Domain-Specific Query Translation for Multilingual Information Access using Machine Translation Augmented With Dictionaries Mined from Wikipedia

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

Abstract

Accurate high-coverage translation is a vital component of reliable cross language information access (CLIA) systems. While machine translation (MT) has been shown to be effective for CLIA tasks in previous evaluation workshops, it is not well suited to specialized tasks where domain specific translations are required. We demonstrate that effective query translation for CLIA can be achieved in the domain of cultural heritage (CH). This is performed 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 (CLIA) 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. Test collection design at these workshops often ensures that there are a reasonable number of relevant documents available for each query. 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 CLIA. However, this is less likely to be the case when translating the very short queries typically entered by general users of search engines, particularly when they are seeking information in a specific domain.

Online cultural heritage (CH) content is currently appearing in many countries produced by organisations such as national libraries, museums, galleries and audiovisual archives. Additionally, there are increasing amounts of CH relevant content available more generally on the World Wide Web. While some of this material concerns national or regional content only of local interest, much material relates to items involving multiple nations and languages, for example concerning events or groups encompassing large areas of Europe or Asia. In order to gain a full understanding of such things, including details contained in different collections and exploring different cultural perspectives, often requires effective multilingual search technologies.

CH content encompasses various different media, including of course text documents, but also images, videos, and audio recordings which may only be described by very limited metadata labels. Such metadata may include simple factual details such as date of creation, but also descriptive details relating to the contents of the item and interpretation and contextualization of the content. Multilingual

[1] trec.nist.gov

[2] http://www.clef-campaign.org/

[3] http://research.nii.ac.jp/ntcir/

searching using metadata content requires that either the metadata be translated into a language with which the user is able to search or that the search query be translated into the language of the metadata. This alternative of document or query translation is a well rehearsed argument in CLIA, which has generally concerned itself with full text document searching. However, the features of metadata require a more careful analysis. Metadata is typically dense in search terms, while lacking the linguistic structure and information redundancy of full text documents. The absence of linguistic structure makes precise translation of content problematic, while the lack of redundancy means that accurate translation of individual words and phrases between the query and document is vital to minimize mismatch between query and document terms. Developing reliable and robust approaches to translation for metadata search is thus an important component of search for many CH archives.

The EU FP6 *MultiMatch*[4] project is concerned with information access for multimedia and multilingual content for a range of European languages. In this paper we report on the MultiMatch query translation methods we are developing to deal with domain-specific language in the CH domain. We demonstrate the effectiveness of these techniques using example query logs from CH sites in English, Spanish and Italian. We translate the queries and examine the quality of these translations using human annotation. We show how a domain-specific phrase dictionary can be used to augment traditional general MT systems to improve the coverage and reliability of translation of these queries. We also show how retrieval performance on CH image metadata is improved with the use of these improved, domain-specific translations.

The remainder of this paper is organized as follows: Section 2 introduces the translation resources used for this study, Section 3 describes our experimental setup and results, Section 4 summarizes our conclusions, and Section 5 gives details of our ongoing work.

2 Query Translation Techniques

The MT approach to query translation for CLIA uses an existing MT system to provide automatic translation. Using MT systems for query translation is widely used in CLIA when such a system is available for the particular language pair under consideration. Results reported at the standard retrieval 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 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 typically absent from static lexicons. Effective translation of such phrases is therefore 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,

[4] www.multimatch.org

[5] http://wikipedia.org

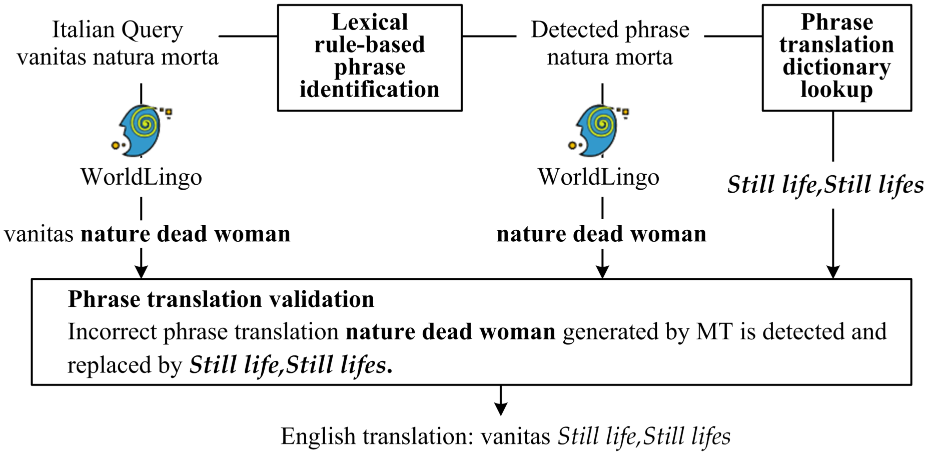
Italian Query
vanitas natura morta
Lexical rule-based phrase identification
Detected phrase
natura morta
Phrase translation dictionary lookup
WorldLingo
WorldLingo
Still life,Still lifes
vanitas nature dead woman
nature dead woman
Phrase translation validation
Incorrect phrase translation nature dead woman generated by MT is detected and replaced by Still life,Still lifes.
English translation: vanitas Still life,Still lifes

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

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 implement a three-stage automatic process to mine wikipedia pages as a translation source and construct phrase dictionaries in the culture heritage domain.

1. First, 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 is 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 obtained, we extracted the hyperlinks to each of the query languages (Italian and Spanish).

3. We then selected the basenames of each pair of hyperlinks (English–Italian, English–Spanish) as translations and added them into our domain-specific dictionaries. The multiple-word phrases were added into the phrase dictionary for each language. These phrase dictionaries are later used for dictionary-based phrase identification.

The dictionaries we compiled 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 play an equally important role in the query translation process as domain-specific named entities.

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. Three methods of multiple-word phrase identification have been commonly used: lexical rule-based (Ballesteros and Croft, 1997; Hull and Grefenstette, 1996), statistical (Coenen et al., 2007; Gao et al., 2001), and syntactical methods (Sharma and Raman, 2003; Gelbukh et al., 2004; Van de Cruys and Villada Moirón, 2007). The lexical rule-based approach with maximum forward matching was adopted in our query translation process due to its robust performance and computational simplicity. 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 part of the query until no matches are found. The performance of this approach depends strongly on the completeness of the coverage of the adopted dictionary. Our experimental 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 accurately identify 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 For each of the phrases previously recognized, we again pass it to the MT system and the translation T_{mt} of this phrase is returned by WorldLingo. T_{mt} is then replaced in the WorldLingo translation of the query by the translations(s) T_{dict} from our domain-specific dictionary, if $\mathrm{T}_{mt} \neq \mathrm{T}_{dict}$. This allows us to correct unreliable phrase translations generated by the MT system.

3 Experimental Investigation

The goal of our experiments was to evaluate the usefulness and the accuracy of the domain-specific translation dictionaries. Instead of using queries from a standard information retrieval test collection, we experimented with queries explicitly seeking CH information from real query log data provided by CH organisations.

3.1 Query Log

The query log data used in this investigation was provided by three European CH organisations participating in the MultiMatch project, and is taken from their archives of real user queries. The data consists of 100 English, 1048 Italian, and 1088 Spanish distinct web queries and the number of hits of each query. The top 200 most popular multiple-word queries in Italian and Spanish were selected as the queries for testing. Due to the smaller size of the English query log, we only obtained English 53 phrasal queries.

	# Detected by dictionaries	# Untranslated by WorldLingo	Proportion
EN–IT	14	11	79%
EN–ES	19	11	58%
IT–EN	83	33	40%
ES–EN	74	33	45%

Table 1: Number of detected phrases using the domain-specific dictionaries.

	Total	# Exactly correct	# + Extra translations	# + Minor noise
EN–IT	14	13	1	0
EN–ES	19	17	1	1
IT–EN	83	40	43	0
ES–EN	74	37	5	32

Table 2: Correctness of the translations of detected domain-specific phrases.

We used two methods of evaluation: first, the dictionary usefulness and the translation effectiveness are judged extrinsically by human assessment; and second, evaluation using a parallel Italian–English metadata document set explored how translation affects the retrieval performance of an information retrieval system.

3.2 Human Judgement Evaluation

The WorldLingo MT system was used to translate Spanish and Italian queries into English and vice versa. Our domain-specific dictionaries were used to translate phrases within the queries into the same target languages. 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 don't have access to the internal WorldLingo dictionaries.

To evaluate the usefulness of our dictionaries, we observed the proportion of domain-specific phrases in the various query sets that can be translated using our domain-specific dictionaries mined from the web, but are incorrectly translated by WorldLingo.

[6] http://worldlingo.com

Original Query	WorldLingo Translation	Improved Machine Translation
EN–IT		
turner east sussex	Turner Sussex orientale	Turner *East Sussex*
still life flowers	fiori di vita tranquilla	fiori di *Natura morta*
francis bacon	Francis Bacon	*Francesco Bacone*
pop art	arte di schiocco	*Pop art*
m c escher	escher di m. c	*Maurits Cornelis Escher*
american 60's	americano 60's	americano *Anni 1960*
EN–ES		
vanessa bell	campana del vanessa	*Vanessa Bell*
turner east sussex	Turner sussex del este	Turner *East Sussex*
henry moore	moore del Henrio	*Henry Moore*
still life flowers	flores de la vida inmóvil	flores de *Bodegón*
guerrilla girls	muchachas del guerrilla	*Guerrilla Girls*
IT–EN		
leonardo da vinci	leonardo from you win	*Da Vinci, Leonardo da Vinci,* *Leonardo daVinci, Leonardo de Vinci*
duomo di milano	dome of Milan	*Cathedral of Milan, Duomo di Milan,* *Duomo di Milano, Duomo of Milan, Milan Cathedral*
beni culturali	cultural assets	*Cultural heritage*
arte povera	poor art	*Arte povera*
san lorenzo	saint lorenzo	*Lawrence of Rome, Saint Lawrence, St Lawrence,*
gentile da fabriano	kind from fabriano	*Gentile da Fabriano*
statua della liberta	statue of the freedom	*Statue of Liberty*
aldo rossi	aldo red	*Aldo Rossi*
arnaldo pomodoro	arnaldo tomato	*Arnaldo Pomodoro*
la cattura di cristo di caravaggio	the capture of caravaggio Christ	*The Taking of Christ* caravaggio
ES–EN		
lope de vega	lope of fertile valley	*Lope de Vega*
literatura infantil	infantile Literature	*Children's book, Children's books,Children's literature*
cantar de mio cid	to sing of mine cid	*Cantar de mio Cid, Lay of the Cid, The Lay of the Cid*
el quijote de la mancha	quijote of the spot	quijote of *La Mancha*
dulce maria loynaz	candy Maria loynaz	*Dulce María Loynaz*
andres bello	andres beautiful	*Andrés Bello*
filosofia del derecho	philosophy of the right	*Philosophy of law*
elogio de la locura	praise of madness	*In Praise of Folly, Praise of Folly, The Praise of Folly*
la regenta	it runs it	*La Regenta*
cristobal colon	cristobal colon	*Christopher Colombus, Christopher Columbus,* *Cristopher Columbus*

Table 3: Some examples of improved translations using the domain-specific dictionaries. (The corrected phrase translations are in italic.)

Namely, we tested the ability of our system to detect and correct the presence of unreliable MT translations for domain-specific phrases. Translated phrases for these queries can generally be judged unambiguously as correct or incorrect by a bilingual speaker of the languages involved, and so we are confident that assessment of translation accuracy here does not involve significant degrees of subjectivity.

As shown in Table 1, we can see that 79%, 58%, 40%, and 45% of incorrect MT-translated phrases were able to be corrected using the domain-specific dictionaries mined from wikipedia, in EN–IT, EN–ES, IT–EN, and ES–EN translation tasks, respectively. Our system leads to a large improvement in MT translation for domain-specific phrases. Some examples of improved query translations are shown in Table 3.

We also conducted an investigation on the correctness of the translation mined from wikipedia, as shown in Table 2. *Exact correct translation* is strictly-correct single translation. *Extra translation* refers to strictly-correct multiple translations, for example, "Cathedral of Milan, Duomo di Milan, Duomo di Milano, Duomo of Milan, Milan Cathedral" (Italian: Duomo di Milano). It is interesting to observe that about 50% of Italian phrases are found to have multiple correct English translations due to multiple English wikipedia pages being redirected to the same Italian pages. Some *minor noise* is observed when the correct translation contains some related additional words, such as "Alfonso XII of Spain" (Spanish: Alfonso XII). When used for information retrieval, this additional information can sometimes improve effectiveness.

We are not able to manually evaluate the accuracy of all translation pairs in our bilingual dictionaries due to limited resources. However, our results for sample queries from user logs demonstrate that our translations are generally highly accurate.

3.3 Intrinsic Evaluation Using IR System

Our information retrieval experiments were performed on a database of metadata associated with a collection of 5000 CH photographs. The metadata to describe each artifact in the collection is available in English and in Italian. Each photograph is described identically in both languages. We formed a separate search index for English and Italian. Search was carried out using the Lucene search engine[7]. We carried out an evaluation based on this collection which proceeded as follows:

1. Submit the original queries to the index and record the ranked list of references returned.
2. Submit the translated queries to the appropriate index and record the ranked list of references returned.
3. Find the correlation between the lists returned for the native language queries and the queries translated to that language.
4. The better translation will have the stronger correlation with the native language list.

Due to the fact that the corpus was only complete in the Italian and English versions, we were unable to include the Spanish queries in this part of the evaluation. Also, while this collection is based in the CH domain, some of the queries yield no relevant documents due to their specialist nature. The collection of queries for which meaningful retrieval results are returned is too small to allow for a quantitative analysis of retrieval effectiveness. Therefore, we present a qualitative analysis of some of the more interesting cases.

3.3.1 Italian–English translations

The Italian queries cover a wide range of Italian interests in CH. We present here a sample of some of the more interesting results.

Arnaldo Pomodoro This refers to an Italian artist, but the name "Pomodoro" is translated to "Tomato" in English by WorldLingo. While there were no references to the artist in the collection, all documents returned contained the term "tomato" (referring to the vegetable) which are irrelevant to the query. The dictionary-based translation recognized the name and therefore left it untranslated. It is preferable to retrieve no documents rather than to retrieve irrelevant ones.

Amore e Psiche This refers to the sculpture entitled "Cupid and Psyche" in English. This phrase was matched in our phrase dictionary and translated correctly. The MT system translated this as "Love, Psyche". The dictionary translation was observed to retrieve relevant documents with greater precision since it matched against the more specific term "Cupid", as opposed to the more general term "Love".

David Michaelangelo This query provided a counterexample. The phrase dictionary added the term "statue" to the translated query. This led to retrieval of a large number of non-relevant documents.

3.3.2 English–Italian translations

As with the Italian queries, there was not much overlap between the query log and the document collection. Some of the interesting translations include:

pop art This phrase was recognized by our domain-specific dictionary, and so was left in its original form for searching in Italian. Interestingly, this led to an improvement in search accuracy for the query compared to that in the English language collection. For the English index, this phrase matched many non-relevant documents which contained the word "art". However, when searching in the Italian index, where "art" is not a word encountered in the

[7] `http://lucene.apache.org/`

general vocabulary, the phrase retrieves only 7 documents, of which 5 were relevant.

Turner East Sussex The place name "East Sussex" was correctly recognized and translated by our phrase dictionary. However the MT system again failed to recognise it and translated the partial term "East" to "Orientale". The presence of the term "Orientale" in the translated query resulted in many non-relevant documents being retrieved, reducing the precision of the query.

The examples given in this section provide anecdotal evidence to support the view that the automatically mined domain-specific phrase dictionary improves the performance of the retrieval system. Query sets and relevance judgements are being created for the MultiMatch document set by domain experts who compiled the original collections. Thus we will be able to ensure that the query sets are a good representative sample of the information needs of the typical user. These test collections will allow us to conduct full quantitative analysis of our system.

4 Conclusions

We have presented an automatic mining system developed for construction of domain-specific phrase dictionaries. Phrases not translated by a general MT system are shown to be translated effectively using these dictionaries. The extracted translations were evaluated by human assessment and shown to be highly accurate. We have also demonstrated a way to combine these dictionaries with MT for topical phrases in the culture heritage domain. Our experimental results show that we were able to detect and correct a large proportion of domain-specific phrases unsuccessfully translated by MT, and thus improve information retrieval effectiveness and facilitate MLIA.

5 Ongoing Work

In our ongoing work we plan to further extend the coverage of our dictionaries by exploring the mining of other translations pairs from within the linked *Wikipedia* pages. While the method described in this paper has been shown to be effective for query translation, we have so far only demonstrated its behavior for a very small number of queries to our CLIA system. We are currently developing test collections based on several CH data sets to evaluate the effectiveness of our hybrid query translation method.

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. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

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, Illinois, United States. ACM Press.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 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 the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91, Philadelphia, PA, USA. 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 *Working Notes for the Cross Language Evaluation Forum 2006 Workshop*, Alicante, Spain.

Frans Coenen, Paul H. Leng, Robert Sanderson, and Yanbo J. Wang. 2007. Statistical identification of key phrases for text classification. In *Machine Learning and Data Mining in Pattern Recognition*, volume 4571 of *Lecture Notes in Computer Science*, pages 838–853. Springer.

Thierry Declerck, Asunciòn Gòmez Pèrez, Ovidiu Vela, Zeno Gantner, and David Manzano-Macho. 2006. Multilingual lexical semantic resources for ontology translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy. ELDA.

Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, and Changning Huang. 2001. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in information retrieval*, pages 96–104, New Orleans, Louisiana, United States. ACM Press.

Alexander F. Gelbukh, Grigori Sidorov, Sang-Yong Han, and Erika Hernández-Rubio. 2004. Automatic syntactic analysis for detection of word combinations. In *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2945 of *Lecture Notes in Computer Science*, pages 243–247. Springer.

David A. Hull and Gregory Grefenstette. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57, Zurich, Switzerland. ACM Press.

Rupali Sharma and S. Raman. 2003. Phrase-based text representation for managing the web documents. In *Proceedings of the International Conference on Information Technology: Computers and Communications*, page 165, Washington, DC, USA. IEEE Computer Society.

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.