M1.3.1 Evaluation plan for CLUBS project

Juliane Stiller & Vivien Petras
Humboldt-Universität zu Berlin

– v2.0 –
July 2017

Abstract

This document describes the different evaluation studies which will be executed during the course of the project. The studies assess the performance of different MT approaches for cross-lingual retrieval in the bibliographic search engine PubPsych.

Contents

1 Introduction

The goal of the project is to determine which of the following approaches is best for cross-lingual information retrieval (CLIR) of metadata records in digital libraries:

- Record translation
- Query translation
- Mapping of controlled vocabularies in different languages
- English as a pivot language (for record and query translation)
- Merged approaches from the ones above

1.1 Intrinsic and Extrinsic Evaluation

We distinguish between intrinsic and extrinsic evaluation and follow the concepts of Dorr et al. [**?**]. For the intrinsic evaluation, we test the quality of the automatic translation or mapping approaches themselves. In the extrinsic evaluation, we determine the impact of the approach on retrieval or other tasks related to the access of documents.

Merging the results of these two types of evaluation will provide an answer to the question, which CLIR solution is the best for metadata portals. Figure **??** shows evaluations for the different solutions which enable cross-lingual retrieval of metadata records.

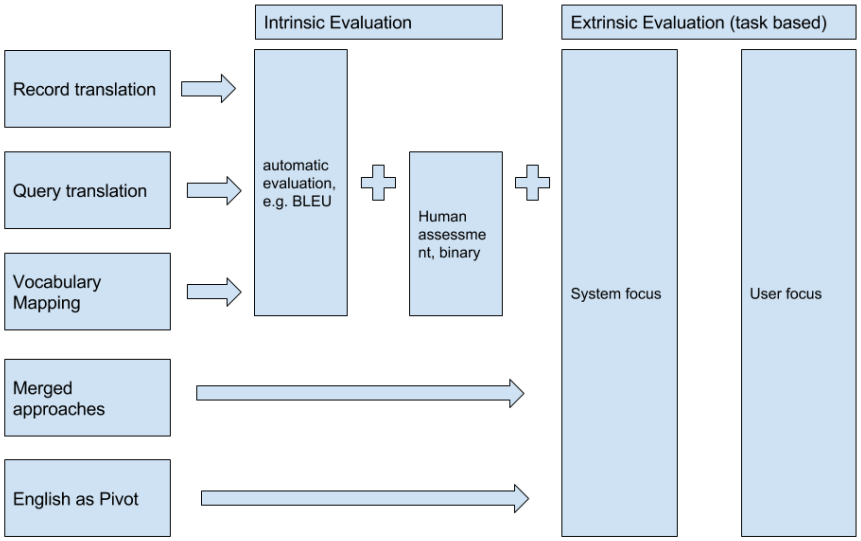

Figure 1: Evaluation plan for the different solutions tested in the project.

First, the intrinsic value of each approach (there might be several approaches to a solution, e.g. different technical options for record translation) will be tested. For example, does the automatic record translation with STM produce appropriate (i.e. correct) translations? This can be assessed automatically or be judged by a human. The next step is the extrinsic evaluation. The extrinsic evaluation consists of different methods, which are either system- or user-focused.

In the system-focused extrinsic evaluation, we will look at the impact of the different translation approaches on retrieval performance, e.g. differences in result numbers and in the document sets retrieved. The system-focused evaluation will be performed using the set-up in Figure **??**. Either the queries or the documents in the information retrieval system (here Solr) will change for the different translation approaches as described in the succeeding sections. Result lists from these evaluations will be compared to a baseline of result lists retrieved with untranslated queries and / or documents.

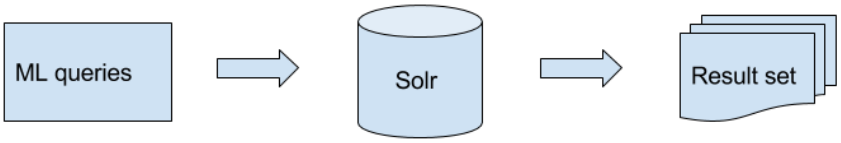

Figure 2: Set-up for the extrinsic evaluation approaches.

The user-focused approach will only be applied to the best technical approach for each solution.It tries to incorporate the perspective of the users. This usually requires a judgment regarding the relevance of the retrieved documents. Additional user-focused evaluation could also take GUI changes with regard to multilingual features into account, shifting the evaluation approaches into user experience evaluations.

The intrinsic evaluations will be done by DFKI/SU. Test corpora will be provided by Humboldt and ZPID.

1.2 The Process of Evaluation

For each of the technical approaches to translate different components of the metadata, we will perform the following evaluations:

- Intrinsic evaluations determining measures such as BLEU and other automatic measures.
- Additionally, extrinsic evaluation of the approaches will be performed (the different runs are detailed in the later sections).
- Analyzing results from the intrinsic and extrinsic evaluation, we will determine winning approach for each solutions (numbers 1-5).
- These approaches (one per solution) will then be compared against each other to determine the best solutions for bibliographic cross-lingual search.
- For the comparison, we will determine the relevance of retrieved results for each solution.

  Table **??** list the different approaches for each of the five solutions and details the process of evaluation.

1.3 Producing the Baselines for Extrinsic System-focused Evaluations

The basic set-up is shown in figure **??**. In its essence, it will be used to produce the different baseline for the extrinsic system-focused evaluations. Since we want to exclude dynamic factors of collection or system updates (i.e. new relevant documents or new search features), which are independent from the translation, we will perform the evaluations on

Solution	Nr.	Method / Approach	Intrinsic	extrinsic system-focused	relevance /extrinsic user-focused
Query translation	1a	CV mapping	X	X	Winner method
	1b	CV mapping + MT aligned chunks	X	X	
Abstract translation	2a	SMT / NMT	X	X	X
Knowledge-based solution	3a	CV mapping	X	X	Winner method
	3b	CV mapping + MT aligned chunks	X	X	
English as Pivot	4a	1a + 2	X	X	Winner method
	4b	1a + 3a	X	X	
	4c	1a + 3b	X	X	
	4d	1b + 2	X	X	
	4e	1b + 3a	X	X	
	4g	1b + 3b	X	X	
Merging	5a	2 + 3a	X	X	Winner method
	5b	2 + 3b	X	X	

Table 1: Approaches for each solution and the evaluations.

stable (frozen) Solr instances with a stable collection of documents. A set of 50 queries manually translated into the four languages will be sent to the frozen Solr instance of PubPsych (see figure **??**). The retrieved result lists will serve as a baseline for monolingual retrieval results.

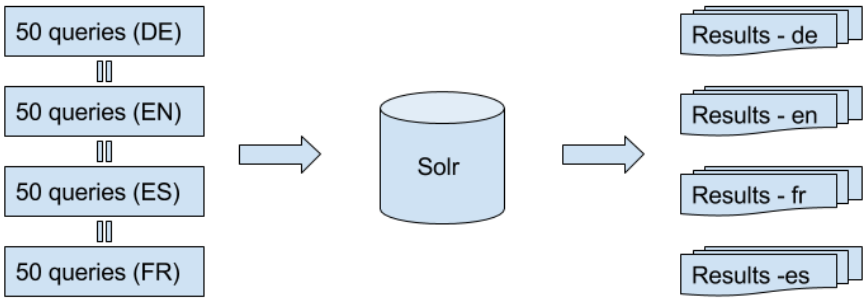

Figure 3: Producing the baseline result sets.

For each of the 200 queries, which are send to the frozen Solr version, the following information is saved:

- the results list
- the number of documents retrieved
- the ranked list of documents from the first result page

From the four different result lists of the aligned (i.e. parallel) queries in English, French, German and Spanish, merged result lists can be created that contain all results of the four result lists. After removing the duplicates among the merged list, it can serve again as a baseline for assessing relevance in multilingual retrieval.

2 Abstract Translation

A record in PubPsych can have the fields aggregated in table **??** with the respective candidates for MT. The fields *key phrase* and *additional descriptor* might contain relevant text for translation but are only provided by certain data sources or for a small fraction of records. Therefore, we decided to not include them in the machine translation.

Solution	Intrinsic Evaluation		Extrinsic Evaluation	
	Automatic	Human Judgement	System-focused	User-focused
Query translation	250 queries in 4 languages	N/A	50 queries in 4 languages	N/A
Abstract translation	800 records in 4 languages	N/A	50 queries in 4 languages	N/A
Knowledge-based	tbd	tbd	tbd	tbd
English as Pivot	same as for record and query translation	N/A	250 queries in English	N/A
Merging				

Table 2: Corpora used for the different experiments.

2.1 Intrinsic Evaluation with Automatic Measurements

Here, we use automatic measures, such as BLEU [**?**], to determine the quality of translation. Humboldt and ZPID provide an aligned corpus for this with **800** records that is manually translated from English into French, German, Spanish.

2.2 Intrinsic Evaluation with Human Judgment

Note: As the record translations are the means to improve cross-lingual retrieval, we will not evaluate translation with human judgements. If an intrinsic evaluation is required, it could look like this: Annotators evaluate fluency and adequacy of the MT record. Fluency measures the translation based on its grammatical correctness, lack of spelling errors and natural language. It should be assessed by native speaker which can be monolingual. For the evaluation, a Likert-scale is often used [**?**, Ch. 4].

Accuracy measures how well the original meaning is transported in the translation. For this, bilingual annotators are needed. For both assessments, the inter-annotator agreement should be calculated. Two states of MT output could be compared ranking the different translations [**?**].

2.3 System-focused Extrinsic Evaluation

Here, we will evaluate the record translation approaches using volume and result set differences in PubPsych by comparing the baseline result sets from the frozen Solr instance (see section **??**) with the result set retrieved with record translation. The query evaluation corpus consists of **50** queries aligned in English, French, German and Spanish.

For the record translation evaluation, queries are fixed over the course of the project. The change is the translation of the record.

Methodology:

- Each query (out of 50) is sent against the (untranslated) record collection and result sets are retrieved. These four result sets from the four languages are saved and used as the baseline resulting in 200 result lists.
- After the translation, the experiment is repeated with the same queries sent against the corpora translated with different methods. (see Fig. **??**). The evaluation com-

Field name	Translation Candidate	Comments
Title	X	
Subtitle	X	
Title translation		
Author		
Affiliation		
Country		
Source		
Year		
Journal title		
Media Type		
ISBN		
ISSN		
Thesis title	X	Only present in source Psyndex
Language		language of publication, should be translated but not with MT
Abstract	X	
Additional abstract	X	
Classification	X	
Keyword/Controlled term	X	from controlled vocabularies
Additional Descriptor		
Age group		
Origin of Population		
Key phrase		assigned freely by the cataloguer
Controlled Method		
Document type		
Level of Evidence		
Segment		

Table 3: Fields of a PubPsych record and the candidate fields for translation.

pares the new result sets to the baseline result set. Each query should ideally retrieve more and different (better) documents than the baseline.

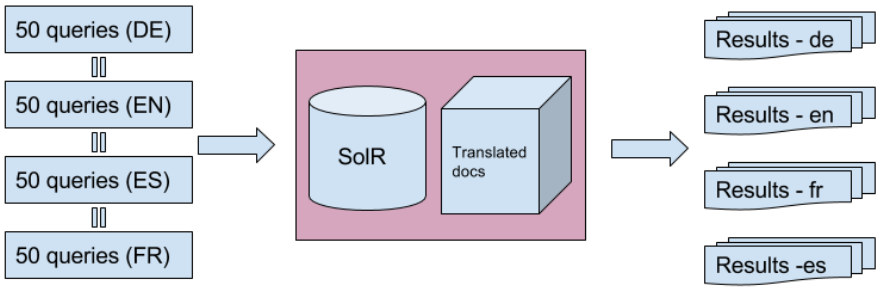

Figure 4: Retrieving result sets with translated records.

Requirements:

- Aligned query corpus of 50 queries in English, German, French, Spanish

- Frozen Solr to produce baseline and new result sets
- Translated PubPsych records (same as in frozen baseline corpus)

Evaluation Steps:
We first compare result sets of same language queries using the non-translated records vs the translated records. For example, the result set of a query in English (without record translation) is compared to the result set of a query in English (with record translation).

- Compare results sets retrieved with the same language query, we look at the size of the new result set (with record translation) compared to the baseline result set (without record translation).
- Compare retrieved results to baseline, e.g. differences in found data sources.
- Compare ranking of the first x results.

In a second step, we compare the merged results lists of each of the four aligned queries and compare it to the results list retrived with record translation. Here, the following measures are possible:

- Compare the size of the new result set (with record translation) compared to the baseline result set (without record translation) retrieved from merging the results of four aligned queries.
- Compare retrieved results to baseline, e.g. differences in found data sources.
- Compare ranking of the first x results.

Example:
Table **??** shows an example query "youth unemployment" aligned in the relevant languages. The result sets retrieved with no record translation can be compared to the result sets for each query when the records are translated (comparing the values in each column = 4 comparisons). Additionally, each result set retrieved with record translation can be compared to the merged list retrieved with no record translation (4 comparisons).

	"youth unemployment"	Jugendarbeitslosigkeit	"chômage des jeunes"	"desempleo juvenil"	merged list
no record trans.	106	126	3	4	181
with record transl.	183	180	185	185	

Table 4: Example of an aligned query and baseline number of retrieved documents.

2.4 User-focused Extrinsic Evaluation

The user-focused extrinsic evaluation will be used to determine which of the translation solution (as shown in table **??** is the best for bibliographic metadata. This evaluation will assess the impact of the translation on retrieval effectiveness determining the best approach.

Note: For all user-focused evaluations, we will leverage synergies by merging results list of the runs and removing duplicates. These evaluations are very costly and through thoughtful planning, we should minimize the effort here. One option is to decide for one user test which covers several scenarios. A more detailed workflow can be found under section **??**.

3 Query Translation

3.1 Intrinsic Evaluation with Automatic Measurement

Here, we use automatic measures, such as BLEU, to determine the quality of translation. Humboldt and ZPID provides an aligned corpus for this with **250** queries translated in English, French, German and Spanish.

3.2 Intrinsic Evaluation with Human Judgement

This could be a binary judgement: correct or incorrect. This could be also independent of an evaluation corpus, just sample some random queries and decide if the MT is correct or not. The feasibility of this approach depends on the complexity of the queries. Similar to the record translation, the intrinsic human evaluation is accompanied with increased effort. As the main concern is the impact on CLIR, we decided to forego this experiment.

3.3 System-focused Extrinsic Evaluation

Figure **??** shows how the process could be done: the (frozen) Solr with the untranslated records is used and only the queries are translated. Each query is translated via MT into the respective other three languages (see table **??**, the cells marked in green) resulting in 12 queries used for retrieval.

Three different evaluations will be performed:

- **Improvements of multilingual retrieval vs monolingual retrieval.** The results set retrieved in a monolingual search are compared to the one with query translations. For example, the result set of an English query ("youth unemployment") is compared to the one which were retrieved with the English query and the translation equivalents in German, French and Spanish ("youth unemployment" OR "Jugendarbeitslosigkeit" OR "chômage des jeunes" OR "desempleo juvenil"). Next to the baseline runs, 4 runs with the multilingual queries are needed.

- **Indirect assessment of performance of query translation.** In table **??**, each row contains 3 translations for queries (translation from the aligned corpus). The translated queries can be compared to the manually translated queries. If queries are not identical, they are both sent to the Solr instance and results sets are compared. If the queries are the same translated by machine or by human, the results sets will be identical, too. Next to the baseline runs, 12 runs with each translated query are needed.

- The last step would be to compare the merged list of each quadruple of queries which were manually translated to the one which was translated with MT.

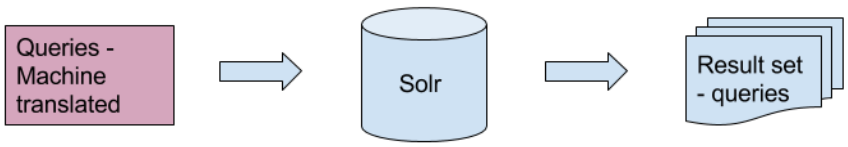

Figure 5: Evaluation approach for query translation.

Aligned queries	From EN into	From DE into	From FR into	From ES into
EN	-	EN	EN	EN
DE	DE	-	DE	DE
FR	FR	FR	-	FR
ES	ES	ES	ES	-

Table 5: MT translation of query sets into respective languages.

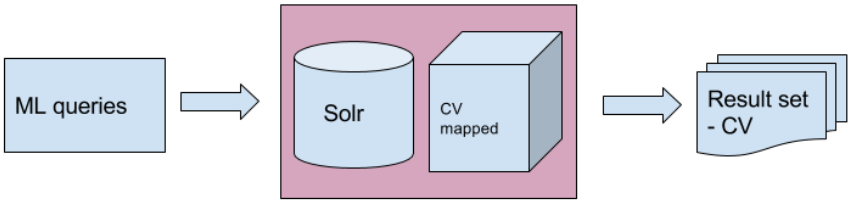

Figure 6: Evaluation approach for mapping of controlled vocabularies.

3.4 User-focused Extrinsic Evaluation

See Note under **??** and section **??**.

4 Vocabulary Mapping

For the vocabulary mapping approaches, we will align the controlled vocabularies ideally in the four target languages. The vocabularies have to be identified. Existing mappings can be used for training and testing.

4.1 Intrinsic Evaluation with Automatic Measurement

Could be a precision-based measure based on already available mapped vocabulary terms, i.e. how many of the terms in one vocabulary are mapped to another one?

4.2 Intrinsic Evaluation with Human Judgement

This could be a binary judgement of the created mappings of a subset of the controlled vocabularies.

4.3 System-focused Extrinsic Evaluation

The methodology is very similar to the one we use for record translation, except that only the CV terms are "translated", i.e. mapped to the other languages (see figure **??**).

4.4 User-focused Extrinsic Evaluation

Testing the ranking for this might be very interesting. Overlap between queries and CV is crucial here. Also, see Note under **??** and section **??**.

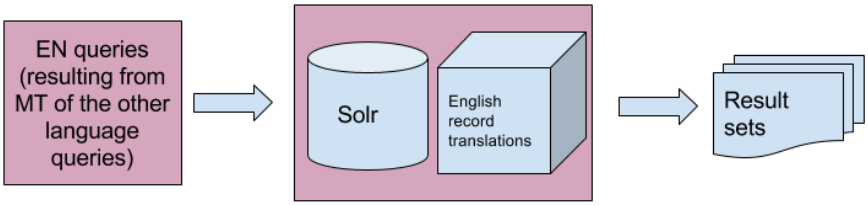

Figure 7: Evaluation approaches for English as a pivot.

5 English as a Pivot Language

5.1 Intrinsic Evaluation with Automatic Measurement

We can probably reuse the test we did for the record and query translations just using the English test sets.

5.2 Intrinsic Evaluation with Human Judgment

We can probably reuse the test we did for the record and query translations just using the English test sets.

5.3 System-focused Extrinsic Evaluation

We can probably reuse the test we did for the record and query translations just using the English test sets, see figure **??**.

5.4 User-focused Extrinsic Evaluation

See note under **??** and section **??**.

6 Merged Translation Approaches

As determined in the document on the different scenarios for translation approaches [1], we identified one merged approach. The translation of the abstract (abstract and title field) can be combined with the knowledge based approach where multilingual keywords from controlled vocabularies are mapped or translated.

The evaluation set-up for this approach is similar to the evaluation of its components - intrinsic evaluation with automatic judgment. For the user-focused extrinsic evaluation, we would look at different results sets. These could be merged to create synergies between the different evaluation steps.

7 Comparing the Different Approaches

So far, the different approaches were assessed comparing their performance to a system without any translation. We now need to define ways in which we compare the approaches against each other. One approach is to juxtapose the different evaluation results and rank the approaches according to their performance. Another option is to assess indirectly each solution with the retrieval performance.

[1] https://github.com/clubs-project/documentation/wiki/Scenarios

7.1 Proposal for Comparing Solutions

For each winning approach (5 in total) a relevance assessment will be executed, so the approaches can be compared against each other and against the Baseline (or default ranking). For each query (50 per language, 200 in total) the top10 results will be stored as well as the position of the documents. The result list from 6 set-ups will be saved and converted to one list (with a removal of duplicates) for each query quadruple. The resulting list will be assessed for relevance per language for each respective query. Measures will show if translation solutions improve the precision of results (top10) across languages. We will measure P@10. Relevance will be assessed on a three-point-scale.

What we need for this:

- Relevance assessors able to assess in our four languages.
- Topic descriptions for the 50 queries used in the evaluation need to be done (also translated into each language) + descriptions of different relevant degrees.