M1.3.2 CLUBS - Testing retrieval performance

Juliane Stiller & Vivien Petras
Humboldt-Universität zu Berlin

– v3.0 –
March 2018

Abstract

This document will detail the experiment that will be conducted to determine, which of the five approaches performs best with regard to retrieval performance.

Contents

1 Introduction

To identify which of the translation approaches work best with regard to the retrieval performance in PubPsych, we perform a retrieval performance test. The different translation approaches chosen to be tested are detailed in the Deliverable M1.3.1 CLUBS - Evaluation Plan and in table 1. This document will detail the approach for one part of the extrinsic user-focused evaluation: which translation approach produces better results for user queries? How is the retrieval performance influenced by the different approaches?

Solution	Nr.	Method / Approach	Intrinsic	extrinsic system-focused	relevance / extrinsic user-focused
Query translation (QT)	1a	CV mapping	X	X	Winner method
	1b	CV mapping + MT aligned chunks	X	X	
Abstract translation (TR1 + TR2)	2a	SMT / NMT	X	X	X
Knowledge-based solution (KA)	3a	CV mapping	X	X	Winner method
	3b	CV mapping + MT aligned chunks	X	X	
English as Pivot (EP)	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 (MA)	5a	2 + 3a	X	X	Winner method
	5b	2 + 3b	X	X	

Table 1: Approaches for each solution and the evaluations.

2 Topic creation

50 queries were chosen from the English query corpus based on the range of topics they represent as well as a representation of query categories.

Domain experts will determine the information needs for each query and describe them in textual form.

Example:

Query: Bullying and teacher Description: Documents on bullying in schools and other educational institution and teachers' reactions and management strategies, also documents on bullying committed by teachers.

A record is relevant, when it fulfills the information need represented by the original query and by the suggested information need description. For some measures, we will use binary relevance, for others, there will be a scale.

The information need descriptions will not be translated from English assuming that all assessors will know English sufficiently to understand the information need and be able to assess documents in accordance with the information need.

3 Ad-hoc retrieval task

The 50 queries and their translations (200 in total) will be sent to the Solr instance to retrieve results. For each query in each language, we store the top 10 documents (depth 10): for 200 queries that would be 2000 documents at the most. For each document, all the fields (or the equivalent information) that will be shown in the portal will be extracted and presented to the assessors. For each run, we store the top 10 list with relevant fields and the rank in the search result list.

The following runs will be set-up:

Queries	**Baseline**	**QT**	**TR**	**KA**	**EP**	**MA**
BL-DE	BL-1	QT-BL-1	TR1	KA-1	EP-1	MA-1
BL-EN	BL-2	QT-BL-2	TR2	KA-2	EP-2	MA-2
BL-FR	BL-3	QT-BL-3	TR3	KA-3	EP-3	MA-3
BL-ES	BL-4	QT-BL-4	TR4	KA-4	EP-4	MA-4

There are 24 runs. For each run, 50 result lists are produced with 10 documents - one for each query. If there is no overlap between the 28 result lists for each query (one list per 28 runs), we have to assess 14,000 documents. We estimate a big overlap between the result lists, so there will be less than 12,000 documents that need to be assessed.

4 Pooling for relevance assessment - pooling by language

We assume that the language of the source document is in general equal to the language of the title and the abstract. So based on the language field value, the documents are split by language. Documents in languages different than German, English, Spanish or French are ignored in the assessment.

5 Assessing results

For determining the relevance of a document for a given query, we need assessors with German, French or Spanish and additional English language skills (the descriptions for each query will be only in English). The assessors will get extensive relevance assessment guidelines. Due to budget constraints, each document will be only assessed by one judge. Although we could think about a double assessment of English documents to calculate inter-annotator agreement.

Results are assessed based on the description for each query. For each document, the following three-point scale for relevance applies:

1. Not relevant – the record does not fulfill the information need, the information is not relevant.

2. Partially relevant – the record partially fulfills the information need, but there are some doubts as to whether the whole information need is covered.

3. Highly relevant – the record as represented fulfills the information need and is highly relevant.

Problem: Different assessors assess relevance of documents coming from one query. This will contribute to inter-annotator differences within a query's relevance assessments. Due to the multilinguality of this task, we cannot provide an assessment for all documents of a query coming from a single judge.

6 Comparing results

For calculating metrics on retrieval performance, we will look into the following measures:

- R-precision: If the number of relevant documents is r<=10, we only look at the documents up to the r-th rank of the list. If the number of relevant documents is >10, then we calculate R-precision based on 10, which would result in R-precision of 1 if all 10 result documents are relevant.

- P@10: We calculate the precision at 10 for the result list. As we only look at the first 10 results, precision will not separately calculated
- Recall(10): How many of the relevant documents were found? If r<=10, recall is measured based on the actual number r. If r>=10, the recall is measured based on r=10 because only 10 result documents will be looked at. Recall will be 1 if the result list contains only relevant documents and r>=10.
- nDCG: ranked-based measures (discounted cumulative gain): Here the relevance of each documents is important, more relevant documents are more important than less relevant documents. Highly relevant documents should also occur high up in the ranked results lists.

7 Time plan

Month	**Description**	**Requirement**
Dec. 17	Determine 50 topics	
March 18	Descriptions for topics	1 English speaking domain experts
	Assessment guidelines ready	
September 18	Assessment software ready	
June 18 - December 18	Recruiting assessors	1 judges for each language = 3 judges
January 19	Assessments	
March 19	Calculating results	

A Example

Following an example for a given query x: Assessments showed that in the pool are 15 highly relevant documents, 50 partially relevant document and 200 non-relevant documents. So there is the ideal result list possible which could retrieve 10 highly relevant documents.

The Ideal DCG metrics can serve as baseline for comparing to rankings: Ideal DCG based on the above numbers:

Rank	Graded relevance	Relevance value	Cumulated Gain	IDCG
1	Highly relevant	2	2	2
2	Highly relevant	2	4	4
3	Highly relevant	2	6	5,261859507
4	Highly relevant	2	8	6,261859507
5	Highly relevant	2	10	7,123212623
6	highly relevant	2	12	7,896918238
7	Highly relevant	2	14	8,609332612
8	Highly relevant	2	16	9,275999279
9	Highly relevant	2	18	9,906929032
10	Highly relevant	2	20	10,50898902

The results list of query x for the baseline run:

Rank	Doc	Binary Relevance	Graded relevance	Relevance value	CG	DCG
1	Doc1	relevant	Highly relevant	2	2	2
2	Doc2	relevant	partially relevant	1	3	3
3	Doc3	relevant	Highly relevant	2	5	4,2618595
4	Doc4	relevant	partially relevant	1	6	4,7618595
5	Doc5	relevant	Highly relevant	2	8	5,6232126
6	Doc6	relevant	highly relevant	2	10	6,3969182
7	Doc7	relevant	Highly relevant	2	12	7,1093326
8	Doc8	non-relevant	non-relevant	0	12	7,1093326
9	Doc9	relevant	partially relevant	1	13	7,4247975
10	Doc10	relevant	Highly relevant	2	15	8,0268575

System TR1 produces the following result list:

Rank	Doc	Binary Relevance	Graded relevance	Relevance value	CG	DCG
1	Doc1	relevant	Highly relevant	2	2	2
2	Doc3	relevant	Highly relevant	2	4	4
3	Doc52	relevant	partially relevant	1	5	4,63092975
4	Doc14	non-relevant	non-relevant	0	5	4,63092975
5	Doc25	relevant	Highly relevant	2	7	5,49228287
6	Doc16	relevant	partially relevant	1	8	5,87913568
7	Doc7	relevant	Highly relevant	2	10	6,59155005
8	Doc5	relevant	Highly relevant	2	12	7,25821672
9	Doc39	non-relevant	non-relevant	0	12	7,25821672
10	Doc10	relevant	Highly relevant	2	14	7,86027671

Comparing System TR1 to Baseline (as an example only for one query). R-precision, P@10 and Recall are equal for r=>10.

Retrieval metric	**Baseline**	**TR1**
R-precision (10)	0,9	0,8
P@10	0,9	0,8
Recall(10)	0,9	0,8
nDCG at rank 10	0,7638	0,7480