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

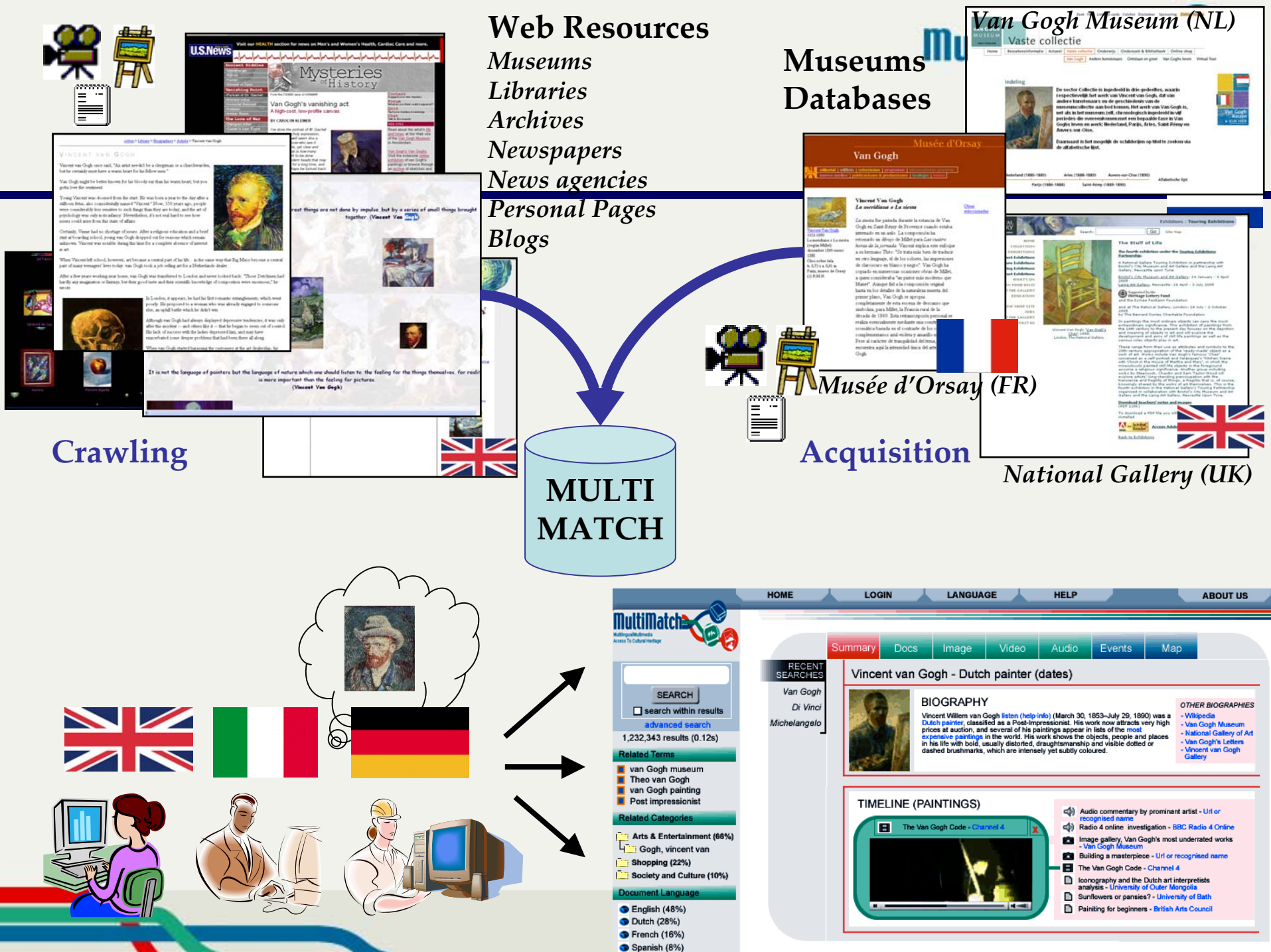
Outline

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

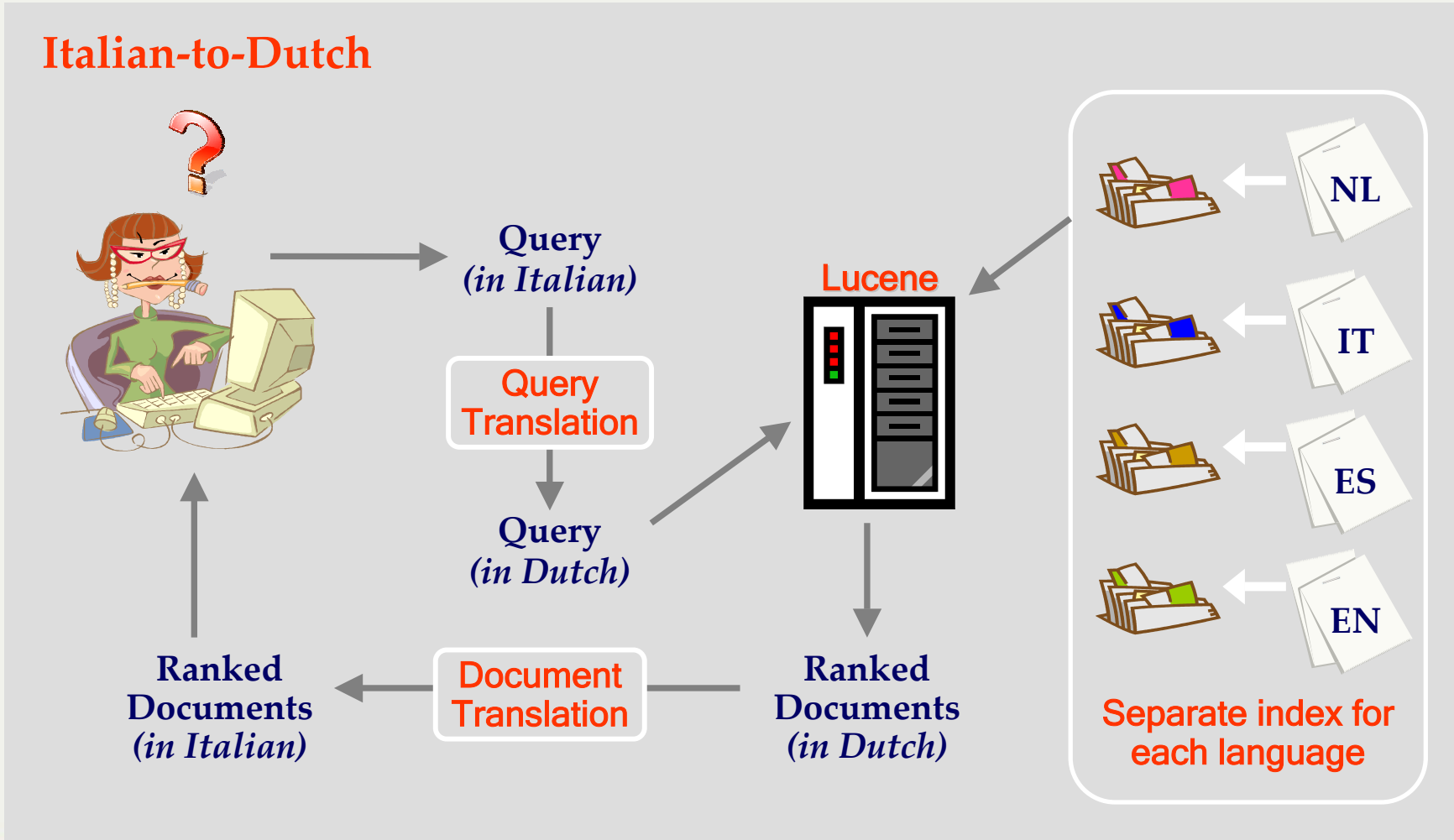
MultiMatch Search System

- 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.


Web Resources
Museums
Libraries
Archives
Newspapers
News agencies
Personal Pages
Blogs
Museums
Databases
Van Gogh Museum (NL)
Musée d'Orsay (FR)
National Gallery (UK)
Crawling
Acquisition
MULTI
MATCH
HOME
LOGIN
LANGUAGE
HELP
ABOUT US
MultiMatch
SEARCH
search within results
advanced search
1,232,343 results (0.12s)
Related Terms
van Gogh museum
Theo van Gogh
van Gogh painting
Post impressionist
Related Categories
Arts & Entertainment (66%)
Gogh, vincent van
Shopping (22%)
Society and Culture (10%)
Document Language
English (48%)
Dutch (28%)
French (16%)
Spanish (8%)
Summary
Docs
Image
Video
Audio
Events
Map
RECENT SEARCHES
Van Gogh
Di Vinci
Michelangelo
Vincent van Gogh - Dutch painter (dates)
BIOGRAPHY
OTHER BIOGRAPHIES
- Wikipedia
- Van Gogh Museum
- National Gallery of Art
- Van Gogh's Letters
- Vincent van Gogh Gallery
TIMELINE (PAINTINGS)
The Van Gogh Code - Channel 4

MultiMatch Cross-lingual Search



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

Research Goal

- 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.

Domain-specific Dictionary Construction

- 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

Mona Lisa - Wikipedia, the free encyclopedia - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://en.wikipedia.org/wiki/Mona_Lisa

article | discussion | edit this page | history

Sign in / create account

Your continued donations keep Wikipedia running!

From Wikipedia, the free encyclopedia

For other uses, see Mona Lisa (disambiguation).

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.

in other languages

- العربية
- Bosanski
- Български
- Česky
- Dansk
- Deutsch
- Eesti
- Esperanto
- فارسی
- Français
- Galego
- 한국어
- Hrvatski
- Ilokano
- Íslenska
- עברית
- Kurdî / كوردي
- Lëtzebuergesch
- Lietuvių
- Magyar
- Bahasa Melayu

```
<a href="http://en.wikipedia.org/wiki/Mona_Lisa">English</a>
[EN]Mona Lisa

<a href="http://es.wikipedia.org/wiki/La_Gioconda">Español</a>
[ES]La Gioconda

<a href="http://it.wikipedia.org/wiki/La_Gioconda_(dipinto)">Italiano</a>
[IT]La Gioconda

<a href="http://nl.wikipedia.org/wiki/Mona_Lisa">Nederlands</a>
[NL]Mona Lisa
```

Mona Lisa

Leonardo Da Vinci, circa 1503–1507

Title of the painting [edit]

Dictionary Construction Process

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.)

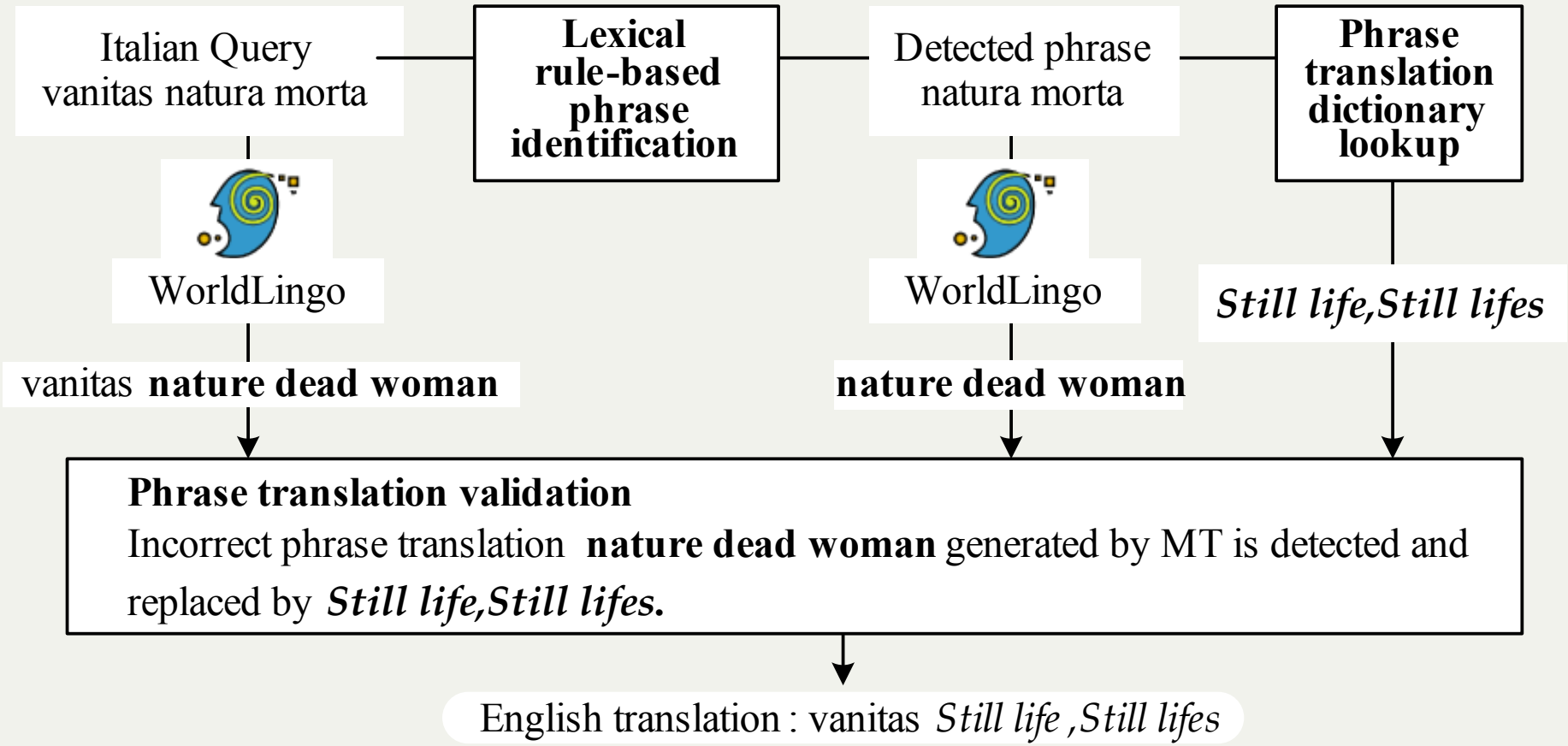
Experimental Investigation

- 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



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

	# 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%

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

	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

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

Italian Query	**WorldLingo English translation**	**Domain-specific English translation**
leonardo da vinci	leonardo from u win	Leonardo da Vinci Leonardo de Vinci Leonardo daVinci
beni culturali	cultural assets	Cultural Heritage
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
arnaldo pomodora	arnaldo tomato	Arnaldo Pomodoro

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 ☺