Domain-Specific Query Translation for Multilingual Access to Digital Libraries

Gareth J. F. Jones, Fabio Fantino, Marguerite Fuller, Eamonn Newman, Ying Zhang
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

Abstract

Accurate high-coverage translation is a vital component of reliable cross language information access (CLIR) systems. This is particularly true of access to 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 information retrieval evaluation workshops, it is not well suited to specialized tasks where domain specific translations are required. We demonstrate that effective query translation in the domain of cultural heritage (CH) can be achieved by augmenting a standard MT system with domain-specific phrase dictionaries automatically mined from the online *Wikipedia*. 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

Reliable translation is a key component of effective Cross Language Information Access (CLIR) systems. Various approaches to translation have been explored at evaluation workshops such as TREC[1], CLEF[2] and NTCIR[3]. Experiments at these workshops have been based on laboratory collections consisting of news articles or technical reports with "TREC" style queries with a minimum length of a full sentence. In such cases general purpose translation resources based on bilingual dictionaries and standard machine translation (MT) have been shown to be effective for translation in CLIR.

This approach using general translation resources will often not be suitable for queries to Multilingual Digital Libraries (DLs) which often contain domain-specific terms which must be translated accurately if effective content retrieval is to be achieved. One DL domain of which this is true is cultural heritage (CH). The EU FP6 *MultiMatch*[4] project 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 report on the MultiMatch query translation methods which have been developed to deal with domain-specific language in the CH domain. We demonstrate the effectiveness of these techniques using sample CH query logs in English, Spanish and Italian. We translate the queries and examine the quality of these translations using human assessors. We show how a domain-specific phrase dictionary can be used to augment a traditional general MT system to improve the coverage and reliability of translation of these queries.

The remainder of this paper is organized as follows: Section 2 introduces the translation resources used for this study, Section 3 describes our experimental investigation, and Section 4 summarizes our conclusions.

2 Query Translation Techniques

The MT approach to query translation for CLIR uses an existing MT system to provide a single automatic translation. Results reported at the CLIR evaluation workshops have often shown it to be competitive with other translation methods. However, while MT systems can provide reasonable 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 a specific domain. For example, the Spanish phrase "Canto general" is translated by the standard MT system used in our work

[1] trec.nist.gov
[2] http://www.clef-campaign.org/
[3] http://research.nii.ac.jp/ntcir/
[4] www.multimatch.org

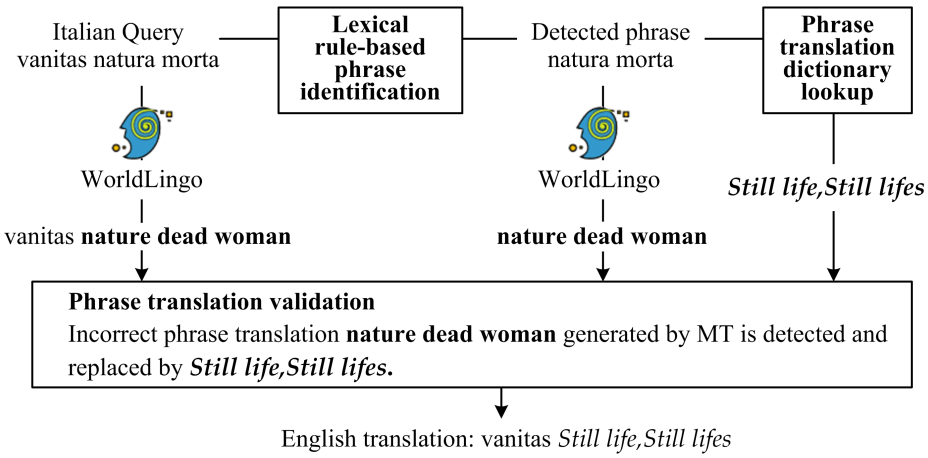

Figure 1: An example of Italian–English query translation.

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 often domain-specific and are typically absent from static lexicons. Effective translation of such phrases is particularly critical for short queries that are typically entered by non-expert users of search engines.

The focus of the research reported in this paper is a method to improve translation effectiveness of phrases previously untranslated or inappropriately translated by a standard MT system. In this work we combine an MT system with domain-specific phrase dictionaries mined from the online *Wikipedia*. The next sections describe the construction of our dictionaries and their combination with the MT system.

2.1 Phrase Dictionary Construction

Our phrase translation system uses domain-specific phrase dictionaries built by mining the online Wikipedia[5]. As a multilingual hypertext medium, Wikipedia has been shown to be a valuable new source of translation information (Adafre and de Rijke, 2005; Adafre and de Rijke, 2006; Bouma et al., 2006; Declerck et al., 2006). 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 terms. 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. 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:

1. Perform 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 was restricted to the category of culture including all of its recursive subcategories. In total, we collected $458,929$ English pages.

2. For each English page the hyperlinks to each of the query languages were extracted (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 are added into the phrase dictionary for each language.

The compiled dictionaries contain about $90,000$, $70,000$, and $80,000$ distinct multiple-word phrases in English, Italian, and Spanish respectively. The majority of the phrases extracted are CH domain-specific named entities and the rest of them are 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 are located in the wikipedia archive, the alternative translations are ranked in the extended bilingual dictionaries by frequency of occurrence in the wikipedia pages.

[5] `http://wikipedia.org`

2.2 Improved MT-based Translation

Figure 1 shows our query translation process which proceeds as follows:

Lexical rule-based phrase identification Given a query, the first task is to locate phrases. To do this, we adopt a lexical rule-based approach with maximum forward matching (Ballesteros and Croft, 1997; Hull and Grefenstette, 1996), since it has been shown to have robust performance and is computationally simple. The query is sequentially scanned to match the phrase dictionary. The longest matched subsequence is taken as a phrase and translated via a domain-specific dictionary lookup. This process is recursively invoked on the remaining parts of the query until no further matches are found. The performance of this approach depends strongly on the completeness of the coverage of the adopted dictionary. Analysis of our test query results showed that at least one phrase is detected in 90% of the testing queries, for example, personal names, geographic locations, and titles of various types of artworks. This indicates that the phrase dictionaries we compiled can be used to identify domain-specific phrases in web queries.

WorldLingo machine translation We translate the original query into the target language using the WorldLingo[6] MT system. WorldLingo was selected for the MultiMatch project because it generally provides good translation between English, Spanish, Italian, and Dutch — the languages relevant to the Multimatch project. In addition, it provides a useful API that can be used to translate queries in real-time via HTTP transfer protocol.

Phrase translation validation Each of the recognised dictionary phrases is passed to the MT system. The translation T_{mt} of this phrase returned by WorldLingo is then replaced in the WorldLingo translation of the complete query by the translations(s) T_{dict} from our domain-specific dictionary, if $\mathrm{T}_{mt} \neq \mathrm{T}_{dict}$. This potentially enables us to correct inaccurate phrase translations generated by the MT system.

3 Experimental Investigation

We performed an experimental investigation to compare query translation accuracy of our domain-specific hybrid approach with the standard MT output. 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 information retrieval test collection, we based our experiments on real query log data provided by organisations providing search to CH archives.

[6] `http://worldlingo.com`

3.1 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 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 Digital Library based in Spain whose focus is on poetry and ancient and modern literature in the Spanish language. The 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 terms), 1088 Spanish queries (3.39 terms on average) and 100 English queries (1.67 terms 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 are interested in the disagreements between the systems. These sets of translations are denoted *Es-En*, *It-En*, *En-Es* and *En-It*. The remaining translations were collated so that the evaluators could have 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 is not possible to directly compare the lexical coverage of our domain-specific dictionaries and the built-in phrase dictionaries of WorldLingo since we do not have access to the internal WorldLingo dictionaries.

3.2 Human Evaluation of Translation Quality

For each query 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 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 result was produced for 1919 leaving 792 to be evaluated.

Table 2 shows that the hybrid translation system was generally regarded as providing a better translation. For Spanish-English, the hybrid translation was correct in 79% of the cases where there was a disagreement be-

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

Table 1: Query translation examples.

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 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	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

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

tween the systems. "No preference" results indicate that the evaluator felt that neither translations was appropriate. For Italian-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 79.34% correctly translated by the hybrid system. Situations where both are deemed "correct" raises 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 any significance to the results, however for the sake of completeness, we can report correct translation rates of 81.82% 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 methods for enhancing an MT system by incorporating domain-specific dictionaries are successful for query translation. By identifying phrases and named entities which have a special meaning within the domain, we were able to improve upon the baseline translation 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. 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 features are described in alternative equally correct ways. 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 reexamined showing the alternative translations contained in the hybrid dictionary to the evaluators. In many cases, one of the alternative hybrid translations matched the MT system translation exactly, or matched when stopwords were removed. Table 3 shows the updated results of in-

cluding 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 they were regarded as both correct arise cases where the output from the two systems was so similar as to effectively be functionally identical.

While we are not able to manually evaluate the accuracy of all translation pairs in our bilingual dictionaries due to limited resources. However, the of our translation results for a set of sample user queries in the CH domain demonstrate that our translations are generally highly accurate.

3.3 Related Experiments

The objective of our hybrid translation system is ultimately to improve CLIR accuracy. Since we did not have access to a suitable set of documents and corresponding relevance data for our user search topics, we conducted a preliminary set of CLIR experiments using a different IR test collection. We used the CLEF 2007 Cross Language Speech Retrieval (CL-SR) task. This consists of a small collection of about 8000 documents and 42 search topics with corresponding relevance data indicating which documents are relevant to each query. This provides an interesting test for search technologies within the MultiMatch project since it is a (non-CH) domain specific cross language multimedia retrieval task. However, the topic statements are generally rather longer than those typically entered into a web search engine. For the CLEF task we trained new bilingual dictionaries for the relevant in the domain of the CL-SR data set (issues relating to World War Two). These were then used in combination with a standard MT system to perform a set of comparative experiments exploring alternative translation strategies. The full results of these experiments were reported in (Zhang et al., 2008). 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 our best submitted monolingual run was slightly less than (although not significantly) the best submission, our submitted result for the cross language task was the best showing the lowest decrease relative to monolingual performance. 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.

4 Conclusions

In this paper we have described and demonstrated our hybrid query translation method suitable for use in multilingual Digital Libraries. 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.

Acknowledgement

Work partially supported by 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

Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering Missing Links in Wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 90–97, Chicago, U.S.A. ACM Press.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding Similar Sentences Across Multiple Languages in Wikipedia. In *Proceedings of EACL 2006*, pages 62–69, Trento, Italy.

Lisa Ballesteros and W. Bruce Croft. 1997. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proceedings of SIGIR 1997*, pages 84–91, Philadelphia, U.S.A. ACM Press.

Gosse Bouma, Ismail Fahmi, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jorg Tiedemann. 2006. The University of Groningen at QA@CLEF 2006 using Syntactic Knowledge for QA. In *Proceedings of CLEF 2005*, Alicante, Spain.

Thierry Declerck, Asunciòn Gòmez Pèrez, Ovidiu Vela, Zeno Gantner, and David Manzano-Macho. 2006. Multilingual Lexical Semantic Resources for Ontology Translation. In *Proceedings of LREC 2006*, Genoa, Italy. ELDA.

David A. Hull and Gregory Grefenstette. 1996. Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. In *Proceedings of SIGIR 1996*, pages 49–57, Zurich, Switzerland. ACM Press.

Ying Zhang, Gareth J. F. Jones, and Ke Zhang. 2008. Dublin City University at CLEF 2007: Cross-Language Speech Retrieval (CL-SR) Experiments. In *Proceedings of CLEF 2008*, pages 703–711, Budapest, Hungary. Springer.