Reconsidering Domain-Specific Cross-Language Information Access in the Age of Distributional Semantics

Gareth J. F. Jones

ADAPT Centre, School of Computing

Dublin City University, Ireland

Overview

- The Information Retrieval Process
- Cross-Language Information Retrieval (CLIR)
- Frameworks and Translation Strategies for CLIR
- CLIR and Distributional Semantics
- Concluding Remarks

The Information Retrieval Process

- Information retrieval systems seek to address a user's *information need*.
- An information need arises from an *anomalous state of knowledge (ASK)*.
- Process of resolving an ASK is a cognitive process on the part of the user.
- Conceptually trying to resolve an ASK makes IR can be hard, since the user must create a search query to look for information that you don't know.

  This can be stated more formally as the idea that there is a "non-specifiability of need" problem.

The Information Retrieval Process

- The ASK means that the user may not know or use language correctly to form a search query which properly describes their information need.
- This presents a fundamental challenge to IR systems.
- If users are not able to accurately describe their information need; then how can an IR system which is designed to return documents which match the query, reliably resolve an ASK?
- Fortunately, people can recognise relevant and useful information more easily than describe it.
  - IR systems return items matching the query, and rely on the user identifying useful items among the non-relevant ones.

Cross-Language Information Retrieval (CLIR)

- In general, people are able to read a language better than write it.
- Since user's have difficulty expressing their information need in their native language, this problem is likely to worse in another language with which they are less familiar.
- A CLIR system seeks to enable the user to express their information need in a language with which they are familiar, to retrieve documents in language with which they are less familiar.
- So, we need some form of automated translation.

Frameworks and Translation Strategies for CLIR

- Translate the queries - generally short
- Translate the documents - traditionally longer, but consider social media

Early research focused on three broad query translation strategies:

- Bi/Multilingual Dictionaries
- Aligned Corpora
- Machine Translation
  - Common knowledge: it doesn't work for CLIR!

Document translation using machine translation

Bi/Multilingual Dictionaries

- Replace each word in query with all possible translations from the dictionary.
  - Many will not be the sense of the word as used in the query.
  - Many translation errors introduced into the query.
  - Some error words may have significant IR impact, e.g. high idf weights.
- Query has good coverage (helps recall), but errors (impacts precision).
- Work on techniques for effective IE with these queries, e.g.
  - Use only first entry in dictionary.
  - "Pirkola's method" - union sum of n values of all translations for a word, and use single combined idf values for all translations.

Machine Translation

English - Japanese CLIR for news collection.

Query translation - from dictionary to machine translation:

- DICT: All dictionary entries
- POS: Matching Part-of-Speech
- DEF: Default - most popular transation
- SYN, Synonyms - machine translation without final lexical choice
- FMT: Full machine translation

MAP: $DICT < POS, DEF, SYN < FMT$ FMT wins!

(Jones et al, SIGIR 1999)

Multilingual IR

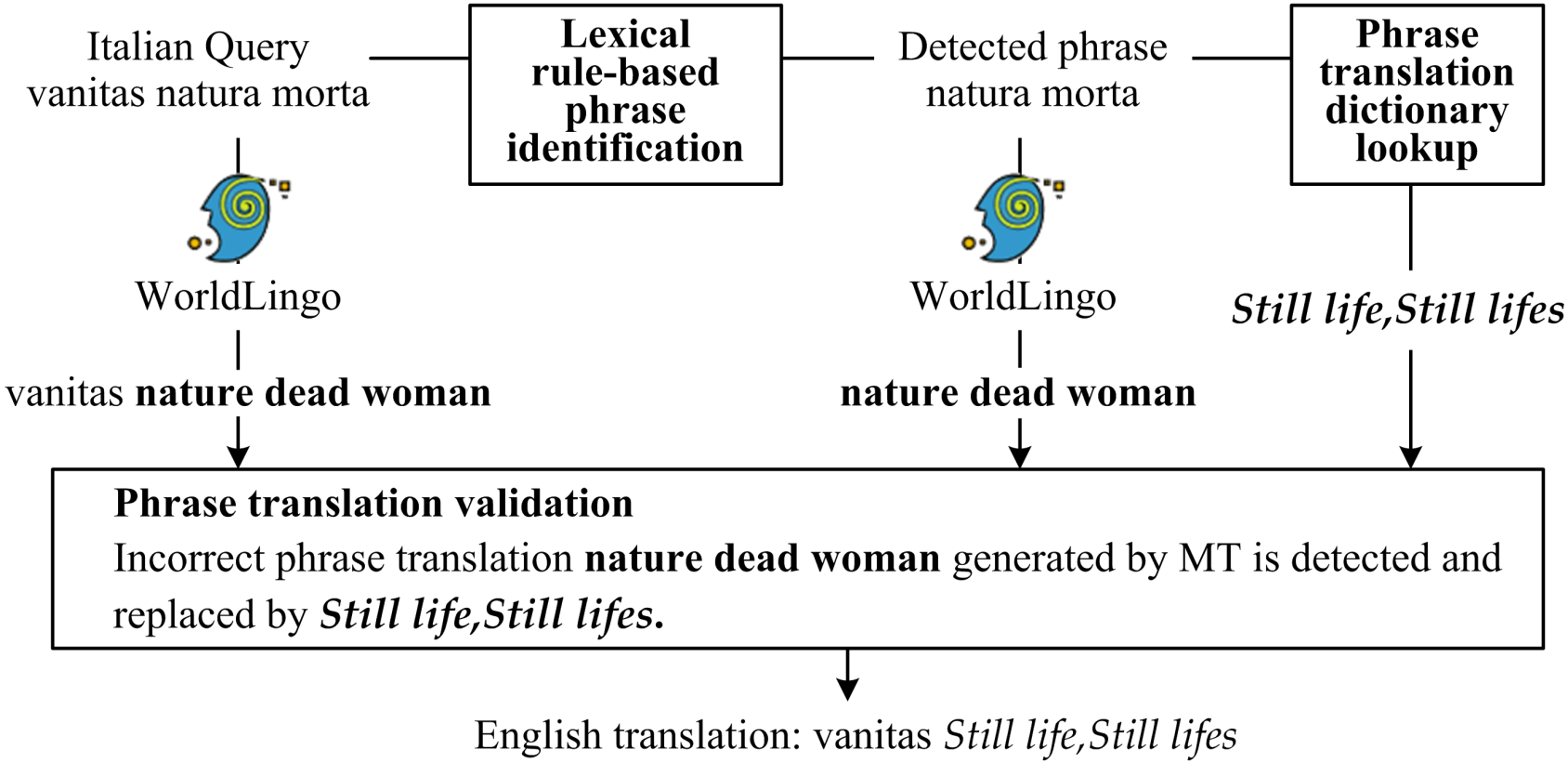
- Query in one language to retrieve documents from multiple other languages.
- Query translation - generate a separate ranked list for each language and merge to create final list (“data fusion”)
- Document translation - single search index in query language, single retrieval listed created, no merging necessary.
- Document translation generally more effective. (CLEF 2004)

Domain-Specific Translation

- Early work focused on CLIR for archives of newspaper articles.
  - This is basically an easy retrieval task.
- General purpose translation resources are effective for these tasks.
- but are less effective for specialist domains, e.g. medicine.
  - vocabulary outside the coverage of general resources 2
  - domain-specific translation likelihoods
- Use of domain-specific translation methods more effective.

(Rogati & Yang, SIGIR 2004)

Domain-Specific Translation: Cultural Heritage



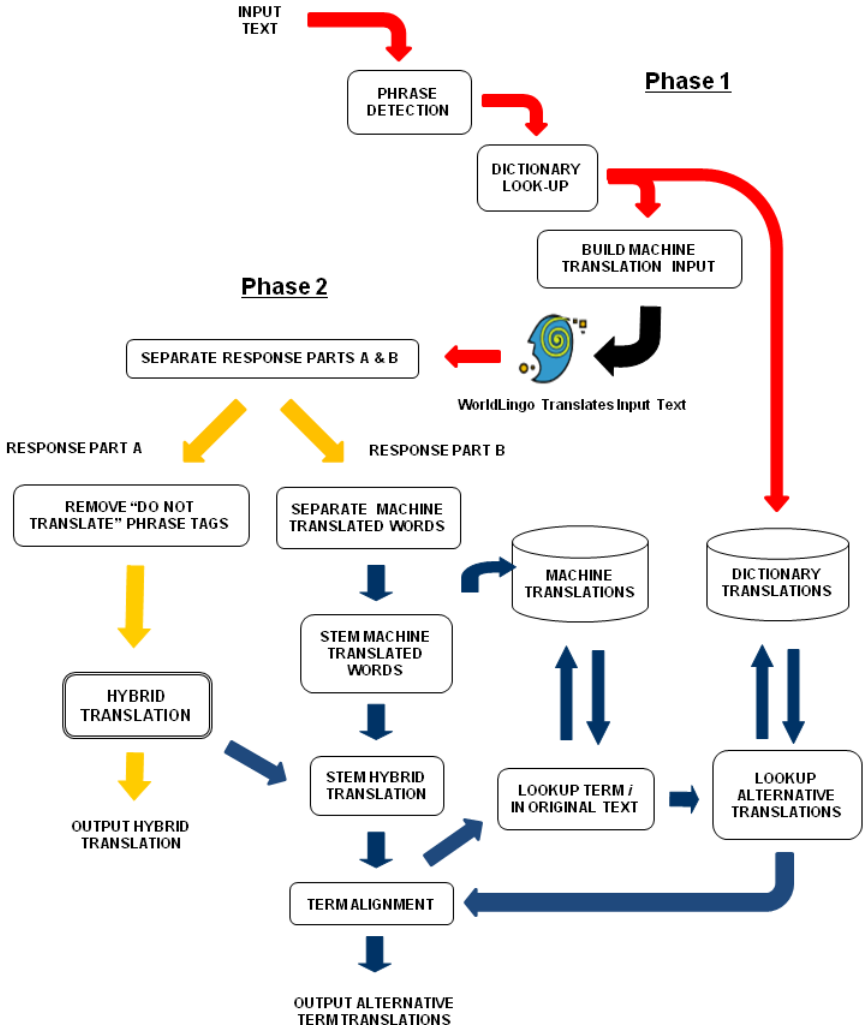
Example of Italian–English hybrid translation of a search query. from *MultiMatch* (Jones et al, CL-IA 2008)

Domain-Specific Translation: Cultural Heritage

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

Query translation examples.

Interactive Hybrid Translation Process: Cultural Heritage



(Jones et al, 2010)

Closer Integration of Information Retrieval and Machine Translation

- Detailed study of combination of the components of IR and statistical machine translation (SMT) for CLIR.
- Well designed integration of elements of IR and SMT improved CLIR effectiveness.

(Ture, PhD, University of Maryland, 2013)

Contemporary Online Machine Translation

But surely, state of the art MT systems do not have these problems!

- Compared CLIR effectiveness for standard online Google translate and Bing translation systems.
- Standard news CLIR task from CLEF 2000.
  - English language news collection.
  - Queries in: English, Italian, Spanish, German, Finnish, Dutch, Swedish.
- On average:
  - Google translate significantly better for Spanish
  - Bing translate significantly better for Swedish
  - No significant difference for other language pairs.

(Vahid et al., MWA @ WWW 2015)

Contemporary Online Machine Translation

Original Spanish Query	Google Translation	Bing Translation
Drogas en Holanda	**Holland** Drug	Drugs in the **Netherlands**
Ingreso en la Unión Europea	**Join** in the European Union	**Income** in the European Union
Bajas entre bomberos	Casualties among **firefighters**	**Fire** casualties
Uso de la energía eólica	Use of wind **energy**	Use of wind **power**
Premio Nobel de Economía	Nobel Prize in **Economics**	Nobel Prize **winner**
Turismo en E.E.U.U.	Tourism **E.E.U.U.**	Tourism in **USA**
Accidentes de aviones en pista	**Aircraft** accidents on track	**Planes** in track crashes

Examples of Spanish-English CLEF query *title* field translations.

CLIR and Distributional Semantics

- Distributional semantics in the form of embedding models can provide semantic interpretations of words in semantic space.
- Significant improvements in performance in MT.
- Much interest in neural-IR methods incorporating word embedding methods.
- How might embedding models be used in CLIR?
  - Very interesting initial study explores monolingual and crosslingual word embedding for Dutch-English CLIR.

(Vulić & Moens, SIGIR 2015)

CLIR and Distributional Semantics

- MT will (always) make translation errors which will impact on IR effectiveness.
- NeuralMT methods require large amounts of training data which is not available in all CLIR settings.
- Bi/Multi-lingual dictionaries give coverage of possible translations, but also introduce errors.
  - Alternative translation error problem likely to be lower for domain-specific concepts.

CLIR and Distributional Semantics

- Can we use semantic models to make more effective use of dictionary translation to maintain coverage,
- but while reducing translation errors;
- and to give effective CLIR where it is not possible to build a domain-specific NMT system?
- Can such translation methods be elegantly integrated with Neural IR methods for CLIR?

Concluding Remarks

- Overview of the history of research in CLIR.
- MT currently the dominant technology for CLIR.
  - Does not provide a secure or fully reliably CLIR solution.
  - Not always possible to develop MT system for all language pairs and/or domains.

    NIST OpenCLIR
- Neural methods effective for MT and IR. What is the potential for them to be used in CLIR?