Automatic Multilingual Subtitling in the eTITLE project

Maite Melero
(maite.melero@upf.edu)
GLiCom (Universitat Pompeu Fabra)

Antoni Oliver
(aoliverg@uoc.edu)
LPG (Universitat Oberta de Catalunya)

Toni Badia
(toni.badia@upf.edu)
GLiCom (Universitat Pompeu Fabra)

Abstract

This paper presents the Multilingual Translation Service of eTITLE, a European eContent project, which has produced tools to assist in the multilingual subtitling of audiovisual material through the web. The eTITLE Translation Service combines state-of-the-art Machine translation and Translation memories, which may be tailored to the customer needs. The user can choose to use only Machine Translation, only a certain Translation Memory (or all available) or to use an integration of both approaches, which is the default option. In this latter option, only if no translation above a certain threshold is found in the Translation Memory, the original sentence is Machine Translated. The language pairs that have been developed in the framework of the eTITLE project are: English-Spanish, Spanish-English, English-Czech, Catalan-Spanish, Spanish-Catalan English-Catalan and Catalan-English.

1. Introduction

eTITLE is a two-year project that ended in February 2006 and has created web-based solutions that allow media content owners to exploit it internationally, through multilingual and cross-platform localisation. The project builds on a spectrum of newly available technologies for Digital Asset Management, Automated speech-to-text, Machine Translation, Sentence Compression, Subtitling Automation and Metadata Automation to provide a much more cost-effective digital workflow.

Subtitling is, together with dubbing, the main form of translation or "language transfer" in television and other audio-visual environments. Language transfer involves more than facilitating the viewer's comprehension of unfamiliar language. The European Commission has, for example, recommended subtitling as a means of improving knowledge of foreign languages within the European Union. With the continual ramping up of quantities of subtitling produced within the EU as a combination of regulatory changes and recent technological advances creating new channels requiring subtitling, broadcasters and media owners are aggressively seeking lower prices for the subtitle service.

In this framework, the eTITLE approach may increase the efficiency of current subtitling by automating various processes within the workflow. The eTITLE project

has been a joint collaboration of the following partners: TMR, a UK-based media services company that has coordinated the whole project and has been in charge of putting in place the eTITLE prototype; GLiCom, a Computational Linguistics group at the Universitat Pompeu Fabra (UPF, Barcelona) that has implemented the Translation Service part; and the Interactive Technology Group, also from the UPF, which has taken care of the Compression and Subtitle Placement modules. The rest of the partners are: TVC (Catalunya, Spain), MTV Network Europe (Spain) and LS Productions (Czech Republic) as content providers and user partners.

In its multilingual aspect, one of the most appealing features of the eTITLE system is the integration of Machine Translation in the subtitling environment. Attempts at automatically translating subtitles are quite rare, for a recent example see Musa[1] (Piperidis et al.), despite some claims about the appropriateness of the subtitling domain for the current state of the technology in Machine Translation (Popowich et al, 2000).

Machine Translation systems are generally large scope and robust, but their automatically generated output can sometimes be far from acceptable. Translation Memories are large databases of previously translated sentences that may occur again in successive source texts. Since the output is originally generated by humans, the typical errors and noisy output that MT systems sometimes produce are avoided. However, they are less robust than MT, that is, more often they are unable to provide a translation (i.e. their coverage is smaller).

Because the shortcomings and advantages of TM and MT systems are complementary, a considerable improvement of the translation results is to be expected from an integration of both.

2. Multilingual subtitling user requirements

Subtitling may be multilingual (or interlingual) and monolingual (or intralingual). In multilingual subtitling, the subtitle renders the translation in the target language of the dialogue or speech in the source language. Its main use is the internationalisation of audiovisual material while keeping the original audio version, as opposed to dubbing, which replaces the original audio with audio recorded in the target language. In monolingual subtitling, the subtitle transcribes same language speech. This is also called closed captioning, especially in the US. It is mainly done for the benefit of the deaf and hard-of-hearing, as well as for improving reading skills of children or Second Language learners. From a technical point of view both types of subtitling share similar requirements. They are constrained by time and space, and they have to meet certain legibility constraints that affect their size, position on the screen, number and length of lines and segmentation of lines in self-contained units. Their main difference is that multilingual subtitling requires translation to happen at a certain point in the process.

2.1 Scenarios considered by eTITLE

A range of scenarios involving the needs of several classes of users have been considered in the context of eTITLE. These requirements cover the cases of the user partners of the project:

a. Broadcasters with strong regional identities and local interest material, such as TVC, Televisió de Catalunya (Spain);

[1] http://sifnos.ilsp.gr/musa/

b. Media groups that both originate and deliver content in may European countries, such as MTV Europe;
c. Companies localising multilingual media (currently in the form of DVD) for a large territory covering several language areas, such as LS Productions.

With each of the partners operating in a distinct location, market and environment, it was felt they constituted a good range of sample data to provide a framework for the user requirements to start from. With this aim, each partner's workflow, systems and expectations were examined in detail, combining onsite analysis with the results from technical surveys. A number of industry experts were also consulted. Hence, the analysis of the main partners was supplemented with numerous discussions between subtitlers, subtitle software providers and video professionals.

2.1.1. TVC

TVC uses subtitling in its live newsroom broadcasts as well as in its day-to-day programming. Overall, TVC currently subtitles 28% of programming, but would like to substantially increase this figure. All subtitling is performed in-house, using bespoke systems created around the TVC workflow. The bulk of the material to subtitle is already in Catalan; the subtitles are predominately used for hearing impaired viewers and output via teletext. However, multilingual subtitling is also envisaged in several scenarios, such as emissions broadcasted in original version (mainly English) and subtitled in Catalan; or web-based distribution of in-house produced material to other markets, such as Europe or the Spanish speaking countries.

Therefore, eTITLE addresses two problems facing European broadcasters: how to reuse their very considerable archives and exploit them in new markets; and how to draw effectively on global information resources.

2.1.2. MTV Europe

The Spanish MTV channel airs a mixture of domestically produced programming and foreign shows, such as "Jackass", thus the main subtitling requirement is for English to Spanish. In 2003, MTV Spain transmitted 65 hours of material it had translated and subtitled, along with 170 hours of content subtitled in Latin America. MTV has a unique style of expression that immediately represents its brand and image. The language employed is continually adapting and is interspersed with slang. It is vital this essence is captured and reproduced in the subtitling of its material.

The MTV case study indicates that a number of matters need to be addressed:

- The need for customisation of viewing content to suit the geographic location in which the programme is being broadcast;
- The importance of programme impact on the subtitle viewers;
- The digital transference of programming across regions;

The issue of web-based media delivery needs to be addressed to allow subtitling to be brought in house and create a more efficient system.

2.1.3. LS Productions

LS Productions offers a dedicated subtitling service for broadcasters, film distributors and media owners. It transcribes from Czech for the hearing impaired but most frequently it subtitles English to Czech. Its main work is subtitling DVD titles and

subtitling for 35mm cinema print but it also has the facilities to provide subtitling straight to video using a character generator.

2.2. eTITLE as a web-based subtitling service

The business and technological models developed by eTITLE have looked to provide automation into the subtitling process of the different scenarios. The possible solution scenario demonstrates that a client-server architecture is the best choice.
In the initial stages of the eTITLE project, a client ingest station capable of encoding from tape and uploading directly into the subtitling system was envisaged. However, the study of real users necessities, as well as the increasing prevalence of files in digital format observed in all environments, recommended a shift in that model.
Therefore, rather than compete in the ingest stations market, eTITLE defines itself as a web-based multilingual subtitling service. The system, accessed by the client from any Internet connected machine, stipulates a script and video file is ingested before any processing takes place. After all the interactive iterations of the subtitling process have taken place through the web interface, the system outputs its own native export format (ETL) as well as common subtitle formats (STL).
The idea is that the resulting 'localised' media can then be distributed to any suitable platform to maximise the potential audience and revenue, and to enable the possible migration from traditional broadcasting to IP-based delivery. Thus, where a client has an existing subtitling station, the eTITLE product operates up-stream, slashing the preparation time required in the subtitling process. Where no existing subtitle software is present, eTITLE exports finished subtitles suitable for the client's application.

3. Architecture of the eTITLE system

The eTITLE server-side system is composed of an enhanced set of modules integrated together using the SOAP protocol. Figure 1 below shows graphically the main layout of the different modules. There exists also a module controller, which is responsible for the co-ordination and control of the jobs in the system and the processes being carried out on them.

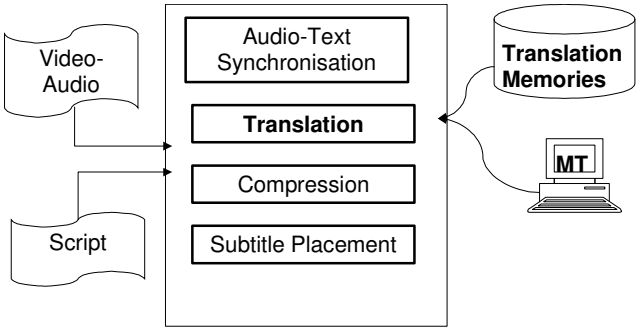

Figure 1: Architecture of the eTITLE system

3.1. Script Synchronisation

The Transcript Alignment module takes a text and a media file as an input and locks the script to the media file by assigning timecodes to the different elements in the script.

The resulting file contains a start and end timecode and an index for each word in the script. This gives a large amount of flexibility to the next processes that are carried out as it allows for the segmentation of the script to occur wherever it is needed, as the timecodes exist for each word.
The transcript alignment module is implemented only for English and makes use of the Virage[2] video analysis technology.

3.2. Translation

The translation engine combines input from two sources:

- Machine translation and
- Translation memories

In the current eTITLE Translation service, the user can choose to use either only Machine Translation, only a certain Translation Memory (or all available) or else use an integration of both approaches, which is the default option. In this latter option, only if no translation above a certain threshold is found in the Translation Memory, the original sentence is Machine Translated. The threshold of similarity is a parameter that must be set by the user, the default being 90% (the lower the threshold, the greater is the probability of finding a translation in the TM; the higher is the threshold, the higher the degree of confidence on the translation).
The language pairs that have been developed in the framework of the eTITLE project are: English-Spanish, Spanish-English, English-Czech, Catalan-Spanish, Spanish-Catalan, English-Catalan and Catalan-English.

3.2.1. Machine Translation

The machine translation service retrieves a translation made by a third party MT system. This service can call different web services depending on the language pair, or use licensed systems installed in the server.
The MT systems used for the available language pairs are:

- WordMagic, for the pairs involving English and Spanish (Cancelo, 2000)
- Internostrum, for the pairs involving Spanish and Catalan (Canals et al., 2001).
- Microton, for the English-Czech pair[3].

For the pairs including English and Catalan, we have taken advantage of the typological similarity between Spanish and Catalan and have used Spanish as pivot, so that translation is achieved by feeding WordMagic with the output of Internostrum (Catalan-English) and the other way round for the other direction.
The translation unit or segment is usually a sentence or a subtitle. Larger pieces of texts can be sent, but for language pairs using a free web service there is a limit in the length of the piece of text.
Internostrum and Microton are accessed as a web service via the SOAP protocol (this requires to have the SOAP Lite package installed[4]). WordMagic is locally installed in the server and is accessed through an SDK[5].

[2] http://www.virage.com/
[3] http://www.eurotran.cz/
[4] http://www.soaplite.com/
[5] http://www.wordmagicsoft.com/

3.2.2. Translation Memories

This module returns one or more matches of a given segment from a translation memory, with a similarity index higher than a given minimum. If no minimum similarity is given a default of 90 % is used.
The initial Translation Memories developed for the eTITLE translation module, out of parallel text are the following:

- Catalan-Spanish / Spanish-Catalan: The Catalan-Spanish Translation Memory has 263,721 segments, totalling 14 million words (7 per language). The corpus comes from two different sources: the bilingual version of the newspaper El Periódico de Catalunya, and the bilingual version of the official bulletin from the Catalan government (Diari Oficial de la Generalitat de Catalunya). The texts have been automatically downloaded from the Web and have been aligned using only HTML marks as reference. In this case, the manual revision of a sample (2899 segments) shows that the confidence score is 95,8%.
- Spanish-English / English-Spanish: The English-Spanish Translation Memory has over 4 million words (2 per language). It consists of thousands of bilingual United Nations documents downloaded from the Web and aligned using an adapted version of Moore's algorithm (Moore, 2002). To evaluate the confidence rate of the alignment, a sample of 713 aligned segments has been manually revised. This manual revision has detected only 8 segments misaligned, yielding a confidence rate of 98,2% for the overall alignment.
- English-Czech: The English-Czech Translation Memory has over 13 million words. It consists of bilingual subtitles downloaded from the Web, corresponding to 857 films. The subtitles have been aligned using a hybrid strategy: first taking into consideration the time codes of both versions and then, using an adapted version of Moore's algorithm for the rest of the subtitles where the encoding of time codes was not compatible. A manual revision of a sample of 1000 segments shows a confidence score of 97%.

The Translation Service foresees that the users develop with time their own Translation Memories, which benefits the system's customisability. A more detailed technical description of this Service is presented in section 4.

3.3. Compression

The compression module (Bouayad-Agha et al., 2006) has been implemented for Catalan and English and comprises the following processes:

1. Number to digit conversion
2. Non-destructive compression (e.g. acronyms)
3. Destructive compression (e.g. elimination of repetitions, proper names and interjections; reduction of subordinating sentences and verbal expressions,...)

The compression modules takes as input the raw text which has been augmented in a previous stage by subtitle and line delimiters based on speaker demarcation and word-based timecodes. It consists of four main phases that apply in turn:

- Linguistic pre-processing: The input is tokenized, tagged and chunked.
- Compression rate determination: This phase involves calculating the compression rate per subtitle and per line based on maximum number of characters as specified in the compression control parameters. This information can be added as attributes to the subtitle and line delimiters.

- Compression candidates production: For each subtitle/line requiring some compression, a series of independent programs, each of which corresponding to a different compression strategy, is applied. Each program outputs a set of text chunks with compression suggestions.
- Compression candidates selection: This module is in charge of recomposing each sentence or subtitle as a list of compression suggestions in decreasing order of confidence, given the various compression suggestions of its constituents, the compression rate per line and per subtitle, and the confidence level of each strategy (or even possibly of each suggestion). The output of this module is a list of segments with suggestions in decreasing order of preference.

3.4. Subtitle Placement

Subtitle placement is carried out after every process to ensure that users are able to view the most accurate subtitle layout for the processed material. The display in the edit is divided up based on the subtitle demarcations suggested in the subtitle placement module.
The basic geometric segmentation of subtitles splits up subtitles every 70 characters. In the balanced mode, splitting is performed by balancing the number of characters per subtitle.
Lines and subtitles can also be segmented using linguistic factors, combined with the balanced geometric mode of segmentation. At present, linguistic factors have been implemented for English, Catalan and Spanish. For other languages only the geometric segmentation applies. The segmentation algorithm uses a set of keywords (one per language) that bounds words or penalizes bad boundaries. The best boundary point is computed in combination with geometrical rules.
An undefined number of subtitles can be grouped in a single subtitle following certain constraints, such as same speaker, short temporal distance and size restrictions.
For subtitle placement it is necessary to identify temporal conflicts and solve them in the best way possible, in order to control duration and temporal collisions between subtitles. Colour assignment to speakers tries to minimize colour collisions. Each speaker has an associated colour. The most standard colours are assigned to speakers with a large number of occurrences. If the number of speakers and/or colours is high, the best combination is calculated.

4. Translation Service description

In this section we offer a detailed technical description of the Translation Service. The main characteristics of this service are the combination of Translation Memories and Machine Translation, and the fact that is configured as a distributed environment, i.e. translation memories and machine translation engines do not need to be installed in the same computer. In addition, different machine translation engines can be used to translate different language pairs.
The main components of the Translation Service are:

- **Translation Controller**: controls the translation task and communicates with the Machine Translation Service and the Translation Memory Service
- **Machine Translation Service**: performs a translation request to a Machine Translation System

- **Translation Memory Service**: returns one or more matches from a given segment with a given similarity or higher from one or more translation memories

In the next figure we present a schematic view of the Translation Service:

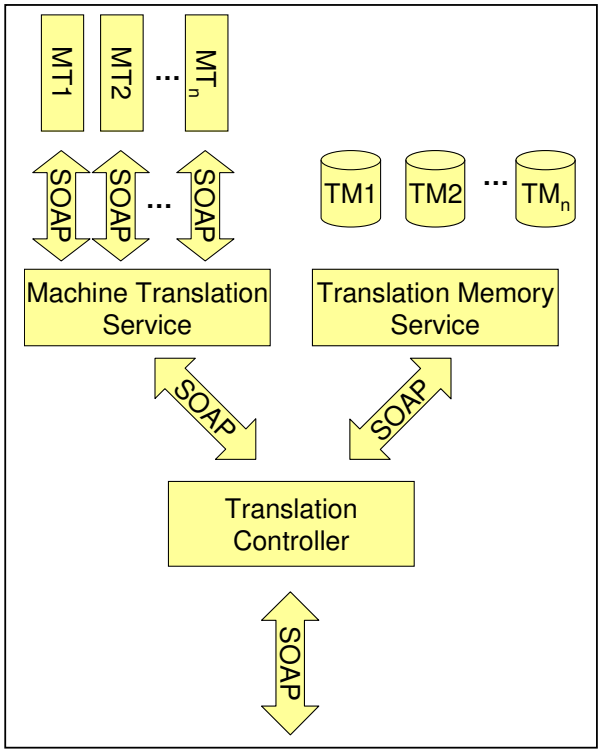

Figure 2. Schematic view of the Translation Service

The client application, in our case the automatic subtitling system, sends a translation request to the Translation Controller using the SOAP protocol. Typically, the Translation Controller will send a request to the Machine Translation Service and to the Translation Memory Service and will send a set of translation candidates to the client application. There are many other options, for example, the Translation Controller can decide which candidate is the best one and return only one translation; it's also possible to use only the Machine Translation Service or the Translation Memory Service.

Using a distributed system has many advantages. The Machine Translation System can physically be in different computers and can also be replicated, so the overall system is very robust. New Machine Translation systems may be easily added and some free translation services may also be used. The Translation Memory Service can reside in the same computer where the Translation Controller is or in a different one. More than one Translation Memory Service can be used. This possibility is useful to speed up the process when a big number of requests are made simultaneously. An unlimited number of Translation Memories can be used at the same time, and, as we will explain later in this document, the technology employed allows for the use of very big translation memories.

Therefore, the design of the system is very flexible and allows its application in multiple situations. The programs are written in Perl and use standard libraries so they work well in multiple platforms (Windows, Linux, etc.).

4.1. Distributed translation environment using SOAP protocol

SOAP (Simple Object Access Protocol) is a lightweight protocol intended for exchanging structured information in a decentralized, distributed environment using XML technologies. The client program sends a request to the server program and the server returns an answer. Both the request and the answer are XML documents. In the next two figures we can see a simple example of a client request and the server's answer.

```
<SOAP-ENV:Body>
       <namesp2:translate_segment xmlns:namesp2=
        "http://sindarin.upf.es/MTservice">
              <c-gensym4 xsi:type="xsd:string">demo</c-gensym4>
              <c-gensym6 xsi:type="xsd:string">
                     Hello, good morning
              </c-gensym6>
              <c-gensym8 xsi:type="xsd:string">eng2spa</c-gensym8>
       </namesp2:translate_segment>
</SOAP-ENV:Body>
```

Figure 3. A SOAP request example. The parameters are marked in bold letters: user "demo", a sentence to translate "Hello, good morning" and a language pair eng2spa".

```
<SOAP-ENV:Body>
       <namesp2:translate_segmentResponse xmlns:namesp2=
        "http://sindarin.upf.es/MTservice">
              <s-gensym8 xsi:type="SOAP-ENC:base64">
                     Hola, buenos días
              </s-gensym8>
       </namesp2:translate_segmentResponse>
</SOAP-ENV:Body>
```

Figure 4. A SOAP answer example. The server returns the translation "Hola, Buenos días"

The use of the SOAP protocol has several advantages. One major advantage over other distributed protocols is that SOAP works well using HTTP as a layer protocol, thus avoiding problems with network firewalls. XML was chosen as the standard message format because of its widespread acceptance. Additionally, a wide variety of freely available tools significantly ease the transition to a SOAP-based implementation. One of its major weaknesses is that, because of the lengthy XML format, SOAP can be considerably slower than other technologies such as CORBA.

4.2. Translation controller

The Translation Controller is the module capable of controlling the translation task, and communicates with the Machine Translation Service and the Translation Memory Service. The client application can make different kinds of requests:

- Machine translation only.

- Translation memory only. In this case the client application can request for just one translation, which is the best translation according the similarity index; or for a set of candidate translations.
- Machine translation and translation memory simultaneously. In this case the client application also has two options: to request for only one translation, which is the best translation according the similarity index; or for a set of translations. The machine translated candidate has a fixed value of similarity that can be specified for each machine translation system and will get the order in the set of candidates according this given similarity.

This module has the following functions:

- info: returns information about the module
- translate-segment: translates one segment, using either the Translation Memory Service or the Machine Translation Service

4.2.1. The translate-segment function

This is the main function of the overall system. It returns a list of translation candidates ordered by similarity. The list contains a given number of candidates; if the number of candidates is set to one, it returns the best translation. This function can use the Machine Translation Service or the Translation Memory Service, or both at the same time. In this latter case, the machine-translated candidate has a fixed similarity, which can be specified individually for each machine translation system. This function requires the following parameters:

- User ID: the system is protected from inappropriate use by means of a list of allowed users.
- Language pair: This parameter is used for the selection of the Machine Translation system. If the parameter is empty, no Machine Translation system will be used.
- Translation Memory list: this list contains a set of paths and names of available translation memories. The translate-segment function can use an unlimited number of translation memories. If this list is empty, no translation memory will be used.
- Minimum similarity: it is the minimum similarity for a segment to be retrieved from the translation memory. By default is set to 90%.
- Number of candidates: the number of candidates to be retrieved. If the number is one only the best candidate will be retrieved.

This function returns an XML message as the one presented in figure 5. For each candidate we can get its similarity index, the translation memory from which it has been retrieved, and both the source and target segments. Please, note that candidate 2 is machine translated as stated in the attribute “from”.

```
<translation-controler-result>
   <option id="1" sim="100" from="memo1engspa">
             <sl>Technical assistance provided by the Department of Economic and
Social Affairs</sl>
             <tl>Asistencia técnica proporcionada por el Departamento de Asuntos
Económicos y Sociales</tl>
   </option>
   <option id="2" sim="90" from="MT">
             <sl>Technical assistance provided by the Department of Economic and
Social Affairs</sl>
             <tl>La asistencia técnica prevista por el Departamento de Asuntos
Económicos y Sociales</tl>
   </option>
   <option id="3" sim="81.8181818181818" from="memo3engspa">
             <sl>2. Programme of technical assistance of the Department of Economic
and Social Affairs</sl>
             <tl>2. Programa de asistencia técnica del Departamento de Asuntos
Económicos y Sociales</tl>
   </option>
   <option id="4" sim="64.2857142857143" from="memo2engspa">
             <sl>Substantive services for the session were provided by the
Department of Economic and Social Affairs.</sl>
             <tl>El Departamento de Asuntos Económicos y Sociales presta servicios
sustantivos en la celebración del período de sesiones.</tl>
   </option>
   <option id="5" sim="54.5454545454545" from="memo1engspa">
             <sl>Population Division, Department of Economic and Social
Affairs;</sl>
             <tl>División de Población del Departamento de Asuntos Económicos y
Sociales;</tl>
   </option>
</translation-controler-result>
```

Figure 5. Example of output of the Translation Controller module.

4.3. Machine Translation Service

This service performs a SOAP request to a Machine Translation system and returns the machine translated segment. This module has two functions:

- info: returns information about the service, available language pairs and machine translation systems.
- translate-segment: translates one segment using a given Machine Translation system

4.3.1. The translate-segment function

This function has the following parameters:

- User ID: to control the access to the system
- Segment: the segment to be translated
- Language pair: this parameter is used to select the correct Machine Translation system

The function returns a machine translated segment.

4.4. Translation Memory Service

The Translation Memory Service returns one or more matches of a given segment with a given similarity or higher, from one or more translation memories. This service has these available functions:

- info: returns information about the module
- retrieve_best: retrieves the best segment if its similarity is higher than the given minimum similarity.
- retrieve_candidates_xml: retrieves an XML containing the given number of candidates with a similarity equal or higher than the given similarity.

4.4.1. Translation Memories

The translation memories are implemented as Berkeley databases. A Berkeley Database is a high-performance, embedded database library that allows manipulating huge databases on a wide variety of systems including most UNIX-like and Windows operating systems. This kind of implementation allows to use very large translation memories with very fast retrieval times.
In our system one translation memory is a set of three Berkeley databases:

- An index database: stores for each word a list of the segments containing that word. Only the words with more than three characters are indexed. Each segment is identified by a unique number.
- A source language database, relating the segment identifier with the source segment.
- A target language database, relating the segment identifier with the target segment.

As a consequence, our translation memories are bilingual and unidirectional.

4.4.2. The *retrieve_best* function

This function retrieves the best segment from a translation memory, which has a similarity index equal or over a given value. The function has the following parameters:

- User ID: to control the access to the service
- Segment: The source language segment
- Translation Memory name and path.
- Minimum similarity

The function returns a string formed by the similarity index, the source segment and the target segment separated by a tabulator

4.4.3. The *retrieve_candidates_xml* function

This function retrieves a XML object containing the given number of candidates with a similarity index equal or higher than a given value. The function has the following parameters:

- User ID: to control the access to the service
- Segment: the source language segment
- Translation Memory name and path.

- Minimum similarity

The function returns a XML object like the one showed in figure 6.

```
<tmresult>
   <option id="1" sim="100">
      <sl>Technical assistance provided by the Department of Economic and Social
Affairs</sl>
      <tl>Asistencia técnica proporcionada por el Departamento de Asuntos
Económicos y Sociales</tl>
   </option>
   <option id="2" sim="81.8181818181818">
      <sl>2. Programme of technical assistance of the Department of Economic and
Social Affairs</sl>
      <tl>2. Programa de asistencia técnica del Departamento de Asuntos
Económicos y Sociales</tl>
   </option>
   <option id="3" sim="64.2857142857143">
      <sl>Substantive services for the session were provided by the Department of
Economic and Social Affairs.</sl>
      <tl>El Departamento de Asuntos Económicos y Sociales presentó servicios
sustantivos en la celebración del período de sesiones.</tl>
   </option>
   <option id="4" sim="54.5454545454545">
      <sl> Population Division, Department of Economic and Social Affairs;</sl>
      <tl>? División de Población del Departamento de Asuntos Económicos y
Sociales;</tl>
   </option>
   <option id="5" sim="43.75">
      <sl>Such consultations should take place in the context of the information
strategy of the Department of Economic and Social Affairs;</sl>
      <tl>Dichas consultas deberían tener lugar en el contexto de la estrategia
de información del Departamento de Asuntos Económicos y Sociales;</tl>
   </option>
</tmresult>
```

Figure 6. Example of output of the retrieve_candidates_xml function

4.5. Translation memory management

A set of functions guarantees the maintainability and scalability of the remote translation memories. Functionalities to export from and import to TMX and tab-separated files are provided:

- import-tab-txt(tm,file,reverse): imports a file (file) containing a translation memory in tab txt format into an existing translation memory (tm). The parameter "reverse" allows importing the second field of the tab txt file as a language 1.
- export-tab-txt(tm,file, reverse): exports an existing translation memory (tm) into a file (file) containing a translation memory in tab txt format. The parameter "reverse" allows importing the second field of the tab txt file as a language 1.
- import-tmx(tm,file,l1-code,l2-code): imports a file (file) containing a translation memory in TMX format into an existing translation memory (tm). The source language TMX code (l1-code) and the target language TMX code (l2-code) must be provided.
- export-tmx(tm,file,l1-code,l2-code,sl): exports an existing translation memory (tm) in a file (file) in TMX format. The source language TMX code (l1-code) and the target language TMX code (l2-code) must be provided.

- tmx-info(file): returns information about the language codes present in a TMX translation memory.

4.6. Evaluation of the translation module

The evaluations performed on the eTITLE Translation module were twofold:

- An evaluation of the combination of machine translation and translation memories compared to using only machine translation.
- An evaluation of the usability of the system, in terms of time saving for a human translating subtitles with no aid, or with the aid of the system.

4.6.1. Evaluation of the integration of MT and TMs, vs. MT alone

Objectives
The goal was to evaluate the impact of integrating a Translation memory with a commercial Machine Translation system, as opposed to using only a commercial Machine Translation system.

Procedure
The testing materials were bilingual pair of texts (Spanish, English and Catalan, respectively) belonging to the same domain as the corpora used to build the Translation Memories. Let us note that these texts had not been used to build the Translation Memories.
We compared the results obtained by the Translation service using the option that allows us to use both TM and MT with the results obtained by each of the MT applications used by the different language pairs. To evaluate the English-Spanish translation we have compared eTitle against WordMagic. To evaluate Catalan-Spanish, we have compared eTitle against Internostrum.
The results obtained by the eTitle Translation service and each of the MT applications were then automatically evaluated using the metric BLEU, one of the most well-known automatic evaluation procedures (Papineni et al.). Additionally we provide the NIST score.

Results
The results of the evaluation of the English to Spanish translation are shown in the table below. The number added to the eTitle columns (i.e. eTitle100, eTitle95, …) indicates the minimum similarity index for the translation memory. If no segment with this similarity or higher is found in the translation memory, the machine translated segment is used. The darker cells indicate the best results.

eng2spa	WordMagic	eTitle100	eTitle95	eTitle90	eTitle85	eTitle80	eTitle75
BLEU	0.2913	0.3134	0.3690	0.3635	0.3544	0.3454	0.3294
NIST	7.5380	7.8174	8.4795	8.4188	8.3098	8.1997	8.0099

Table 1. Results of English to Spanish

All the results obtained with the eTitle system, for all values of similarity, are consistently better than the results obtained by WordMagic used alone. The best results are obtained using a similarity threshold of 95% for both metrics (Bleu and Nist).

The results of the evaluation of the Catalan to Spanish translation are shown in the table below. The darker cells indicate the best results.

cat2spa	WordMagic	eTitle100	eTitle95	eTitle90	eTitle85	eTitle80	eTitle75
BLEU	0.7866	0.7896	0.7791	0.7861	0.7921	0.7943	0.7936
NIST	13.4477	13.4465	13.2238	13.3290	13.4204	13.4744	13.4852

Table 2. Results of Catalan to Spanish

The Catalan to Spanish translation differs from the English to Spanish in that the results obtained by the MT program alone (i.e. Internostrum) are much better, due to the similarity between the two languages involved. For this reason, the improvement due to the integration of TM and MT is less consistent. However, the best results overall are still obtained by eTitle: with a similarity of 80%, according to BLEU and a similarity of 75%, according to NIST.
The results of the evaluation of the English to Czech translation are shown in the table below.

eng2cze	WordMagic	eTitle100	eTitle95	eTitle90	eTitle85	eTitle80	eTitle75
BLEU	0.2100	0.2103	0.2103	0.2103	0.2103	0.2109	0.2108
NIST	4.7474	4.7505	4.7505	4.7505	4.7505	4.7574	4.7506

Table 3. Results of English to Czech

eTitle, with a similarity of 80%, shows the best results, which are only a slight improvement on the benchmark, presumably due to the variability of the text type chosen, namely movie subtitles belonging to all genres.

4.6.2. Usability evaluation

Objectives
In this evaluation exercise we have not measured the accuracy of the translation module, or the "goodness" of the resulting translations. We have simply evaluated the advantages of using the functionality over not using it, in terms of the time that it is able to save to the translator in her task. We have chosen to perform this evaluation in a simple environment (filling a Word table) so as to isolate it from other factors (stability of the system, etc.).
The language pair chosen is English to Czech, and the type of text is movie subtitles. Please, note that this language combination achieved the worse results of BLEU and NIST metrics, seen in the last section. For the other language pairs the time saving is expected to be better.

Procedure
As testing material we have chosen the first half of an American movie: "Analyze that". We have subdivided this material in four parts of similar length and put them in four different files. Parts 2 and 4 have been pre-translated from English to Czech using the eTITLE translation system. Parts 1 and 3 have been left in their original English version.

	Pages	words	characters (no blanks)	subtitles	Translation mode
Part1	6	1423	6408	177	HT
Part2	6	1502	6512	173	MT+PE
Part3	5	1356	6254	182	HT
Part4	5	1408	6182	158	MT+PE

Table 4. HT: Human Translation; MT: Machine Translation; PE: Post-Editing

The Czech translator has been asked to translate the movie, starting with part 1, following with part 2 and then, successively, parts 3 and 4.

	How did he know about the money?	
	And how did he know|that Tony Cisco got popped?	
	We didn't find out about it|till this morning.	
	- I don't know.|- It was Peezee. Gotta be.	

Figure 7. Part 1 (sample)

The text has been organized following the layout shown in figure 7. Column 2 contains the original subtitles, column 3 is used to write the translation, and column 1 is used to keep track of times if the task needs to be interrupted.
The parts that have been pretranslated (Parts 2 and 4) have the 3rd column already filled with the result of the eTITLE translation.

	- Don't say that to me. - Shut up.	- Neříkají, že ke mně. - zmlkl.
	You go to hell! I am so out of here!	Vy dostanete se do pekla! Já jsem tak ven tady!
	Go on, get out of here, you fucking pain in the ass!	Pokračovat, vyjít ven tady, vy soulož bolest v oslu!
	You crazy pain in the ass! Get the fuck out of here!	Vy bláznivá bolest v oslu! Dostat souložit ven tady!

Figure 8. Part 2 (sample)

For parts 1 and 3, the task of the translator is to translate the original text (using standard translation aids, such as dictionaries, etc.), while for parts 2 and 4, her task is to post-edit the automatically translated output: adapt, change, delete, rewrite, etc.

Results of the test
These are the total times for each of the parts:

- Part 1: 3 hours 17 minutes
- Part 2: 2 hours 48 minutes
- Part 3: 3 hours 06 minutes
- Part 4: 2 hours 31 minutes

Total time for the unassisted translation: 6 hours 23 minutes
Total time for the translation assisted by the eTITLE translation system: 5 hours 19 minutes
Total time saved: 1 hour 4 minutes (c. 17 %)

Comments
The translator noted that the text to translate was quite difficult –both for the MT system and for the human- due to abundance of colloquial expressions and language. The MT performed particularly poorly in generating the correct morphological Czech form (tense, etc.), and was better at short sentences than at long sentences. In many occasions the words proposed by the eTITLE translator were correct lexical choices and even if they needed morphological adaptations, they saved time of lexicon look-up.

5. Conclusions

We have presented the Translation module of eTITLE, a web-based multilingual subtitling service.

- The Translation module benefits from the integration of Machine Translation systems and Translation Memories, whose shortcomings and advantages are complementary. As the preliminary evaluation presented in section 4.6.1 shows, their integration allows for better results than Machine Translation alone. Moreover, Translation Memories provide the system with costumization capabilities.
- The module has been implemented using a distributed environment, which allows for a high degree of flexibility. The different Machine Translation systems, the Translation memories, and the Translation controller itself may physically be in different computers and can also be replicated, adding to the overall robustness of the system. New MT systems may be easily added and free online translation services may also be accessed.
- The technology employed for the Translation Memory service (Berkeley Databases) allows for the use of very big translation memories, while guaranteeing a very fast access and retrieval, as well as portability.

6. Bibliography

Aizawa, T., Ehara, Uratani and Tanaka (1990). A Machine Translation System for Foreign News in Satellite Broadcasting. In Proc of Coling-90, Vol. 3, pp. 308-310.

Nadjet Bouayad-Agha, Angel Gil, Oriol Valentin and Victor Pascual (2006). A sentence compression module for machine-assisted subtitling. Proceedings of the Seventh International Conference on Intelligent Text Processing and Computational Linguistics (CICLING'06).

Canals R., Esteve, A., Garrido, A., Guardiola, M.I., Iturraspe-Bellver, A., Montserrat, S., Pérez-Antón, P., Ortiz, S., Pastor, H., y Forcada, M. (2001). InterNOSTRUM: a Spanish-Catalan Machine Translation System. Machine Translation Review, 11:21-25.

Cancelo, Pablo (2000), «[Reseña a] Herramientas Mágicas / Word Magic Tools de Word Magic Software (Word Magic Tools 2000 Deluxe 2.1.)», Revista de Lexicografía, VI (1999-2000), 235-238

Czuba, K. (2000). Efficient parsing strategies for syntactic analysis of closed captions. NAACL'00 Student Workshop, Seattle.

Moore, Robert C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In Machine Translation: From Research to Real Users (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, pp. 135-244.

Nyberg, E. and Mitamura, T. (1997). A real-time MT system for translating broadcast captions. In Proceedings of MT Summit VI, Machine Translation Past Present Future, San Diego, CA, pp. 51–57.

Papineni, K.A., Roukos, S., Ward, T., Zhu, W.J. (2001) Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center

Piperidis S., Demiros I., Prokopidis P., Vanroose P., Hoethker A., Daelemans W., Sklavounou E., Konstantinou M. and Karavidas Y. (2004). Multimodal Multilingual Resources in the Subtitling Process, in the Proceedings of the 4th International Language Resources and Evaluation Conference (LREC 2004), Lisbon.

Popowich, F., McFetridge, P., Turcato, D. and Toole, J. (2000). Machine Translation of Closed Captions, in Machine Translation, vol 15, (311-341), Kluwer Academic Publishers, Netherlands.

Siohan et al.(2001), A Real-Time Japanese Broadcast News Closed-Captioning System, Submitted to Eurospeech'01.

Turcato, D., Popowich, F., McFetridge, P., Nicholson, D. and Toole, J. (2000) Preprocessing Closed Captions for Machine Translation. In C. Van Ess-Dykema, C. Voss, F. Reeder (eds.), Proceedings of the Workshop on Embedded Machine Translation Systems , Seattle, May 2000.