

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)
Vaste collectie

De sector Collectie is ingedeeld in drie groottes, waarin respectievelijk het meest van het werk van Van Gogh, dat van Jansz en Raaijmakers's en de geschilderde van de meesters die de laatste jaren van Van Gogh's leven, met alle het meest recente werk, de meest ingedeeld in vijf perioden die overeenstemmen met een bepaalde fase in Van Gogh's leven en werk: **Marktstraat, Paris, Arles, Saint Emery en Arles'se tijd.**

Doelstelling is het mogelijk de schiedrijzen op te helderen via de afbeeldingen hier.

Arles (1888-1889) Arles van Oost (1890) Arles van West (1890) Arles van Zuid (1890)

Arles (1888-1889) Arles van Oost (1890) Arles van West (1890) Arles van Zuid (1890)

Arles (1888-1889) Arles van Oost (1890) Arles van West (1890) Arles van Zuid (1890)

Arles (1888-1889) Arles van Oost (1890) Arles van West (1890) Arles van Zuid (1890)

U.S. News
Van Gogh's vanishing act
A highest-law-profile canvas

Mysteries of History
Van Gogh's vanishing act
A highest-law-profile canvas

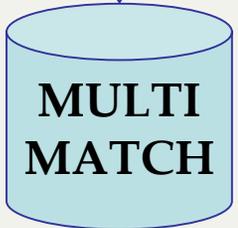
Vincent van Gogh
"I never saw Gogh once said, 'So what search for a description in a characteristic, but he certainly must have been known for his work here, but you give him the credit."

Musée d'Orsay
Van Gogh
Van Gogh's vanishing act
A highest-law-profile canvas

National Gallery (UK)
The stuff of life
The stuff of life
The stuff of life

Crawling

Acquisition



UK, Italy, Germany

Computer icons

MultiMatch
Access To Cultural Heritage

HOME LOGIN LANGUAGE HELP ABOUT US

Summary Docs Image Video Audio Events Map

RECENT SEARCHES
Van Gogh
Di Vinci
Michelangelo

Vincent van Gogh - Dutch painter (dates)

BIOGRAPHY
Vincent Willem van Gogh [listen \(help info\)](#) (March 30, 1853-July 29, 1890) was a Dutch painter, classified as a Post-impressionist. His work now attracts very high prices at auction, and several of his paintings appear in lists of the most expensive paintings in the world. His work shows the objects, people and places in his life with bold, usually distorted, draughtsmanship and visible dotted or dashed brushmarks, which are intensely yet subtly coloured.

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

Audio commentary by prominent artist - [Uri or recognised name](#)
Radio 4 online investigation - [BBC Radio 4 Online](#)
Image gallery, Van Gogh's most underrated works - [Van Gogh Museum](#)
Building a masterpiece - [Uri or recognised name](#)
The Van Gogh Code - [Channel 4](#)
Iconography and the Dutch art interpreters analysis - [University of Cluj-Napoca](#)
Sunflowers or pansies? - [University of Bath](#)
Painting for beginners - [British Arts Council](#)

Related Terms
van Gogh museum
Theo van Gogh
van Gogh painting
Post impressionist

Related Categories
Arts & Entertainment (68%)
Gogh, vincent van
Shopping (22%)
Society and Culture (10%)

Document Language
English (48%)
Dutch (28%)
French (18%)
Spanish (8%)

MultiMatch Cross-lingual Search

Italian-to-Dutch

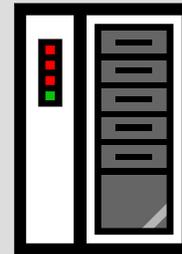


Query
(in Italian)

Query
Translation

Query
(in Dutch)

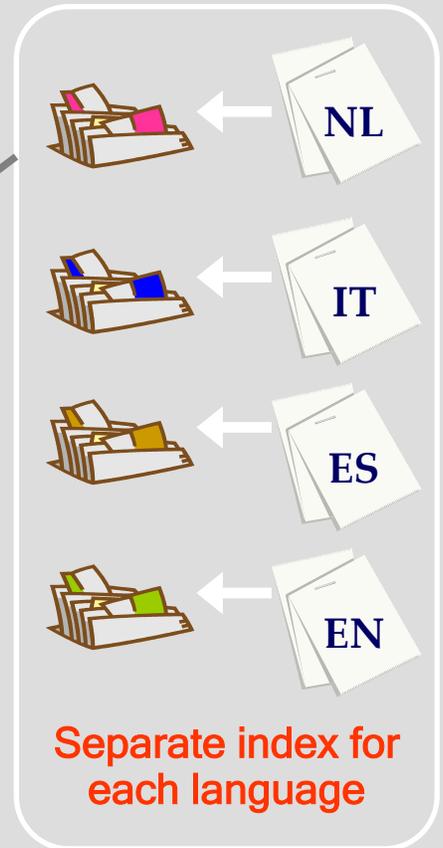
Lucene



Ranked
Documents
(in Dutch)

Document
Translation

Ranked
Documents
(in Italian)



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

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

Go Mona Lisa Leonardo

Sign in / create account

article discussion edit this page history

Your continued donations keep Wikipedia running!

in other languages

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

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.

[English](http://en.wikipedia.org/wiki/Mona_Lisa)

[EN]Mona Lisa

[Español](http://es.wikipedia.org/wiki/La_Gioconda)

[ES]La Gioconda

[Italiano](http://it.wikipedia.org/wiki/La_Gioconda_(dipinto))

[IT]La Gioconda

[Nederlands](http://nl.wikipedia.org/wiki/Mona_Lisa)

[NL]Mona Lisa

Mona Lisa

Leonardo Da Vinci, circa 1503–1507

Title of the painting [edit]

The title *Mona Lisa* stems from the Giorgio Vasari biography of Leonardo da Vinci, published 21 years after Leonardo's death. In it

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

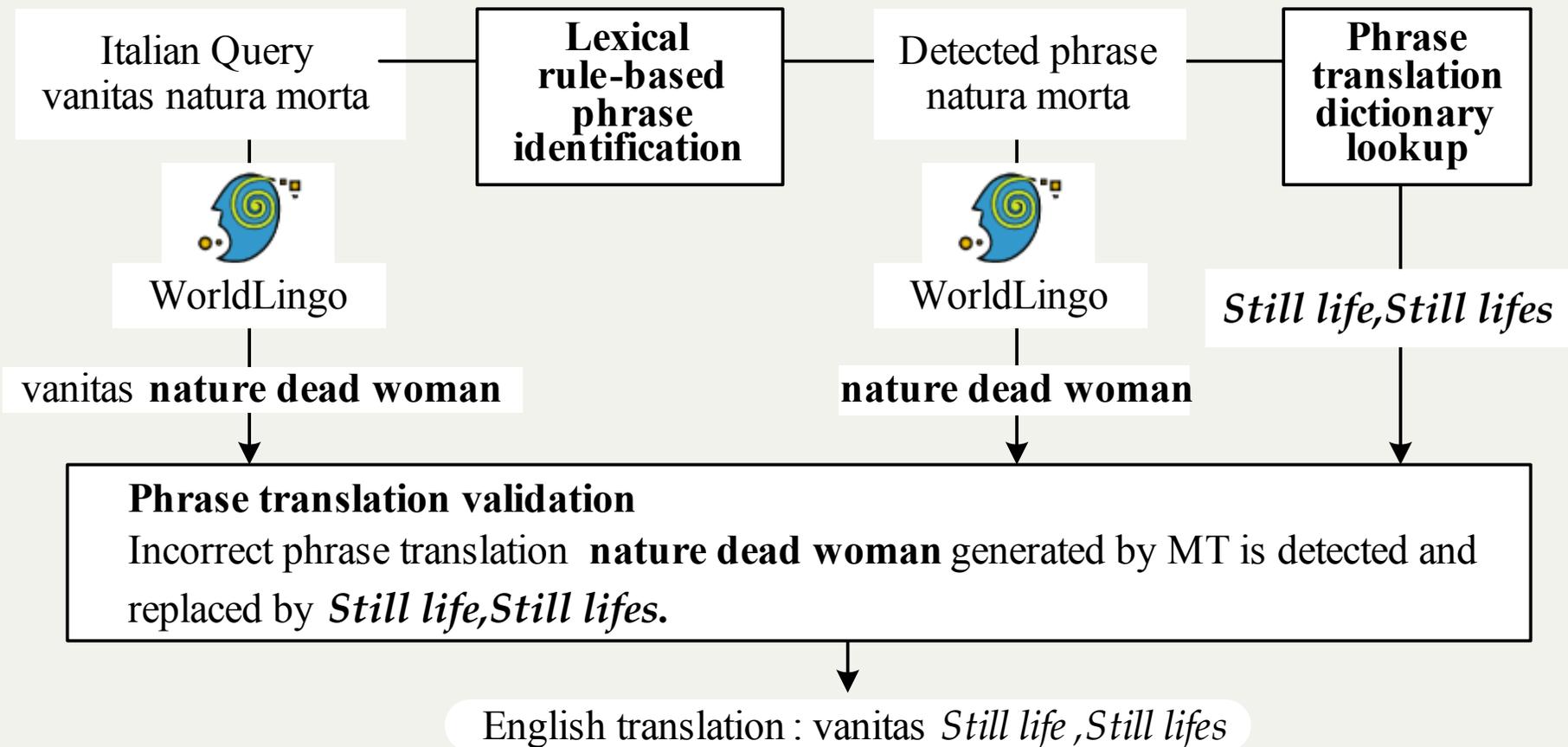
- To evaluate the usefulness and the accuracy of the domain-specific translation dictionaries.
- On sample query logs from users of cultural heritage websites.
- To test the ability of our system to detect and correct the presence of unreliable MT translations for domain-specific phrases.

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

<i>Italian Query</i>	WorldLingo English translation	Domain-specific English translation
<i>leonardo da vinci</i>	leonardo from u win	Leonardo da Vinci Leonardo de Vinci Leonardo daVinci
<i>beni culturali</i>	cultural assets	Cultural Heritage
<i>san lorenzo</i>	saint lorenzo	Lawrence of Rome Saint Lawrence St Lawrence
<i>gentile da fabriano</i>	kind from fabriano	Gentile da Fabriano
<i>statua della liberta</i>	statue of the freedom	Statue of Liberty
<i>arnaldo pomodora</i>	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 😊