D1.5 – Mapping Approaches for Vocabularies and Queries in PubPsych

Cristina España-Bonet
Universität des Saarlandes

– v1.1 –
March 2017

Abstract

This document describes the in-domain vocabularies the project has available, and how their multilingual counterparts are built from them. It also proposes several public resources we can use to complement and extend this data with general-domain vocabulary. Finally, application of the multilingual lexicons to controlled terms and query translations is sketched.

Contents

1 Introduction

As outlined in the description of work for the project, we want to approach our cross-language information retrieval task by using machine translation, either by translating the full content of the digital repository into all the other languages and/or by translating only the query. In both cases, multilingual lexicons are required. They are especially relevant for two of the four approaches envisaged:

1. *Query translation approach*: When a search expression is entered, the portal extends the query by injecting translations of the terms or phrases found in it and runs the modified version against the search index. The quality of the translation might be relevant for the accuracy of the system, although this approach is expected to improve the recall.

2. *Knowledge-based approach*: A multilingual thesaurus provides the controlled terms that are used in the metadata records and maps them to the target languages. A search using a multilingual thesaurus requires the searchers' terms to be mapped to the thesaurus. However, the terms used to construct a query can remain in the searchers' own language. Since PubPsych records utilize more than one controlled vocabulary, a mapping of index terms (concordances between controlled terms) is also necessary. This approach should improve retrieval results because the index terms (using the appropriate scientific language) from the documents are used.

Both approaches, of course, are related because using the multilingual thesaurus for the queries allows their translation. Next, Section 2 describes the available resources and the procedure considered to build an English-French-German-Spanish quad-lexicon with both in-domain and general-domain entries. Later, Section 3 explains how to combine this quad-lexicon with machine translation tables and their use for query translation.

2 Multilingual Vocabularies

The construction of our multilingual vocabulary will be based on the Medical Subject Headings Thesaurus (MeSH), because it is easily available in all for languages relevant for the project. Also important for our project is the Thesaurus of Psychological Index Terms of the American Psychological Association (APA Thesaurus). The PubPsych database segments PSYNDEX, ERIC, PASCAL and ISOC-Psicología use controlled terminology to tag the content of their articles. The APA Thesaurus is similar to the term sets of the MeSH and of ERIC. PSYNDEX Terms is the authorized German translation of the APA Thesaurus. The PASCAL thesaurus is trilingual (French, English, Spanish) and has MeSH (English) integrated. ISOC-Psicología is a thesaurus in Spanish with a similar structure to APA, so a mapping could be obtained. Notice that only PSYNDEX/APA Terms and the Tesauro ISOC de Psicología are domain specific.

The alignment of these resources will allow to build a quad-lexicon in English-French-German-Spanish. Given the intersection of APA, MeSH and PSYNDEX thesauri, we expect to align different languages by pivoting into English and using the IDs to relate all the resources or, at least, the subset of common elements. A quad-lexicon built in this way is expected to cover the translation of controlled terms used in PubPsych well. However, it will lack out-of-domain vocabulary important for query translation. In order to alleviate this problem, we will enlarge our quad-lexicon with other external resources and build and extended vocabulary with:

- Wikipedia titles and category names
- Combination of public bilingual lexicons
- In-domain translation tables from in-house SMT systems

3 Mapping Vocabularies to Queries

Queries in PubPsych are neither guided by a strict grammar nor limited to a controlled vocabulary but allow the user to search with free text. Therefore, translation with only a basic in-domain quad-lexicon such as that extracted from MeSH is not enough for the purpose.

When mapping vocabularies to queries, we will not confine ourselves to technical terms but also attempt to cover frequently used query terms. For this reason, we described in the previous section additional resources to build an extended lexicon which additionally will be used for the translation of the PubPsych's controlled terms. In addition, for query translation, we plan to add translation tables coming from statistical machine translation systems (SMT) so that we can check the best approach given that we use:

- Extended quad-lexicon
- SMT translation tables
- Extended quad-lexicon + SMT translation tables

An SMT translation table maps sequences of words (*phrases*) from one language into another and associates a score to the pair:

```
...
headache ||| Bestimmung ||| 9.40918e-05 0.0018484 0.000202066 0.0028329
headache ||| Kopfschmerz ||| 0.59375 0.438596 0.103825 0.0708215
headache ||| Kopfschmerzen , ||| 0.00168082 0.446541 0.000202066 0.0118873
headache ||| Kopfschmerzen an ||| 0.0369781 0.446541 0.000202066 0.000583065
headache ||| Kopfschmerzen ||| 0.411348 0.446541 0.31694 0.201133
headache ||| Kopfschmerzes ||| 0.225531 0.384615 0.00492965 0.0141643
headache ||| Mischungen ||| 0.00462226 0.0714286 0.000202066 0.0028329
...
```

These tables are bilingual and pivot techniques will be needed to build the multilingual phrase table. The quality of such approaches for obtaining the quad-lingual mappings remains to be checked.

The vocabulary of a phrase table depends on the corpus used to train the SMT system. For the query translation task, we will use the parallel corpus extracted from the PubPsych database as described in D1.2. This way, the table will contain entries related to psychology but also common vocabulary present in text of any origin.

Finally, we will need to take into account the difficulty to detect the language of a query. In order to partially solve this issue, we will build the quad-lexicon in a (*key*, *value*) format. The *key* stores all the vocabulary of the four languages without specifying the language, whereas the *value* stores the translation into the other three languages, this time with the language information as seen in the following example:

```
(Cefalgia, en:Headache|||de:Kopfschmerz|||fr:Céphalée)
(Céphalée, en:Headache|||es:Cefalea|||de:Kopfschmerz)
(Headache, es:Cefalea|||de:Kopfschmerz|||fr:Céphalée)
(Kopfschmerz, en:Headache|||es:Cefalea|||fr:Céphalée)
```

This way, we expect to detect the language simply as the one that matches our lexicon. In case of ambiguity, the languages will be chosen according to the most frequent language used in PubPsych or by using the language of the other terms in the query when possible.

Assuming the existence of the quad-lexicon in the (*key*, *value*) format, we will implement the most general query translation engine as follows:

1. Let Solr/Lucene parse the full query and extract the terms.
2. Try to match each term against the quad-lexicon, then against SMT tables. If a term matches, this term is translated. If not:
   2.1. Split query into decreasingly small parts (chunks of tokens).
   2.2. For each chunk of length n first try to match it against the quad-lexicon, then against SMT chunks of size n. If we match either the quad-lexicon or a SMT phrase, we translate the chunk. If not, we repeat the process with chunks of length $n - 1$.

This methodology has two main advantages: (*i*) there is no need for language identification and (*ii*) it is very fast because it is just a look-up in a table, meaning it can be used online.

4 Conclusions

We have introduced the main methodology envisaged to deal with translating controlled terms and queries within PubPsych. In both tasks, we rely on the construction of a quad-lingual lexicon in English-French-German-Spanish with both in-domain and general domain vocabulary.

Notice that the real-world quality of the multilingual vocabularies built either coming from thesauri or from translation tables can require changes to the final strategy for translation of controlled terms and queries.