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, DCU, Ireland
Outline

- MultiMatch search system
- MT-based query translation
- Domain-specific dictionary construction
- Experimental investigation
- Results and discussion
- Conclusion and future work
MultiMatch Search System

- Characteristics of Culture Heritage achieves
  - Multilingual
  - Multimedia (texts, images, videos, audio)

- MultiMatch project
  Provide information access for multimedia and multilingual CH content for a range of European languages.
MultiMatch Cross-lingual Search

Italian-to-Dutch

Query (in Italian) → Query Translation → Query (in Dutch) → Lucene

Ranked Documents (in Italian) → Document Translation → Ranked Documents (in Dutch)

Separate index for each language

NL
IT
ES
EN
Machine-based Translation

- WorldLingo commercial machine translation system used under licence
- Supports all 12 language pairs for the four selected languages
- Easy to use and integrate into prototype
- Well-documented API
Motivations

- MT is able to provide reasonable translations for general terms.
- **Not sufficient** for domain-specific terms (in particular, multiple-word phrases).
  - Personal names
  - Organization names
  - Location names
  - Titles of art works
To improve translation accuracy of phrases previously untranslated or inappropriately translated by a standard MT system, and thus improve the CLIR effectiveness and facilitate MLIA.

Solution
Augmented MT combining domain-specific dictionaries mined from the web.
Multilingual wikipedia

A wikipedia page written in one language can contain hyperlinks to its counterparts in other languages: *titles* and *basenames* are translation pairs.

For example …
Hyperlink Feature of Wikipedia

Hyperlink Feature of

In other languages

Mona Lisa, or La Gioconda (La Joconde), is a 16th century oil painting on poplar wood by Leonardo da Vinci, and is arguably the most famous painting in the world. Few works of art have been subject to as much scrutiny, study, mythologizing and parody. It is owned by the French government and hangs in the Musée du Louvre in Paris. The painting, a half-length portrait, depicts a woman whose gaze meets the viewer's with an expression often described as enigmatic.

Title of the painting

A 3-stage automatic process

1. **Crawling** the English wikipedia, *Category: Culture* (pages and subcategories).
2. **Extracting** hyperlinks to query languages (Italian and Spanish).
3. **Generating** translation pairs using hyperlink basenames.

(The multiple-word phrase were added into the *phrase dictionary* for each language.)
Experimental Investigation

- To evaluate the usefulness and the accuracy of the domain-specific translation dictionaries.
- On sample query logs from users of cultural heritage websites.
- To test the ability of our system to detect and correct the presence of unreliable MT translations for domain-specific phrases.
Hybrid Query Translation Process

- Dictionary-based phrase translation
  - Lexical rule-based phrase identification
  - Phrase translation
- WorldLingo machine translation
  - For both the query and the phrases detected.
- Phrase translation validation
  - For each of the recognized phrases, replaced its WorldLingo translation by the translation(s) from our domain-specific dictionary, if they are not identical.
Hybrid Query Translation Example

Italian Query: vanitas natura morta

Lexical rule-based phrase identification

Detected phrase: natura morta

Phrase translation dictionary lookup

WorldLingo

vanitas nature dead woman

nature dead woman

Still life, Still lifes

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
Evaluation Methodology

- The top 200 popular multiple-word queries in Italian and Spanish.
- English 53 phrasal queries (Due to a smaller English query log).
- Human assessment
- How translation affects the retrieval performance of an IR system (our collection is too small to allow for a full quantitative analysis).
Human Judgement Evaluation

<table>
<thead>
<tr>
<th></th>
<th># Detected by dictionaries</th>
<th># Untranslated by WorldLingo</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN–IT</td>
<td>14</td>
<td>11</td>
<td>79%</td>
</tr>
<tr>
<td>EN–ES</td>
<td>19</td>
<td>11</td>
<td>58%</td>
</tr>
<tr>
<td>IT–EN</td>
<td>83</td>
<td>33</td>
<td>40%</td>
</tr>
<tr>
<td>ES–EN</td>
<td>74</td>
<td>33</td>
<td>45%</td>
</tr>
</tbody>
</table>

Our system leads to a significant improvement in MT translation for domain-specific phrases.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th># Exactly correct</th>
<th># + Extra translations</th>
<th># + Minor noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN–IT</td>
<td>14</td>
<td>13</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EN–ES</td>
<td>19</td>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IT–EN</td>
<td>83</td>
<td>40</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>ES–EN</td>
<td>74</td>
<td>37</td>
<td>5</td>
<td>32</td>
</tr>
</tbody>
</table>

50% of Italian phrases are found to have multiple correct translations due to multiple English wikipedia pages being redirected to the same Italian pages. Minor noise information sometimes can also improve effectiveness.
## Some Translation Examples

<table>
<thead>
<tr>
<th>Italian Query</th>
<th>WorldLingo English translation</th>
<th>Domain-specific English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>leonardo da vinci</td>
<td>leonardo from u win</td>
<td>Leonardo da Vinci</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leonardo de Vinci</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leonardo da Vinci</td>
</tr>
<tr>
<td>beni culturali</td>
<td>cultural assets</td>
<td>Cultural Heritage</td>
</tr>
<tr>
<td>san lorenzo</td>
<td>saint lorenzo</td>
<td>Lawrence of Rome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Saint Lawrence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>St Lawrence</td>
</tr>
<tr>
<td>gentile da fabriano</td>
<td>kind from fabriano</td>
<td>Gentile da Fabriano</td>
</tr>
<tr>
<td>statua della liberta</td>
<td>statue of the freedom</td>
<td>Statue of Liberty</td>
</tr>
<tr>
<td>arnaldo pomodora</td>
<td>arnaldo tomato</td>
<td>Arnaldo Pomodoro</td>
</tr>
</tbody>
</table>
Conclusion and Future Work

- We are able to detect and correct a large proportion of unsuccessfully translated domain-specific phrases by MT, and thus improve CLIR effectiveness and facilitate MLIA.
- We are currently developing test collections based on several CH datasets to evaluate the effectiveness of our hybrid query translation method.
The End

Thank you for your attention 😊