Hybrid and Interactive Domain-Specific Translation for Multilingual Access to Digital Libraries

Gareth J. F. Jones, Marguerite Fuller, Eamonn Newman, and Ying Zhang

Centre for Digital Video Processing
School of Computing
Dublin City University, Dublin 9, Ireland
gjones@computing.dcu.ie

Abstract. Accurate high-coverage translation is a vital component of reliable cross language information retrieval (CLIR) systems. This is particularly true for retrieval from archives such as Digital Libraries which are often specific to certain domains. While general machine translation (MT) has been shown to be effective for CLIR tasks in laboratory information retrieval evaluation tasks, it is generally not well suited to specialized situations where domain-specific translations are required. We demonstrate that effective query translation in the domain of cultural heritage (CH) can be achieved using a hybrid translation method which augments a standard MT system with domain-specific phrase dictionaries automatically mined from *Wikipedia* . We further describe the use of these components in a domain-specific interactive query translation service. The interactive system selects the hybrid translation by default, with other possible translations being offered to the user interactively to enable them to select alternative or additional translation(s). The objective of this interactive service is to provide user control of translation while maximising translation accuracy and minimizing the translation effort of the user. Experiments using our hybrid translation system with sample query logs from users of CH websites demonstrate a large improvement in the accuracy of domain-specific phrase detection and translation.

1 Introduction

The growth in Digital Libraries (DLs) is offering access to increasing numbers of document collections from around the world. The full potential of these resources for applications such as research, study and cultural exchange can only be realised when users have efficient and reliable access to them. Such access poses many challenges for the designers of technologies for DLs. One of these challenges is the development of effective methods to support multilingual access to DLs where the contents may be in multiple languages, one or more of which may be unknown or known only partially to the user of the DLs. In such situations the user must rely on automatic translation technologies to support search

of the content and interaction with retrieved items. In working with these systems user must pose their search queries in a language known to them and rely on automatic translation to render their search request into the document language or languages, and, depending on their reading skills in the target language, possibly rely on automatic translation of retrieved documents. The effectiveness with which their search is conducted depends to a large extent on the quality of the translation of the domain-specific concepts.

Reliable translation is thus a key component of effective cross language information retrieval (CLIR) and multilingual information retrieval (MLIR) systems. Various approaches to translation have been explored at evaluation workshops such as TREC[1], CLEF[2] and NTCIR[3]. While extensive sets of experiments have been reported at these workshops, they have been based on laboratory information retrieval (IR) test collections consisting of news articles or technical reports with "TREC" style search queries[4] with a minimum length of a full sentence. With document sets such as these, general purpose translation resources based on bilingual dictionaries or standard machine translation (MT) have been shown to be effective for translation in CLIR.

This approach to translation using general resources will however frequently not be sufficient for multilingual DLs which often contain domain-specific terms or phrases related to the specific content that the user is seeking to locate within the library. In these cases content, and in particular the sections of the content related to the specific domain of interest, must be translated accurately if effective access to relevant information is to be achieved. One DL domain of which this is true is cultural heritage (CH). The CH domain is of interest to many organisations, including museums and national institutions engaged in the preservation of cultural content. Developing tools to make this content available to larger numbers of potential users than is the case at present is of interest to all such organisations. This desire is driven partially by a desire to increase societal awareness of their cultural assets, but also more pragmatically to justify the resources currently being invested in the development of DLs and their cultural holdings. Search tools for CH content may take the form of standard search engines producing ranked lists for users, but may also look towards more sophisticated applications incorporating personalisation of content selection and delivery of dynamically composed personal responses.

A number of projects in recent years have explored technologies to advance multilingual access to DLs. Among these projects was the EU FP6 *MultiMatch*[5] project which was concerned with information access for multimedia and multilingual content for a range of European languages in the domain of CH. In this paper we briefly review the principle approaches taken to translation in CLIR and MLIR systems, namely dictionary-based methods and machine translation

[1] `http://trec.nist.gov`
[2] `http://www.clef-campaign.org/`
[3] `http://research.nii.ac.jp/ntcir/`
[4] Referred to at TREC as search *topics*.
[5] `http://www.multimatch.org`

(MT). We then use this to motivate our proposal of a hybrid translation service for CLIR and MLIR developed within the MultiMatch project to facilitate effective domain-specific translation in the CH domain. This combines a commercial MT service with a domain-specific dictionary gathered automatically from the multilingual *Wikipedia*. The basic form of this service operates automatically in the form of an augmented MT service which outputs its best available translation of the text input. We demonstrate the effectiveness of this service using sample CH request logs in English, Spanish and Italian provided to us by organisations providing access to in DLs in the area of CH. We translate the requests and examine the quality of the translated output using human assessors. This study demonstrates how using a domain-specific phrase dictionary to augment a general MT system can improve the coverage and reliability of translation of these requests within this domain.

The automatic hybrid service is then extended to provide an interactive translation service enabling users with some knowledge of the target translation language to check the elements of the hybrid translated output and to correct or augment those which they judge to be inaccurate or limited using alternative possible translations taken from the bilingual dictionary.

The remainder of this paper is organized as follows: Section 2 overviews the topic of translation in CLIR and MLIR, Section 3 introduces our hybrid approach to translation and the translation resources used in this study, Section 4 describes our experimental investigation of the effectiveness of the hybrid translation service, Section 5 then describes the extension of the hybrid service to enable interactive user adjustment of the translated output, and finally Section 6 summarizes our conclusions and considers directions for further work.

2 Translation Approaches in CLIR and MLIR

The majority of early work in CLIR concentrated on the translation of search queries using bilingual dictionaries. These were typically the largest general purpose electronic dictionaries available to the investigators. Simple request translation using these dictionaries replaced each word in the source language with all possible alternatives in the target language. The significant ambiguity introduced into the request by this approach was quickly shown to have a significant adverse impact on retrieval effectiveness [9]. Much research in CLIR then focused on methods to remove or reduce the impact of this ambiguity in translation of search queries. One of the most important factors introduced which improved CLIR effectiveness was translation of phrases rather than their individual words [3]. This is particularly important for idiomatic phrases, but also reduces ambiguity in the case of compositional phrases.

A logical alternative translation method in CLIR is the use of MT. It was often argued that search requests lack sufficient grammatical structure to be reliably translated by MT systems, which are traditionally designed for the translation of linguistically well formed text. However, experiments applying MT to CLIR tasks rapidly showed that while the lack of structure in the requests can

result in translation errors, overall CLIR effectiveness is often as good as or better than that achieved by using the most complex dictionary-based methods [10]. Until recently MT systems were only available for a very limited number of language pairs due to the very high cost of development. However, MT systems for many more language pairs are now appearing, greatly increasing its appeal for CLIR. For the translation of documents either for use in the retrieval process (by translating documents instead of the queries [11]), or for reading by users after retrieval with query translation, MT is the only realistic option.

While MT systems can provide sufficient translations for general language expressions, they are often not sufficient for domain-specific phrases that contain personal names, place names, technical terms, titles of artworks, etc. In addition, certain words and phrases hold special meanings in specific domains. For example, the Spanish phrase "Canto general" was translated by a standard MT system used in our work into English as "general song", which is arguably correct. However, in the CH domain, "Canto general" refers to a book title from Pablo Neruda's book of poems and should be translated directly into English as the phrase "Canto general". Multiple word phrases are more information-bearing and more unambiguously represented than single words; they are also often domain-specific and typically absent from static general lexicons. Effective translation of such phrases is particularly critical for the short search queries that are typically entered by non-expert users of search engines. It should be clear that failure to translate these important expressions correctly will often have a disastrous impact on search effectiveness.

An advantage of dictionary-based translation methods for search queries is that bilingual dictionaries can be constructed for new language pairs or domains at comparatively very low cost, and easily be adjusted to add new translation entries, and, of particular importance for CLIR, new phrase translation pairs.

Overall then it would be desirable to have a translation service for CLIR which was well specified for the domain of interest, e.g. CH, and could be easily further adapted as new vocabulary is encountered, but also did not introduce the ambiguity associated with dictionary-based translation. The next section proposes a hybrid translation method that combines these features.

3 A Hybrid Approach to Translation in Information Retrieval

Our novel hybrid translation service aims to improve translation effectiveness in the CH domain by improving the translation of phrases previously untranslated or inappropriately translated by a standard MT system. In this work we combine a standard non-domain specific MT system with domain-specific phrase dictionaries mined from Wikipedia combined with a small standard bilingual dictionaries. Our hybrid service aims to simultaneously address problems of words or phrases which are outside the domain of the MT system, prevent the problems of introducing translation ambiguity associated with dictionary-based translation models, and to improve the reliability of CH phrase translation. Figure 1

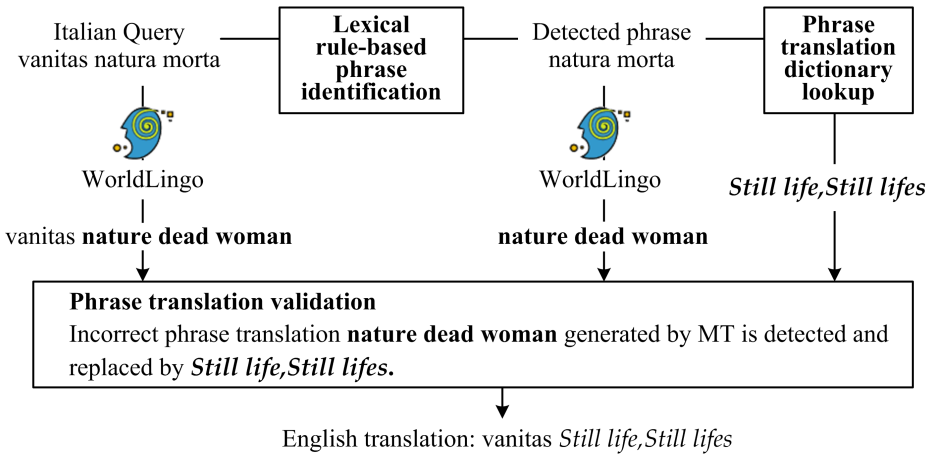

Fig. 1. An example of Italian–English hybrid translation of a search query.

illustrates the stages of our hybrid query translation process for the translation of an Italian search query into English. The basic idea is that rather than passing the text for translation directly to the MT system, we first analyse it to locate phrases in a bilingual dictionary, and handle these specially so that the statistically most likely phrases can be included in a hybrid translation output by combining them with the output of the MT system.

Three methods of multiple-word phrase identification have been commonly used in text analysis: lexical rule-based [3][9], statistical [5], and syntactical methods [5][15]. The lexical rule-based approach with maximum forward matching is adopted in our hybrid translation process due to its robust performance and computational simplicity. The input text is sequentially scanned to seek matches in the phrase dictionary. Where more than one phrase translation is available in the bilingual dictionary, the most frequent translation in the training corpus is selected for inclusion in the final translation. The longest matched sub-sequence is taken as a phrase and translated via a domain-specific dictionary lookup. This process is recursively invoked on the remaining part of the text until no further matches are found. The effectiveness of this approach depends strongly on the completeness of the coverage of the adopted dictionary.

The text for translation is then processed to replace the identified phrases with their corresponding translation from the dictionary-based translation service. The translated phrase is then annotated in the text to be translated to prevent any mistranslation that might occur during translation by the MT system. The demarcation marks indicate to the MT system that content between these marks should not be translated by the MT system. The augmented text is passed to the MT system and its response is processed to remove markup before the combined hybrid output is presented as the translation. One important practical feature for an MT system selected to be used in this service is that it must support text markup to leave marked items untranslated in the output.

After informal analysis of a number of possible online MT systems we selected the WorldLingo[6] system for our work since it provided good support for content markup and translation for a good number of language pairs.

The next section describes the construction of our CH domain-specific bilingual dictionaries.

3.1 CH Domain-Specific Dictionary Construction

Our CH domain-specific dictionaries combine bilingual wordlists downloaded from the XDXF Dictionaries[7] webpage combined with.domain-specific bilingual wordlists built by mining interdocument links from Wikipedia[8] for documents on the same topic. The downloaded XDXF dictionaries contained between 950,000 and 250,000 word pairs depending on the language pair and direction of translation being considered.

In recent years Wikipedia has emerged as a major online source of information. While the largest proportion of content is in English, varying amounts of content are available in other languages. As might be expected since the content is community contributed, the amount of context is somewhat correlated with the number of speakers of the language, but is continuing to grow for all languages. There are many instances of pages on the same topic in different languages within Wikipedia. Although not directly relevant here, it should be noted that while these pages are on the same topic in different languages, they are not generally parallel texts or even close translations of each other, but rather individual pages on the same topic authored separately by speakers of the relevant languages. This means that they generally reflect the cultural perspective and vocabulary use associated with the speakers of the language in question. As a multilingual hypertext medium, Wikipedia has been shown to be a valuable source of translation information [1, 2, 4, 6]. Wikipedia is structured as an interconnected network of articles, in particular, Wikipedia page titles in one language are often linked to a multilingual database of corresponding articles in other languages. Unlike the web, most hyperlinks in Wikipedia have a more consistent pattern and meaningful interpretation. For example, the English Wikipedia page `http://en.wikipedia.org/wiki/Cupid_and_Psyche` hyperlinks to its counterpart written in Italian `http://it.wikipedia.org/wiki/Amore_e_Psiche`, where the basenames of these two URLs ("Cupid and Psyche" and "Amore e Psiche") are an English–Italian translation pair. Thus, the URL basename can be considered to be a term (single word or multiple-word phrase) that should be translated as a unit.

Utilizing the multilingual linkage feature of Wikipedia, we used a three-stage automatic process to mine Wikipedia pages as a translation source and construct phrase dictionaries in the culture heritage domain:

[6] `http://worldlingo.com`
[7] http://xdxf.revdanica.com/down/
[8] `http://wikipedia.org`

1. We performed a web crawl from the English Wikipedia, Category: Culture. This category contains links to articles and subcategories concerning arts, religions, traditions, entertainment, philosophy, etc. The crawl process in the category of culture included all of its recursive subcategories. In total, we collected 458,929 English pages.
2. For each English page the hyperlinks to each of the translation languages to be used were extracted. For the study reported here, the languages mined for links were Italian and Spanish.
3. The basenames of each pair of hyperlinks (English–Italian, English–Spanish) were selected as translations and then added into our domain-specific dictionaries. Multiple-word phrases were added into the phrase dictionary for each language.

Our Wikipedia mined dictionaries contained about 90,000, 70,000, and 80,000 distinct multiple-word phrases in English, Italian, and Spanish respectively. The majority of the phrases extracted were CH domain-specific named entities and the rest of them general noun-based phrases, such as "Music of Ireland" and "Philosophy of history". We did not apply any classifier to filter out the general noun-based phrases since such phrases can be useful additions for accurate query translation.

Where multiple translations of a phrase were located in the Wikipedia archive, the alternative translations were ranked in the bilingual dictionaries by frequency of occurrence in the Wikipedia pages. This ranking enables us to select a single most likely translation for use in the single best output of the hybrid translation system. Combining the Wikipedia mined dictionaries with the general purpose ones gathered from XDXF Dictionaries gave CH-biased dictionaries with good coverage of general and domain-specific concepts.

4 Experimental Investigation

In order to investigate the effectiveness of our hybrid translation service for CH search request translation, we performed an experimental investigation to compare request translation accuracy of our domain-specific hybrid approach with the output of WorldLingo standard MT. The goal here was to measure the degree to which output translations were judged suitable as translated search queries by human assessors. Thus rather than using a standard IR test collection, we based our experiments on real user query log data.

4.1 Query Log Test Sets

The query logs used in our experiments were all provided by real users sending CH related queries to websites provided by or associated with CH organisations. One of the sets consists of queries in Spanish, the second is in Italian and the third is in English. The Spanish queries came from a DL based in Spain whose focus is on poetry and ancient and modern literature in the Spanish language. The

Table 1. Query translation examples.

Original	WorldLingo MT	Hybrid Translation
Plinio il giovane	Plinio the young person	Pliny the Younger
Pittura a tempura	Painting to moderates	Egg tempera
Literatura infantil y juvenil	Infantile and youthful Literature	Children's literature
Al andalus	To andalus	Islamic Spain
Still life paintings	Pinturasde la vida inmovil	Bodegon pinturas

Italian queries are taken from the "Cultural" section of a large Italian Internet Service Provider's website. The queries in English were extracted from the query logs of the website for a well-known art gallery based in London, U.K. There were 1423 Italian queries (with an average length of 2.49 words), 1088 Spanish queries (3.39 words on average) and 100 English queries (1.67 words on average).

Each query was translated separately using the standard WorldLingo MT system and the hybrid system. We translated the Spanish and Italian queries to English (and the English to Spanish and Italian) since we had bilingual evaluators available for these language pairs. When both systems produced the same translation for a given text, the results were discarded since for this evaluation we were interested in the disagreements between the systems. The sets of translations are denoted *Es-En*, *It-En*, *En-Es* and *En-It*. The translations were collated so that the evaluators could make a side-by-side comparison between the original text, the MT output and hybrid translation. Some examples are given in Table 1. A single bilingual evaluator judged the suitability of each translated query set. The details of instructions given to each evaluator for the experiment are described in the following section. It should be noted that it was not possible to directly compare the lexical coverage of our domain-specific dictionaries and the built-in phrase dictionaries of WorldLingo since we did not have access to the internal WorldLingo dictionaries.

4.2 Human Evaluation of Translation Quality

For each query where the WordLingo MT and hybrid outputs differed, the bilingual evaluators were asked to mark which of the two translation results they "considered to be better", that is more accurate to a native speaker. As there was only one evaluator per set, we were not able to consider inter-annotator agreement on this subjective measure. Any possible bias due to a single evaluator will result in a skew of the results for one set, rather than the whole evaluation. Table 2 summarises the results of the experiments. There were 2711 queries to be translated in total. The same translated output was produced for 1919 queries leaving 792 to be examined by the assessors.

The results in Table 2 show that the hybrid translation system was generally regarded as providing a better translation than the WorldLingo MT system. For Spanish-English, the hybrid translation was correct in 79% of the cases

Table 2. Results of analysis of alternative translations.

Language Pair	Number of Translations	Number of Disagreements	Hybrid Correct	Both Correct	WorldLingo MT Correct	No Preference
It - En	1423	482	288	63	75	56
Es - En	1088	281	222	0	58	1
En - It	100	15	9	1	2	3
En - Es	100	14	11	0	3	0

Table 3. Results of analysis of hybrid translations including all dictionary entries.

Language Pair	Number of Translations	Number of Disagreements	Hybrid Correct	Both Correct	WorldLingo MT Correct	No Preference
It - En	1423	482	353 (+65)	71 (+8)	2 (-73)	56
Es - En	1088	281	273 (+51)	0	7 (-51)	1
En - It	100	15	10 (+1)	2 (+1)	0 (-2)	3
En - Es	100	14	12 (+1)	2 (+2)	0 (-3)	0

where there was a disagreement between the systems. "No preference" results indicate that the evaluator felt that neither translation was appropriate. For Italian to English, when we remove "no preference" results and those where both systems were deemed correct (leaving 482-(56+63) = 363 instances), we achieve a very similar score of 79.3% correctly translated by the hybrid system. Situations where both are deemed "correct" raise the interesting issue for CLIR of which one should be preferred in order to be most likely to retrieve relevant documents. The small number of English queries means that we cannot attach significance to the results, however for the sake of completeness, we can report correct translation rates of 81.8% for English to Italian and 78.5% for English to Spanish, which are similar to the results from the larger sets. The similarity of these results, across different language pairs, different evaluators and different set sizes suggests that there was no significant bias inherent in any of the evaluations.

These results show that our method of enhancing MT by incorporating domain-specific dictionaries is successful for query translation. By identifying phrases and named entities with specific interpretations in the CH domain, we are able to improve on standard MT output in around 80% of cases.

Having native speakers as evaluators allows further analysis of the actual quality of the translations, rather than just comparing them to the baseline. In order to make a more detailed comparison, the evaluators were also asked to highlight any translations which they thought were "particularly good" or "particularly bad". For example, the evaluator for translations between Spanish and English thought a translation of "poema del mio cid" was particularly good as it inserted the full name of the work ("Cantar de Mio Cid") into the translation (giving "poem of Cantar de Mio Cid") making it much better than the literal translation provided by the MT system ("poem of mine cid").

In CLIR, unlike conventional MT tasks, there is no need to produce a single best translation, and indeed including multiple possible translations has the potential to retrieve a set of relevant documents where information is described in alternative equally correct ways in different documents. This alternative descriptions of the relevant information may match well with different versions of a query. In order to assess the potential of the hybrid system to be used in CLIR, including all the possible translations available in the domain-specific dictionaries, the results were re-examined showing all the alternative translations available in the hybrid dictionary to the evaluators. In many cases, one of the alternative hybrid translations matched the MT system translation exactly, or matched it when stopwords were removed. Table 3 shows the updated results of adding these alternative translations. The new results show that including the alternative translations produces a large increase in the number of translations produced by the hybrid system deemed correct. In this case where the hybrid system was preferred, the evaluator felt that the expanded output of the hybrid system was better for CLIR than the MT system on its own in almost all cases. The few cases where both results were judged to be correct arose in situations where the output from the two systems was so similar as to effectively be functionally identical.

Analysis of the output of the hybrid translation system showed that at least one phrase is detected in 90% of the evaluation queries. These included, personal names, geographic locations, and titles of various types of artworks. This indicates that our phrase dictionaries have good coverage of phrases to be translated.

While we were not able to manually evaluate the accuracy of all translation pairs in our bilingual dictionaries due to limited resources, our experiments using the hybrid translation tool for sample queries in the CH domain demonstrate that our translations are generally regarded as very accurate by bilingual assessors.

4.3 Related Experiments

The practical objective of our hybrid translation system is to improve CLIR effectiveness in a specific domain of interest. Since we did not have access to a suitable IR test collection consisting of set of documents with corresponding relevance data for the user search topics provided by the CH organisations, we conducted a set of CLIR experiments using a different domain-specific IR test collection. We used the CLEF 2007 Cross Language Speech Retrieval (CL-SR) English language task. This task consists of a small collection of about 8000 spoken "documents" and 42 search queries with corresponding relevance data indicating which of the documents are relevant to each query. The documents were formed from English language interviews with survivors of the Holocaust which were divided into topically coherent segments by subject matter experts. The audio segments were automatically transcribed using automatic speech recognition. The speech recognition was adapted to the domain of the audio recordings, and produced transcripts with an error rate on the order of 20%. This error rate may appear high, but is generally found to be sufficiently accurate to support effective retrieval of the content based on the transcriptions [7]. This test

collection provided an interesting test for search technologies within the Multi-Match project since it is a (non-CH) domain-specific cross language multimedia retrieval task. One limitation of this dataset is that the query statements are generally rather longer than those typically entered into a web search engine. They are typically a full sentence of text, rather than the two or three words often entered into a search engine. However, this task is sufficient to explore the efficacy of our hybrid translation method.

For the CLEF task we trained new bilingual dictionaries in the domain of the CL-SR data set (issues relating to World War Two). These were then used in combination with the WorldLingo MT system to perform a set of comparative experiments exploring alternative translation strategies for search queries originating in French, German and Spanish. The full results of these experiments are reported in [16]. Results from these experiments showed that combining our domain-specific dictionaries with MT methods improves the CLIR effectiveness in terms of Mean Average Precision (MAP) and Precision at rank 10 (P@10) for the CL-SR task. While best retrieval accuracy was achieved using a monolingual evaluation task where the queries were English, our results for the cross language task were the best among those making formal submissions to the CLEF 2007 CL-SR task, showing the lowest decrease relative to monolingual performance when queries were translated from their source language to English [12]. These results are encouraging for us since they demonstrate that our approach can work well for ad hoc retrieval and when working with errorful transcribed output from speech recognition systems, as is often encountered when working with multimedia DL archives.

5 Interactive Hybrid Translation Service

The hybrid translation service described so far provides a single best or most likely translation of the input text. The experimental analysis in the previous section shows that when there is a difference between them, this "best" translation often improves on the standard WorldLingo MT output, and additionally that including alternative translations available in the dictionary improves the coverage of correct translations in the output. Users of CLIR systems typically have differing skills in the languages concerned. Thus users with some knowledge of the language into which the text is being translated will be able to identify some of the mistakes in the hybrid output, i.e. users with some level of reading or at least word recognition skill, but not sufficient productive skills to write the query in the target language. In order to take advantage of these users's language abilities and where possible to eliminate or at least reduce translation errors, an interactive version of the hybrid translation service was developed.

The intention of this system is to provide a translation service which provides an effective integration of the strengths of the separate MT and dictionary-based translation services, and exploits any linguistic knowledge of the users. The MT service provides a single automatic output, similar in form and functionality to the new hybrid service. In this approach the user only has to enter the text

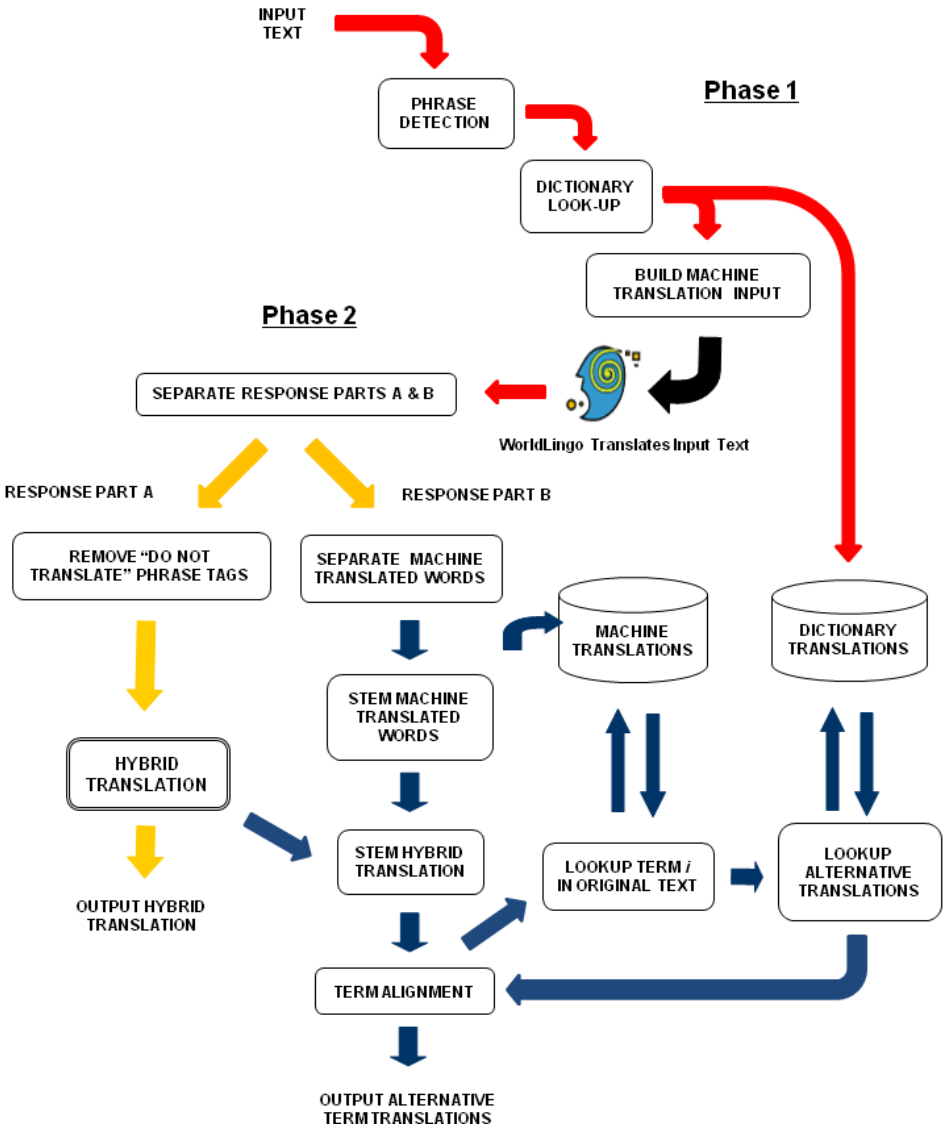

Fig. 2. Interactive Hybrid Translation Process.

which is translated automatically. This is thus low cost to the user and fast, however using this strategy the user has no control over the output and is thus entirely dependent on the suitability of the design and parameter settings of the translation system. A simple dictionary-based approach uses all the available translations of each word, however, as outlined earlier, this has been shown to be ineffective in many experimental studies since it introduces ambiguities, effectively translation errors, some of which can have substantial impact on retrieval behaviour. However, dictionary-based translation potentially offers the user the possibility to select from all the available translations. These translations can be presented to the user in many different ways, for example presenting possible translations in order of frequency or alphabetically, or recommending the first sense in a dictionary as the default translation, with other possibilities shown to the user for selection. Whatever translation ranking design choice is made here, the key point is that the process is interactive with the user having complete free choice of which translations should be used in the CLIR process.

User studies generally show that users who are suitably linguistically qualified like to have control of the translation process in CLIR [8][13]. However, this requires work from the user to perform the translation selections in the CLIR process, and it is generally understood that users do not like to expend more effort than necessary in undertaking the search process. The objective of the interactive hybrid translation service is to increase user control of the translation while maximising translation accuracy and minimizing the average amount of work to be carried by the user to achieve this. To achieve this, the hybrid translation service described in the previous section was extended to facilitate a user's possible desire to find alternative translations for words and phrases within their query to those proposed by automatic translation. The nature of the dictionary translation system lent itself to this extension since many of the terms translated by the MT system also appear in the dictionary with alternative translations. The aim of the hybrid translation system is to provide the single "best" translation to the user as the default translation. If the user is happy with this, they can then proceed directly to the CLIR phase. If, however, they are not satisfied with the accuracy or coverage of the translation, the interactive hybrid system enables them to access the alternative translations available in the dictionary and to select items from among those available to replace or augment elements of the single "best" default translation.

While perhaps appearing a very simple process, incorporating the interactive aspect to the hybrid translation service is actually quite complex due to the use of multiword phrases and the characteristics of MT. For example, if a word sequence is sent to WordLingo for translation, it is often highly problematic to match each word in the original text with its translation as is required for simple dictionary-based translation. There are a number of reasons for this, the word order many differ, a single term in one language may translate into multiple words in another language, multiple words in one language may form compound words in another language, or additional words may be added to the translation which have no equivalent translation in the source language. A simple approach

to overcoming this problem would have been to send each word separately to the MT system for translation. However this would have defeated the purpose of using an MT system since it would have performed simple single word translation of isolated words in the same manner a dictionary-based translation service and any context data contained in the text, important for exploitation of the full linguistic resources of the MT system, would have been ignored. Our solution to this problem is to augment the text sent to the MT system. The augmented text contains the original text fused with a tagged version of the text. The tagged version of the text contains each query word separated by demarcated tags. The MT system translates the text as a whole entity and each word as an entity. This allows a mapping of translated words to original words, this mapping enables the combination of a translation component containing the full hybrid translation along with possible alternative translations for each word in the translation. The complete process to produce the output for the interactive translation system is illustrated in Figure 2.

5.1 Interactive Hybrid Translation Process

This section describes the stages of the complete process for generating the output components of the interactive hybrid translation service. This description assumes use of our CH phrase translation dictionary with the WorldLingo MT system, but the model could in principle be applied with another domain-specific dictionary for an alternative domain or another MT system with similar features.

The process comprises 6 stages at the end of which the output includes the automated single best hybrid translator and the available alternative translations from the CH bilingual dictionary.

Step 0: Request Pre-Processing

Remove excess white space, convert request to lower case.

Step 1: Cultural Heritage Phrase Detection

Detection of words and phrases contained in the input text found in the word and phrase list in the bilingual dictionaries. Greedy-parsing algorithms are used to identify the longest sequences of dictionary words in the input.

Step 2: Dictionary Look-Up

Dictionary look-up is performed on each word in the input text. If the word is present in the dictionary, the word and its corresponding translations are placed in a table. This dictionary table is referenced later during the alignment process ub Step 5(iv). Where a phrase translation is identified in the dictionary, the translation of the phrase replaces the original phrase in the text to be translated.

For example, the text `Mona Lisa Louvre` becomes `<-- La Gionda --> Louvre`, since the phrase `Mona Lisa` appears in our CH dictionary with the translation `La Gionda`.

Step 3: Build Machine Translation Request

The text is formatted for input to the WorldLingo MT system. The formatted text consists of two components:

- First component: the full text with identified CH phrases marked as "do not translate" (the input to the automatic hybrid translation service);
- Second component: two copies are made of each individual word in the text input one of them marked with "do not translate" tags. The purpose of this is to identify the translation of each word generated by the MT service.

Step 4: Formatted text is sent to the WorldLingo MT system.

Step 5: The response from WorldLingo MT system is processed to align the hybrid and alternative translations.

Step 5(i): The two components of the MT response are separated:

- First component: the automatic hybrid translation of the text input (output of the automatic hybrid translation service);
- Second component: individual words and their translations.

Step 5(ii): The tags are removed from the hybrid translation and the individual words and their translations.

Step 5(iii): The words in the hybrid translation and individual translated words are stemmed.

The application of stemming is required on the translated output since word forms in the hybrid translation may be different to those appearing in the translated individual words.

Stemming algorithms are a standard approach in IR which enable alternative word forms, e.g. single and plural, to be matched. Our hybrid system uses the popular rule-based Porter stemming algorithm [14]. The Porter algorithm was originally developed for English, alternative versions for a large number of other languages are now available from Snowball [9].

Step 5(iv): Term Alignment

- Look up each stemmed term in the hybrid translation from the first component in the stemmed individual terms in the second component.
- Look up the corresponding word in the source language.
- Look up the alternative translations of the source word in the dictionary table formed in Step 2.

[9] `http://snowball.tartarus.org/`

Example of Generation of Interactive Translation Output

Query: `Storia del teatro Greco`
Source Language: Italian
Target Language: English

Step 0: Request Pre-Processing

Remove excess white space, convert request to lower case.

```
storia del teatro greco
```

Step 1: Cultural Heritage Phrase Detection

The request is converted to a list of terms and phrases.

```
storia        - single word found in domain-specific dictionary
del           - single word not in domain-specific dictionary
teatro Greco  - phrase found in domain-specific phrase dictionary
```

`del` is a common Italian function word and not found in the CH domain-specific dictionary.

Step 2: Dictionary Look-Up

Form dictionary table of translations found in domain-specific dictionary.

```
storia       - Historie; Historic; History;
teatro greco - Theatre of ancient Greece; Ancient Greek
               theatre; Greek theater; Greek theatre;
```

Where a term is located as a phrase in the domain specific dictionary, it is replaced with its most frequent translation phrase. Other words are left untranslated and untagged in the input to the WorldLingo system.

```
storia
del
Theatre of ancient Greece
```

Step 3: Build Machine Translation Request

The MT request consists of two parts:

Part 1: Complete query for translation, enables use of all available context information in the query using the hybrid translation service.

```
storia del <-- Theatre of ancient Greece -->
```

Part 2: Individual words copied twice. One copy marked "do not translate".

```
<--[storia]--> storia <--[del]--> del <--[teatro greco]-->
<--Theatre of ancient Greece --> <--teatro greco-->
<--[ -->teatro greco <-- ]-->
```

`<-- xxx -->` and `<--[xxx]-->` are WorldLingo markup syntax for pass through unchanged and ignore item.

The two parts are fused together to form a request to be passed to WorldLingo.

Step 4: Send Machine Translation Request to the WorldLingo MT System

Step 5: Process response from WorldLingo MT system to align the hybrid and alternative translations

Step 5(i): Separate Response from WorldLingo

Part 1: history of `<--Theatre of ancient Greece-->`

Part 2: `<--[storia]--> history <--[del]--> of <--[teatro greco]-->`
`<--Theatre of ancient Greece--> <--teatro greco-->`
`<--[ --> Greek theatre <-- ]-->`

Step 5(iia): Remove tags from Part 1

```
history of Theatre of ancient Greece
```

Step 5(iib): Extract words in Part 2

Separate words and phrases into original words and their translations.

```
history - storia
of - del
Theatre of ancient Greece - teatro greco
Greek theatre - teatro greco
```

Step 5(iiia): Stem Hybrid Translation

```
Histori of Theatr of anci Greec
```

Step 5(iiib): Stem Machine Translated Words

```
Histori - Storia
Theatr of anci Greec - teatro greco
of - del
teatro greco - Greek theatre
```

Step 5(iv): Term Alignment

- Split the hybrid translation into its constituent stemmed terms.
- For each term i in the hybrid translation.
- Look up original text of i in the machine translation table.

```
Histori → Storia
of → del
Theatr of anci Greec → teatro greco
Greek theatre → teatro greco
```

Look up alternative translations in the dictionary table.

```
Storia → Historie
```

Storia → Historic
Storia → History

Del → null **Not in dictionary table

teatro Greco → Theatre of ancient Greece
teatro Greco → Ancient Greek theatre
teatro Greco → Greek theater
teatro Greco → Greek theatre

Note: A look-up is also performed on the MT table for the machine translated output of the dictionary translated phrases. This allows for the inclusion of cases where the MT output is different and potentially more appropriate than those contained in the hybrid components of the complete interactive translation.

The automated primary hybrid output shows the selected "best" translation at each point. The alternative translations at each point proposed by the MT system and CH dictionary are also made available to the user. The best translation is shown to user as the selected translation. The user is then free to make use of alternative translations as displayed to them in the user interface.

Source Lanaguage: ITALIAN
Target Language: ENGLISH

Position:	0	1	2
Original Query:	storia	del	teatro greco
Best Hybrid Translation:	history	of	Theatre of ancient Greece

Elements available for use in the interactive translation interface.

position: 0

originalTerm: storia
Type: STANDARD MT
Translation: history
Type: DICT
Translation: Historie
Type: DICT
Translation: Historic
Type: DICT
Translation: History

position: 1

originalTerm: del
Type: STANDARD MT
Translation: of

position: 2

originalTerm: teatro greco
Type: HYBRID MT
Translation: Theatre of ancient Greece

Type: STANDARD MT
Translation: `Greek theatre`
Type: DICT
Translation: `Theatre of ancient Greece`
Type: DICT
Translation: `Ancient Greek theatre`
Type: DICT
Translation: `Greek theater`
Type: DICT
Translation: `Greek theatre`

6 Conclusions

In this paper we have described and demonstrated our hybrid text translation service developed with the MultiMatch project for use in multilingual Digital Libraries. This combines a standard MT system with a domain-specific bilingual dictionary gathered automatically from Wikipedia. An experimental investigation using a query log file from the CH domain illustrated the ability of this approach to improve the suitability of translated queries for this domain. The automatic hybrid translation service was extended to an interactive service enabling users with some knowledge of the translation target language to adjust and augment the "best" automatically generated hybrid translation. The main objective of the interactive service is to incorporate the user's knowledge in order to improve translation quality for their search while minimising the time and effort that they must expend in doing this.

In further work we plan to extend the coverage of our dictionaries by exploring the mining of other translations pairs from within the linked Wikipedia pages. The interactive translation service could also be extended to record the translation adjustments made by the users, and to incorporate these in future translation of similar queries with the objective of increasing the likelihood of more often produced "best" translations which do not require user adjustment. Hence improving the average quality of translations provided to users with no knowledge of the target language who are not able to make corrective adjustments to the proposed translation. The service could be further extended to enable users to add additional entries to the bilingual dictionaries, although this would require participation of users able to suitable dictionary additions.

Acknowledgement

Work supported by the European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD — project MultiMATCH contract IST−033104. The authors are solely responsible for the content of this paper.

References

1. S. F. Adafre and M. de Rijke. Discovering Missing Links in Wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 90–97, Chicago, U.S.A., 2005.
2. S. F. Adafre and M. de Rijke. Finding Similar Sentences Across Multiple Languages in Wikipedia. In *Proceedings of EACL 2006*, pages 62–69, Trento, Italy, 2006.
3. L. Ballesteros and W. B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proceedings of SIGIR 1997*, pages 84–91, Philadelphia, U.S.A., 1997.
4. G. Bouma, I. Fahmi, J. Mur, G. van Noord, L. van der Plas, and J. Tiedemann. Using Syntactic Knowledge for QA. In *Evaluation of Multilingual and Multi-modal Information Retrieval - CLEF 2006*, pages 318–327, Alicante, Spain, 2006.
5. F. Coenen, P. H. Leng, R. Sanderson, and Y. J. Wang. Statistical Identification of Key Phrases for Text Classification. In *Machine Learning and Data Mining in Pattern Recognition*, pages 838–853. Springer, 2007.
6. T. Declerck, A. G. Pèrez, O. Vela, Z. Gantner, and D. Manzano-Macho. Multilingual Lexical Semantic Resources for Ontology Translation. In *Proceedings of LREC 2006*, pages 28–32, Genoa, Italy, 2006. ELDA.
7. J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of RIAO 2000*, pages 1–20, Paris, France, 2000.
8. D. He, J. Wang, D. W. Oard, and M. Nossal. Comparing User-Assisted and Automatic Query Translation. In *Advances in Cross-Language Information Retrieval - CLEF 2002*, pages 400–415, 2002.
9. D. A. Hull and G. Grefenstette. Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In *Proceedings of SIGIR 1996*, pages 49–57, Zurich, Switzerland, 1996.
10. G. J. F. Jones, T. Sakai, N. Collier, A. Kumano, and K. Sumita. A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval. In *Proceedings of SIGIR 1999*, pages 269–270, Berkeley, CA, U.S.A., 1999.
11. A. M. Lam-Adesina and G. J. F. Jones. Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval. In *Comparative Evaluation of Multilingual Information Access Systems - CLEF 2003*, pages 271–285, Trondheim, Norway, 2003.
12. P. Pecina, P. Hoffmannová, G. J. F. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF-2007 Cross-Language Speech Retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval - CLEF 2007*, pages 674–686, Budapest, Hungary, 2007.
13. D. Petrelli, P. Hansen, M. Beaulieu, and M. Sanderson. User Requirement Elicitation for Cross-Language Information Retrieval. *New Review of Information Behaviour Research*, 3:17–35, 2002.
14. M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.
15. T. Van de Cruys and B. n. Villada Moirón. Semantics-Based Multiword Expression Extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June 2007. ACL.
16. Y. Zhang, G. J. F. Jones, and K. Zhang. Dublin City University at CLEF 2007: Cross-Language Speech Retrieval (CL-SR) Experiments. In *Advances in Multilingual and Multimodal Information Retrieval - CLEF 2007*, pages 703–711, Budapest, Hungary, 2007.