Query Translation for Cross-lingual Search in the Academic Search Engine PubPsych

Cristina España-Bonet[1], Juliane Stiller[2], Roland Ramthun[3], Josef van Genabith[1] and Vivien Petras[2]

[1]Universität des Saarlandes & DFKI, Saarbrücken, Germany

[2]Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Germany

[3]Leibniz Institute for Psychology Information (ZPID), Trier, Germany

leibniz-psychology.org

UNIVERSITÄT DES SAARLANDES

Table of contents

Motivation: CluBS - Cross-lingual Bibliographic Search

Motivation of the CLuBS project

- academic discourse happens in several languages
- language barrier ⇒ researchers might not understand or even find relevant articles
- basing research solely on results published in dominant languages such as English bears the risk of drawing conclusions on sub-populations [3]

Goals of the CLuBS project

- aims to improve multilingual access to academic bibliographic information
- develops, implements & evaluates four different approaches to enable Cross-lingual Information Retrieval (CLIR)
- prototypical domain: psychology with PubPsych search engine (https://pubpsych.eu)
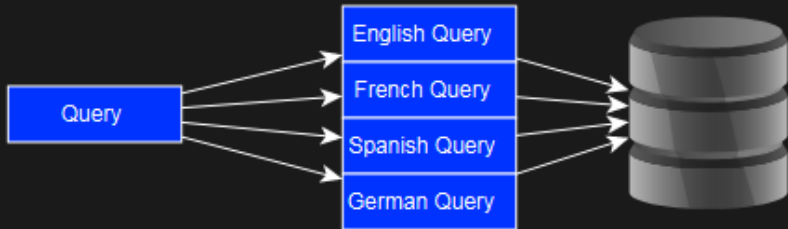
PubPsych

- database of psychological literature, treatments, test and research data
- aggregates bibliographic metadata from nine databases produced by several international partners, e.g. MEDLINE, PSYNDEX and NORART
- metadata mainly in English, German, French and Spanish
  - very uneven distribution due to different indexing practices
  - 20% of all content has no English metadata; 5% of content retrievable with Spanish
- ⇒ results on a topic vary depending on query language
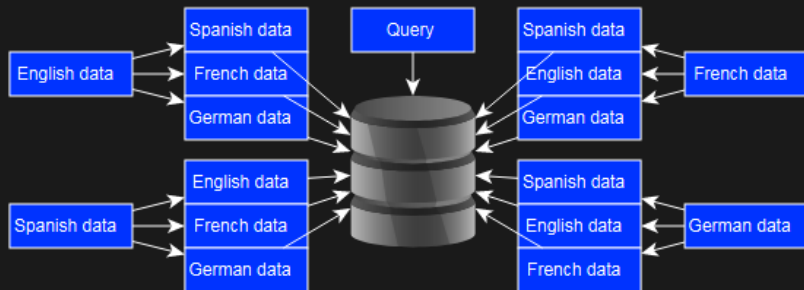
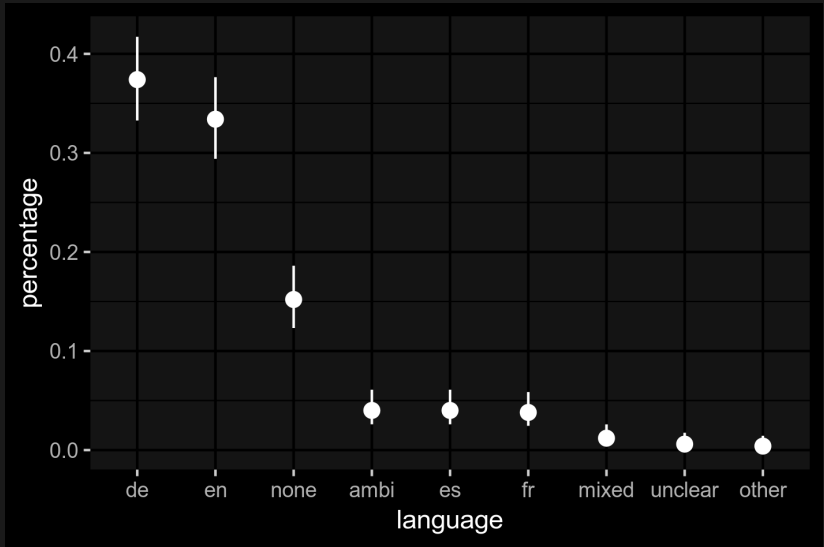- translation of content (metadata in this case)

- translation of queries
- research question: can PubPsych queries be translated into the four target languages by mapping them to purpose-built lexical resources?
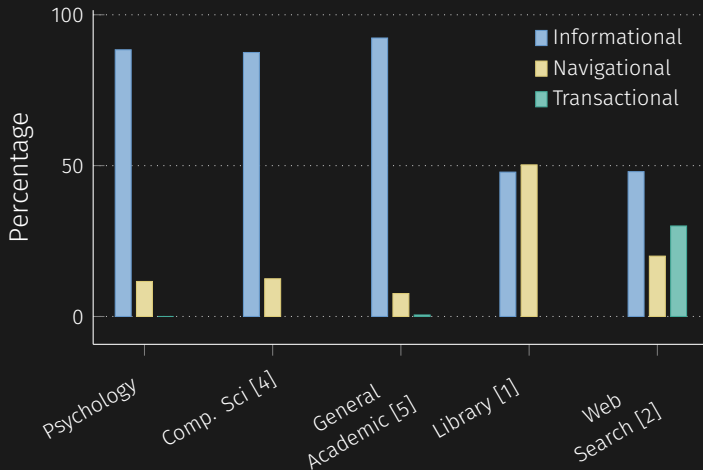
Queries & Languages in PubPsych

Query types and distribution in different domains

88.4% of queries in PubPsych are informational, so we chose an approach of mapping lexical resources to translate them
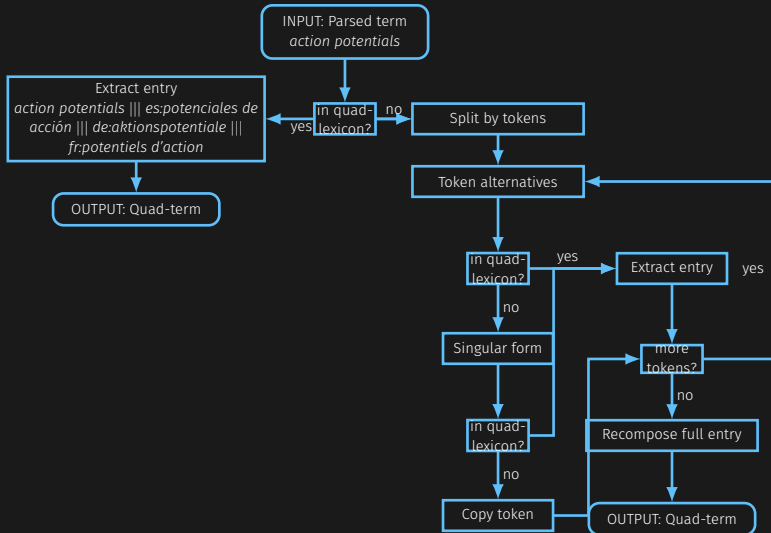
Query Translation

Creation of quadrilingual lexicon

QuadLex: aligned dictionary in English, French, German and Spanish from four sources

	German	English	French	Spanish
MeSH	70,694	175,004	96,333	66,828
WP (titles/categories)		(81,369/38,038)		
Apertium	7,792	5,935	6,020	5,846
Manual	4,262	4,142	4,047	4,081
Total unique (Lex)	202,128	304,277	225,607	195,937

Flowchart for query term translation

Evaluation

How many terms and tokens out of the 536,479 queries can we translate?

- whole terms with MeSH lexicon: **7.7%**
- whole terms with Quadrilingual Lexicon: **14.9 %**
- token level with MeSH lexicon: **64.2%**
- token level with Quadrilingual Lexicon: **85.0 %**

Corpora of 500 queries manually rated by **3 annotators**:

- 100 queries in each language (en, de, fr, es)
- 100 queries without a definite language identification (many named entities)

Evaluation according to adequacy:

- no gold translation existed
- How much of the meaning in the source query was expressed in the translation?
- use of a three-point scale:

  0 none of the meaning was transferred
  1 part of the meaning was transferred
  2 all meaning was transferred

Inter-Annotator Agreement

Fleiss' kappa of the three raters for different language pairs:

source	2*de*	2*en*	2*fr*	2*es*
de	n/a	0.616	0.658	0.598
en	0.442	n/a	0.455	0.521
fr	0.243	0.268	n/a	0.384
es	0.422	0.354	0.472	n/a
none	0.494	0.458	0.513	0.440

disagreement example

- source: "unfinished task" -> DE: "unfinished aufgabe"

Results and Future Work

Results

- average adequacy of 1.4 ➜ most of the queries had at least some of their terms properly translated
- 54%±20% of the queries had the maximum adequacy score when looking at the mean over languages,
- only 14%±8% of the queries got completely incorrect translations;
- remaining 33%±15% were partially well translated.

Results

Languages were quite similar with two clear exceptions:

- translation of German queries had a lower quality (mean adequacy 1.1) mainly because the compound nature of German increases the number of untranslated tokens with respect to the other languages, and
- queries with undetermined language had a very high adequacy (1.8) because they are shorter and, in most of the cases, leaving the source token untranslated resulted in a good translation.

Future work

- increase coverage through multilingual word embeddings
- improve approach by removing systematic errors
- improve translation for German (due to its compound nature)
- next to translation quality we will evaluate the impact on retrieval

Questions?

Cristina España i Bonet <cristinae@dfki.de>
Juliane Stiller <juliane.stiller@ibi.hu-berlin.de>
Roland Ramthun <rr@leibniz-psychology.org>
Josef van Genabith <josef.van_genabith@dfki.de>
Vivien Petras <vivien.petras@ibi.hu-berlin.de>

`https://www.clubs-project.eu/en/`

📑 C. Behnert.
Evaluation Methods within the LibRank Project.
Working Paper, LibRank, 2016.

📑 A. Broder.
A taxonomy of web search.
In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

📑 J. Henrich, S. J. Heine, and A. Norenzayan.
Most people are not WEIRD.
Nature, 466(7302):29–29, jul 2010.

📑 M. Khabsa, Z. Wu, and C. L. Giles.
Towards better understanding of academic search.
In *Joint Conference on Digital Library (JCDL)*, pages 111–114. ACM,
2016.

📄 X. Li, B. J. Schijvenaars, and M. de Rijke.
Investigating queries and search failures in academic search.
Information Processing and Management, 53(3):666 – 683, 2017.