHE

Language Technology Centre Ltd
5-7 Kingston Hill
Kingston upon Thames
Surrey, KT2 7PW
UK

pnozay@langtech.co.uk

arinsche@langtech.co.uk

1 Multilingual applications

1.1 Necessity

With the awesome growth of the Internet, more and more applications and web sites are becoming truly global and have to handle data input from several languages. These applications have to provide some sort of translation to enable people to communicate together whatever their mother tongue.

1.2 Issues

Machine translation is the answer to instant translation. However the output quality is often low and very often difficult to understand if not misleading. Human translation provides excellent output but it very slow and expensive. Other tools in the translation industry such as translation memory provide very good output but need human interaction. Most web sites/applications would like to provide instant, understandable and cheap translation. This is, however, difficult to achieve unless a machine translation system has been tuned to a well defined domain and integrated into a set of tools automating the whole interaction between (in our case) users of a software product with support issues formulated in one language and solution engineers providing solutions in another language.

1.3 LTC Communicator

We have developed the LTC Communicator to provide a practical solution to these questions. It offers multilingualism to any application or web site which requires automatic translation at an intelligible level, due to a combination of customised tools used in a well defined automated sequence. This system is designed as a web service accepting XML requests. It receives questions/answers wrapped in a XML request and returns the translation in an XML message. LTC Communicator is a product that includes the automation server and the LTC translation memory. Systran or any other machine translation system can be plugged in depending on customer preference or the language combinations required. LTC Communicator can be purchased or used as a web service hosted by LTC.

1.4 Fields of application

In order to provide "understandable translation" the system has been tailored to technical questions/answers (trouble tickets/solutions). It is at its best in technical support applications such as a web help desk where a user posts his question in his own language, for instance English. The LTC Communicator then translates the question and the engineer can now read it in his language such as German. When the engineer posts the response in German, LTC Communicator translates it into English for the user to read.

The system can also benefit technical search engines where the user enters his search criteria in his own language, for instance English. Then the LTC Communicator translates the criteria into several languages. The search engine queries its database for any of the translated criteria. If matches appear, the summary of the document as produced by the search engine can be translated back into English so that the user can judge the relevance of the result.

LTC Communicator could be used in any application needing to provide multilingual communication between users. To ensure reasonable output, the vocabulary should be restricted to specific domains if possible.

1.5   System background

The LTC Communicator emerged from a trial carried out within the framework of the EU project IMPACT, which was very successful in terms of system design and output quality in the well defined software domain selected (Helpdesk application for a CRM product).

The full name of the IMPACT project is "Impact of a web based multilingual portal for helpdesk applications", and it was a trial project in the framework of the European "User-friendly Information Society" programme, project number IST-2000-30110, funded by the European Commission in the 5th Framework Programme in the slot for Human Language Technologies. CAS Software AG, Germany, was cooperating with Language Technology Centre (London), the translation and language technology specialist, as well as Infoworld Srl (Milan), the genesisWorld sales partner for Italy, and the systems house New Emphasis (Greece).
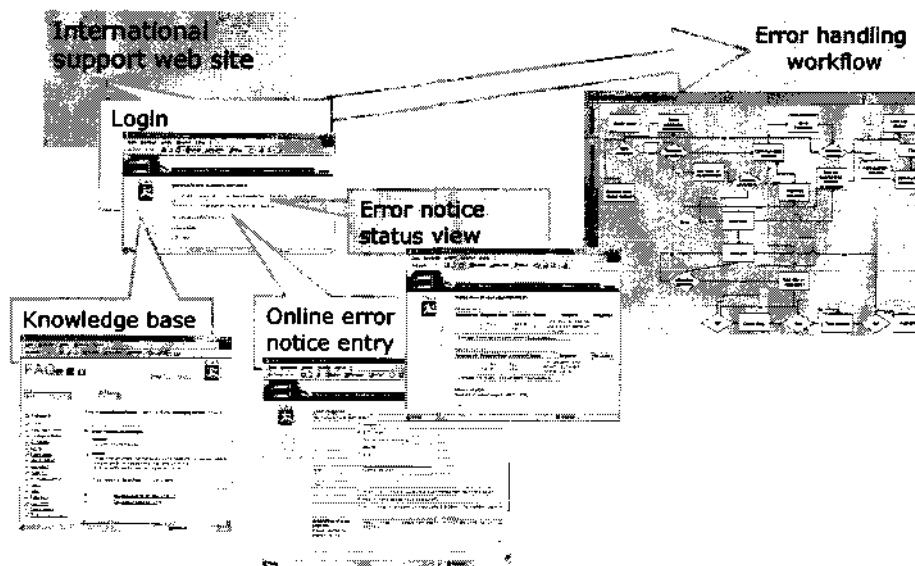

Diagram of the multilingual helpdesk

IMPACT links a helpdesk based on genesisWorld via the Internet to a multilingual web service developed by LTC using Systran MT. A customer can report a problem via the web, entering the information in English. The trouble ticket is automatically translated into German before it is viewed by the helpdesk operative. After processing, the helpdesk agent's solution, which is in German, is automatically translated back into English in the same way and is then available for the customer in a protected area of the web pages. The customer can also find information there on the current status of the problem they reported. The customer additionally has access to a multilingual knowledge database (FAQ) here.

On the basis of its genesisWorld solution, CAS Software operates a helpdesk for processing problems reported by the international partners for genesisWorld solutions. Via the web, the partners can initially search for existing solutions and report problems if required. These are processed directly in the CAS in-house genesisWorld system in accordance with a structured workflow.

In IMPACT, genesisWorld exchanges the content of problem reports with the LTC system for automatic translation, using an interface based on Internet protocols (http, XML). The LTC

system processes the language entry and compares it against existing translations.

The tools used guarantee a language processing quality that makes everything easily understandable for the user and ensures efficient communication. They are also considerably different in terms of architecture and complexity from the consumer web based translations services, the results of which are often unsuitable for professional users. For users who require a translation of publishable quality rather than an understandable result, the system can be set temporarily or permanently to include human revision by professional, human translators.

If completed solutions are to be stored permanently in the solution database, they are translated from German into English by the LTC software and manually revised to yield publication quality, and they are then available in both languages.

The trial was carried out in English and German, the LTC modules were developed into a product called "LTC Communicator", and this product is now available for many language combinations.

2 System Architecture

2.1 Overview

To meet the contradictory requirements of speed versus quality the system makes use of a variety of technologies:

- Machine translation (MT) is used to provide instant translation

- Translation memory (TM) is used to improve the quality of repetitive text

- A spell checker can be used to improve the input quality, which would then improve the chance of TM and MT matches.

- Custom dictionaries for the MT are built to cover the specific domains of application

- Databases populated with bilingual documents on the domain, software, existing faqs... are used by the TM

- Input guidance and validation is used to limit the amount of free text offered to the user and also to reduce the grammatical complexity of the input. The web interface is structured in such a way that the user needs to provide basic information about the version of the software he/she is using, the operating system used etc in drop down lists in order to limit the free text input to the relevant text. In addition, style guidelines were developed for solution engineers and system integrators on the user side to avoid complex syntactic and grammatical structures that are difficult to handle by a machine translation system.

- Optional human revision if publishing quality is required. The raw translation is then automatically forwarded to a reviser for post-edition; recommended when the system

is first introduced to tune the TM memories and the MT dictionaries, but also if the application/web site needs to publish the information in many languages to a wide audience.

- The post-edited MT output is automatically fed back into the TM.

The machine translation system used in the trial was SYSTRAN, the TM system used is an in-house development (LTC), used for batch processing only, and replacing identical matches with target text. Interactive TM processing is not desirable within this workflow as the system automatically processes new text with no TM matches via the MT system. All human post-editing results are fed back into the TM.

A Systran user dictionary was customized with the all the key software terms of the CRM system, and the translation memory was populated with bilingual online help files and the software documentation, and the content of a bilingual Frequently Asked Questions Database.

The output quality was initially monitored by human post-editors who added dictionary entries to the Systran custom dictionary and populated the TM database with post-edited material.

2.2   Simplified architecture
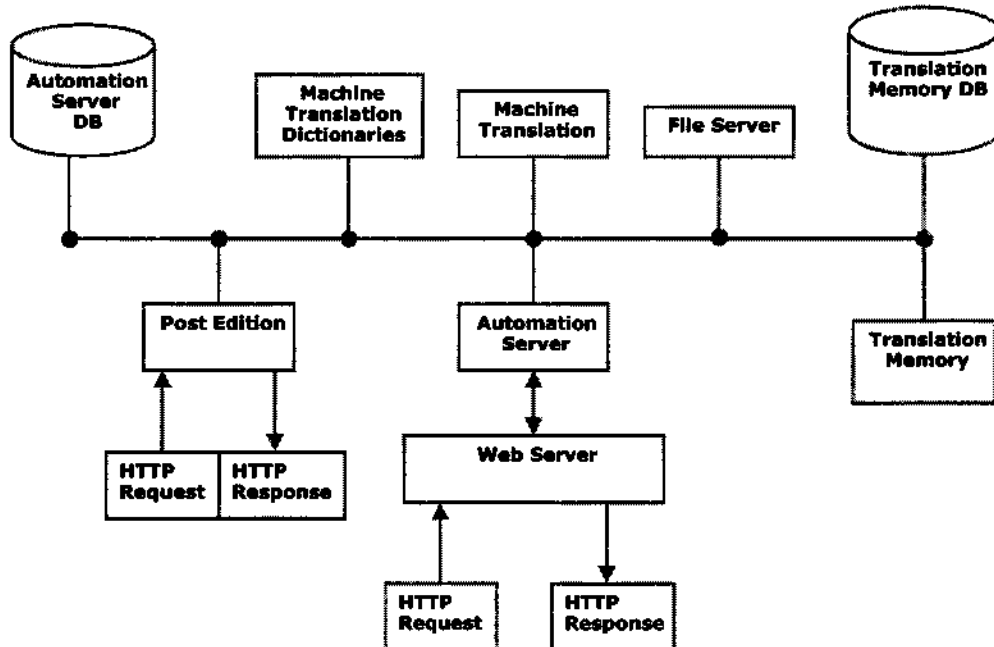
2.3   Workflow

Depending on the request parameters (priority and revision level), one of the following workflows is instantiated. The workflow server decides what the next step is between each

process. The system was optimized so that the duration of a translation processed as described below amounted to only a few seconds in the trial (excluding human post-editing).

2.3.1   Automatic process

1. Request reception on LTC Web server
2. Workflow server
3. TM pre-translation (XML => RTF)
4. Workflow server
5. MT translation (RTF)
6. Workflow server
7. TM Import MT (RTF)
8. Workflow server
9. TM rebuild (RTF => XML)
10. Workflow server
11. Return automation server
12. Workflow Server
13. Resource manager

2.3.2   Semi-automatic process including human post-editing

1. Request reception on LTC Web server
2. Workflow server
3. TM pre-translation (XML => RTF)
4. Workflow server
5. MT translation (RTF)
6. Workflow server
7. TM Import MT (RTF)
8. Workflow server
9. Revision automation server (RTF)
10. Workflow server
11. LTC London Revision Web Server (RTF)
12. Workflow server
13. Revised document on LTC Web Server (RTF)
14. Workflow server
15. TM import Revision (RTF) (import in both language pair)
16. Workflow server
17. TM rebuild (RTF => XML)
18. Workflow server
19. Return automation server
20. Workflow server
21. Resource manager

For example, for an English-speaking help desk supporting German users in the fully automated way described under 2.3.1:

- The user enters a trouble ticket in German, via the software company's support portal;

- The trouble ticket is routed through the LTC-Communicator translation components;

- The request is displayed to the support engineer in English;

- The engineer prepares the solution, also in English, to be automatically routed back through the translation environment;

- The user can then view the solution (or status information) in German.

In addition, the user can set the system to route the translated request and/or solution to an optional human post-editing service either by default or based on some other criteria such as every 5th message or messages with low translation quality.

Our translation service does not use SOAP as a protocol but instead uses XML over the standard HTTP Post protocol. It was necessary to convert the XML messages to rtif as the version of Systran on which the trial was carried out did not accept XML input.

2.4 Technologies used

The following diagram shows the different technologies used by the system and their advantages.
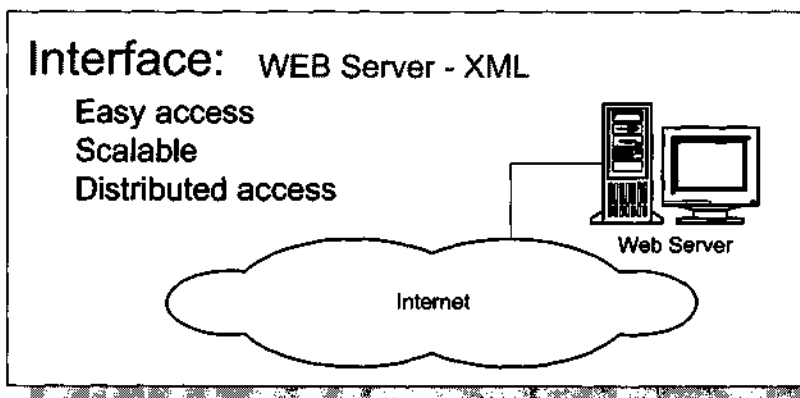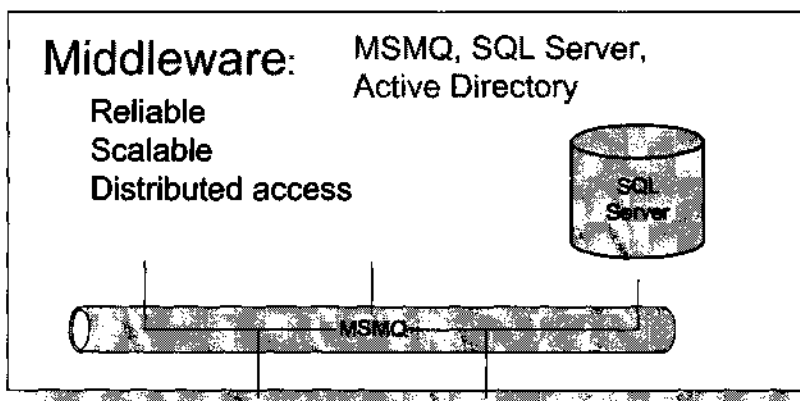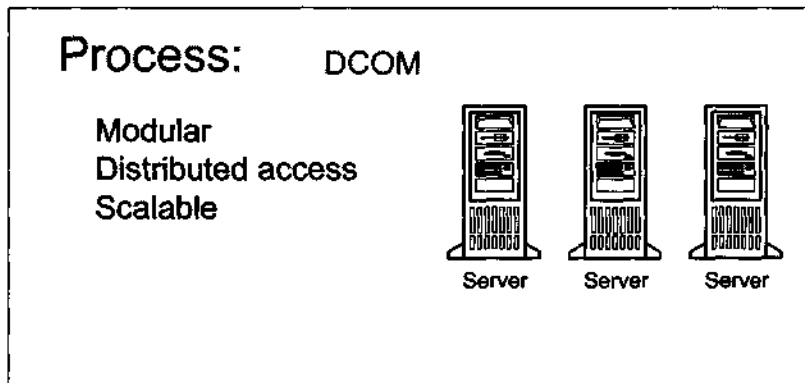
2.5   Interface

2.5.7   *HTTP Protocol*

To access the translation system, the HTTP protocol is used. The application server sends (posts) the text to be translated to the LTC Communicator server through a web server. HTTP allows for

- Distribution of processes: the LTC web server can be hosted anywhere in the world

- Concurrency: the LTC web server can receive many concurrent messages simultaneously

- Workload balancing: multiple LTC web servers can be grouped in clusters to handle more concurrent messages

- Implementation simplicity: HTTP has existed for years and it is nowadays very easy to implement web servers and to generate HTTP invocation (POST) from clients

After translation, the translated text needs to be returned from the LTC automation server to the application server. For the same reasons, we decided to use Http invocation to send back the translation.

2.5.2   *XML message*

The message containing the text to be translated and posted to the web server is in XML format. XML allows for easy exchange of structured information between applications. With XML, it is possible to add tag information describing the data contained in the message avoiding confusion as with text files. New tags can be added in future versions without making older applications incompatible. XML parsers are now widely available and XML is therefore relatively easy to implement.

2.6   Technical Advantages

In addition to the translation requirement, the system has the following advantages:

- Scalable: possible to add/remove servers according to the load

- Load balancing: automatically balance the load between several servers

- Easy support: quick to pinpoint errors if any

- Modifiable workflow: possible to add new steps or replace existing ones, for instance MT

3 Conclusion

LTC Communicator provides an easy solution to enable any web site/application to handle multilingual communication between users by offering:

- An automated support environment for many web-based multilingual application environments

- Plugs into any MT system, depending on customer preference and language combination

- High quality output due to customisation in well defined domain

- Automatic handling of otherwise time-consuming management of typically short and informal communication

- Intelligible output quality; no need for publishable quality, therefore post-editing not needed for most question/answer.

- Style guidelines for authoring to further improve quality of target messages

In the help desk industry, it allows software publishers to run a support centre from headquarters rather than training and paying many support engineers in the different countries or hiring multilingual support engineers centrally.