Interactive Sampling for Systematic Reviews. IMS Unipd At CLEF 2018 eHealth Task 2.

Giorgio Maria Di Nunzio[1], Giacomo Ciuffreda[2], and Federica Vezzani[2]

[1] Dept. of Information Engineering – University of Padua
[2] Dept. of Linguistic and Literary Studies – University of Padua
giorgiomaria.dinunzio@unipd.it, giacomo.ciuffreda@studenti.unipd.it, federica.vezzani@phd.unipd.it

Abstract. This is the second participation of the Information Management Systems (IMS) group at CLEF eHealth Task of Technologically Assisted Reviews in Empirical Medicine. This task focuses on the problem of medical systematic reviews, a problem which requires a recall close (if not equal) to 100%. Semi-Automated approaches are essential to support these type of searches when the amount of data exceed the limits of users, i.e. in terms of attention or patience. We present a variation of the two-dimensional approach which 1) sets the maximum amount of documents that the physician is willing to read, 2) takes into account a sampling strategy to estimate the 95% confidence interval of the number of relevant documents present in the collection.

1 Introduction

In this paper, we describe the participation of the Information Management Systems (IMS) group at CLEF eHealth 2018 [10] Task [1]. This task focuses on the problem of systematic reviews, that is the process of collecting articles that summarise all evidence (if possible) that has been published regarding a certain medical topic. This task requires long search sessions by experts in the field of medicine; for this reason, semi-automatic approaches are essential to support these type of searches when the amount of data exceed the limits of users, i.e. in terms of attention or patience.

The objective of our participation to this task was to:

- include a fixed stopping strategy to simulate the maximum amount of documents that a physician is willing to review in the two-dimensional approach presented in [4, 5];
- add a sampling strategy in the interactive process to estimate the 95% confidence interval of the proportion of relevant documents present in the collection.

2 Approach

In this paper, we continue to investigate the interaction with the two dimensional interpretation of the BM25 model applied to the problem of explicit relevance

feedback [8, 2, 7, 4, 6]. In order to explain how the two-dimensional BM25 space works, in the following sections we present a brief review of the BM25 model.

2.1 BM25

The BM25 is a probabilistic retrieval model where the weight of a term in a document is equal to [9]:

$$w_i^{BM25}(tf) = \frac{tf}{k_1\left((1-b)+b\frac{dl}{avdl}\right)+tf} w_i^{BIM} \tag{1}$$

where w_i is the weight of the i-th term, k_1 and b are two parameters (some default parameters are[3] $k_1 = 1.2$ and $b = 0.75$), tf is the term frequency in the document, and w_i^{BIM} is the Binary Independence Model weight of the i-th term:

$$w_i^{BIM} = \log \frac{\theta_i^{\mathcal{R}}}{(1-\theta_i^{\mathcal{R}})} \frac{(1-\theta_i^{\mathcal{NR}})}{\theta_i^{\mathcal{NR}}} \tag{2}$$

where $\theta_i^{\mathcal{R}}$ and $\theta_i^{\mathcal{NR}}$ are the parameters of the Bernoulli random variable that represent the presence (or absence) of the i-th term in the relevant ($\mathcal{R}$) and non-relevant ($\mathcal{NR}$) documents. The estimate of each parameter is:

$$\theta_i^{\mathcal{R}} = \frac{r_i + \alpha^{\mathcal{R}}}{R + \alpha^{\mathcal{R}} + \beta^{\mathcal{R}}} \tag{3}$$

$$\theta_i^{\mathcal{NR}} = \frac{n_i - r_i + \alpha^{\mathcal{NR}}}{N - R + \alpha^{\mathcal{NR}} + \beta^{\mathcal{NR}}} \tag{4}$$

where R is the number of relevant documents, r_i the number of relevant documents in which the i-th term appears, N is the total number of documents and n_i is the total number of documents in which the i-th term appears. Parameters α and β correspond to the hyper-parameter of the conjugate beta prior distribution of the Bernoulli random variable. For $\alpha^{\mathcal{R}} = \beta^{\mathcal{R}} = 0.5$ and $\beta^{\mathcal{R}} =^{\mathcal{NR}}= 0.5$, we obtain the definition of the well-known Robertson - Spärck Jones weight w_i^{RSJ}. Given a document d, the probability of the document being relevant is proportional to:

$$P(\mathcal{R}|d) \propto \sum_{i \in d} w_i^{BM25}(tf) \tag{5}$$

2.2 Two-Dimensional Model

The two-dimensional representation of probabilities [3, 8] is an intuitive way of presenting a two-class classification problem on a two-dimensional space. Given

[3] `http://terrier.org`

two classes, for example relvant $\mathcal{R}$ and non-relevant $\mathcal{NR}$, a document d is assigned to category $\mathcal{R}$ if the following inequality holds:

$$\underbrace{P(d|\mathcal{NR})}_{y} < m \underbrace{P(d|\mathcal{R})}_{x} + q \tag{6}$$

where $P(d|\mathcal{R})$ and $P(d|\mathcal{NR})$ are the likelihoods of the object d given the two categories, while m and q are two parameters that can be optimized to compensate for either the unbalanced class issues or different misclassification costs.

If we interpret the two likelihoods as two coordinates x and y of a two dimensional space, the problem of classification can be studied on a two-dimensional plot. The decision of the classification is represented by the line $y = mx + q$ that splits the plane into two parts: all the points that fall 'below' this line are classified as objects that belong to class $\mathcal{R}$.

Two-dimensional BM25 In order to link the two-dimensional model to the BM25 model, first we define the BIM weight as a difference of logarithms:

$$w_i^{BIM} = \log \frac{\theta_i^{\mathcal{R}}}{(1-\theta_i^{\mathcal{R}})} - \log \frac{\theta_i^{\mathcal{NR}}}{(1-\theta_i^{\mathcal{NR}})} = w_i^{BIM,\mathcal{R}} - w_i^{BIM,\mathcal{NR}} \tag{7}$$

then, we can define the BM25 term weight accordingly

$$w_i^{BM25}(tf) = \underbrace{\frac{tf}{k_1\left((1-b) + b\frac{dl}{avdl}\right) + tf}}_{w_{tf}} \left(w_i^{BIM,R} - w_i^{BIM,NR} \right) \tag{8}$$

$$= w_{tf} \left(w_i^{BIM,R} - w_i^{BIM,NR} \right) \tag{9}$$

$$= w_i^{BM25,\mathcal{R}}(tf) - w_i^{BM25,\mathcal{NR}}(tf) \tag{10}$$

$$\tag{11}$$

We now have all the elements to define the two coordinates $x = P(d|\mathcal{R})$ and $y = P(d|\mathcal{NR})$ in the following way:

$$P(d|\mathcal{R}) = \sum_{i \in d} w_i^{BM25,\mathcal{R}}(tf) \tag{12}$$

$$P(d|\mathcal{NR}) = \sum_{i \in d} w_i^{BM25,\mathcal{NR}}(tf) \tag{13}$$

where $\sum_{i \in d}$ indicates (with an abuse of notation) the sum over all the terms of document d.

In Figure 1, we show an example of the visualization of a collection of documents using the two-dimensional BM25 model. Relevant and non relevant documents which have already been judged by a user (in our case the physician) are colored in green and red; documents that have not been judged are greyed. The two lines represents two possibile decision lines (see Equation 6) to rank/classify new documents as relevant.

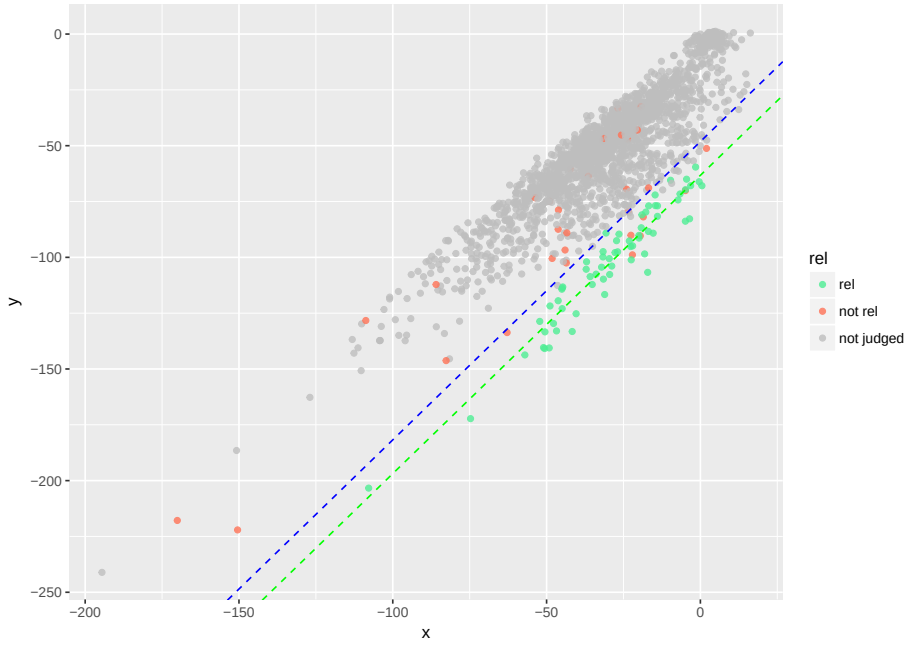

Fig. 1: Example of two-dimensional view of the BM25 model.

3 Method

We focused on the following problems:

1. study the effectiveness of a classifier given a fixed amount of documents that a physician is willing to review;
2. design a sampling strategy to estimate the 95% confidence interval of the number of relevant documents in the collection.

In the experiments, we used the following procedure:

- we set a number n of documents that the physician is willing to read and a number s that tells the algorithm when (every s documents) to randomly sample a document from the collection instead of presenting to the physician the next most relevant document;
- for each topic, we run an optimized (hyper-parameters) BM25 retrieval model and we obtain the relevance feedback for the first abstract in the ranking list;
- from the second document until $n/2-1$, we continuously update the relevance weights of the terms according to the explicit relevance feedback given by the physician (simulated by the qrels available with the test collection);

– for the last half of the documents $n/2$ that the physician is willing to read, we use a Naïve Bayes classifier continuously updated with the explicit relevance feedback [4].

4 Experiments

For all the experiments, we set the values of the BM25 hyper-parameters in the following way:

– $\alpha^{\mathcal{R}} = \alpha^{\mathcal{NR}} = 1.0$
– $\beta^{\mathcal{R}} = \beta^{\mathcal{NR}} = 0.01$

These values are consistent with other experiments and indicate that a beta prior distribution that discounts the 'presence' of a term in favour of its 'absence' (high α and low β) results in a better retrieval performance [5]. The slope m of the decision line is set $m = 1.0$ and $q = 0$ for the first half $n/2$ of the documents; then, m and q are continuously updated according to the relevance information [4].

4.1 Official runs

We submitted three runs by varying the number of documents n that the physician is willing to read per topic: $n = 1000$, $n = 2000$, $n = 3000$. We set the parameter $s = 10$, this means that every ten documents we sample a random document from the collection instead of showing to the physician the next ranked document. The three official runs are named as follows:

– ims_unipd_t500.task2, $n = 1000$
– ims_unipd_t1000.task2, $n = 2000$
– ims_unipd_t1500.task2, $n = 3000$

In Figure 2, we show the recall per topic for each official run. We see that there are two topic in particular that are more difficult than the others: CD009263 and CD012010 with a recall less (or close to) 0.6 for all the runs. Seven topics can be considered as medium difficult (recall between 0.6 and 0.6 for at least one of the experiments): CD008567, CD010213, CD010502, CD012165, CD012179, CD012281, CD012599.

in Figure 3, we compare the results of our three runs with the summary of all the other CLEF 2018 participant. This plot confirms that most of high and medium difficult topics are also topics that, on average, were difficult for most of the participants (barplots more stretched and median far from value 1.00).

Confidence intervals of number of relevant documents During the experiments, every 10 documents we sample a random document from the collection and show the document for relevance assessment in order to estimate the number of relevant documents in the collection. In Table 1, 2, and 3, we show a breakdown

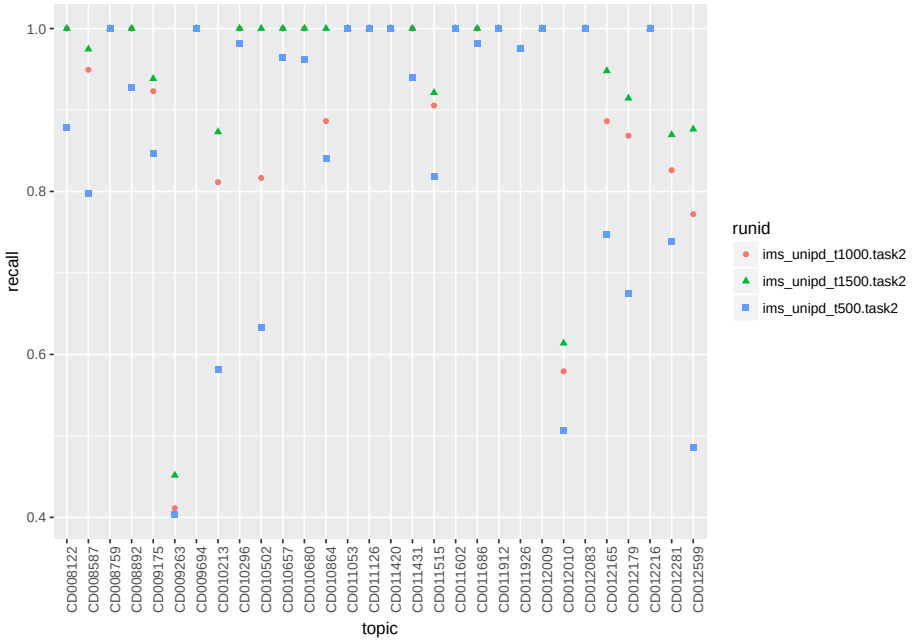

Fig. 2: Recall per topic of the three official runs.

of the number of documents per topic, how many topics were read (explicit relevance feedback), the number of relevant documents, how many documents were randomly sampled, the estimate of the number of relevant documents based on the random sample as well as the 95% confidence interval (minimum and maximum range), and the number of relevant documents found within the limit of the threshold. In most cases, the estimate of the number of relevant documents (and the 95% range) is much larger than the true number of relevant documents. The analysis of the results shown in these table is still under study since we would need a more sophisticated cost-benefit model to understand whether we want to put more effort in the estimate of the number of relevant documents or in the automatic classifier.

4.2 Unofficial runs

In addition to the three official runs, we prepared two unofficial runs in order to study the feasibility of the query rewriting approach based on the work of [5]. We asked two experts in linguistics to rewrite the query, each with a different goal: the first variant is written with the aim of creating a list of keywords resulting from the semic analysis (the study of meaning in linguistic units) of the technical terms contained in the initial query. The second variant is written with the aim of reformulating the information need into a humanly readable sentence using

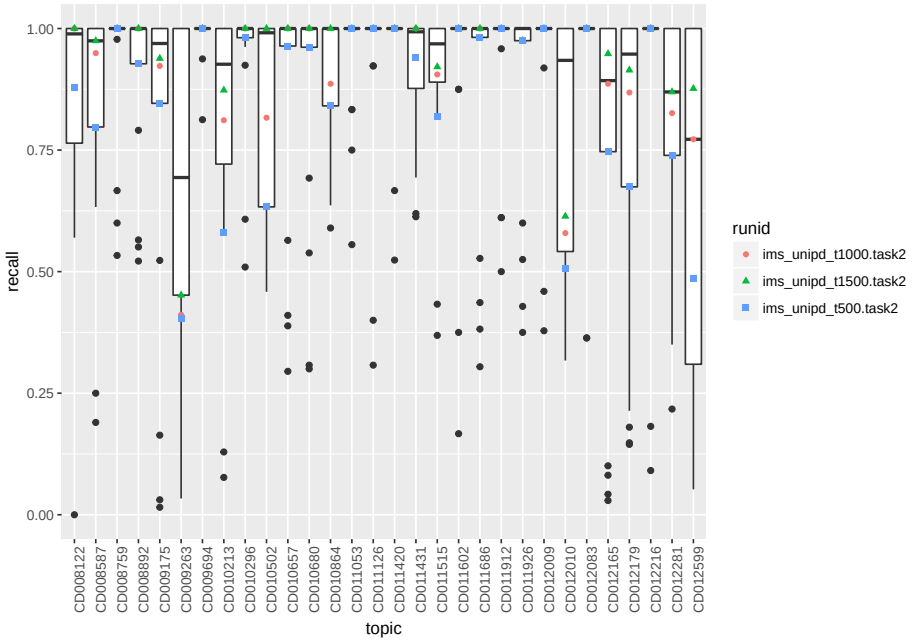

Fig. 3: Boxplot of recall per topic of all the CLEF 2018 runs subtask 2 overlapped with our three official runs.

alternative terms such as synonyms, orthographic variants, related forms and/or acronyms. The two experts worked independently from each other by following a structured linguistic methodology and focusing on different terminological aspects. We name these two experiments with "keyword" and "readable".

Linguistic Methodology: Terminological Record The methodology applied for the process of query rewriting is based on a linguistic and terminological analysis of all the technical terms contained in the information needs provided in the dataset. The approach is divided into the following steps:

1. Recognition of technical terms;
2. Extraction of technical terms;
3. Linguistic and semantic analysis;
4. Formulation of terminological records;
5. Query rewriting.

The core of our methodology is basically a new model of terminological record used for the analysis of medical terminology [11]. This tool is a structured set of terminological data referring to a specific concept and it is used in order to provide linguistic information about the concept itself and the term used for its

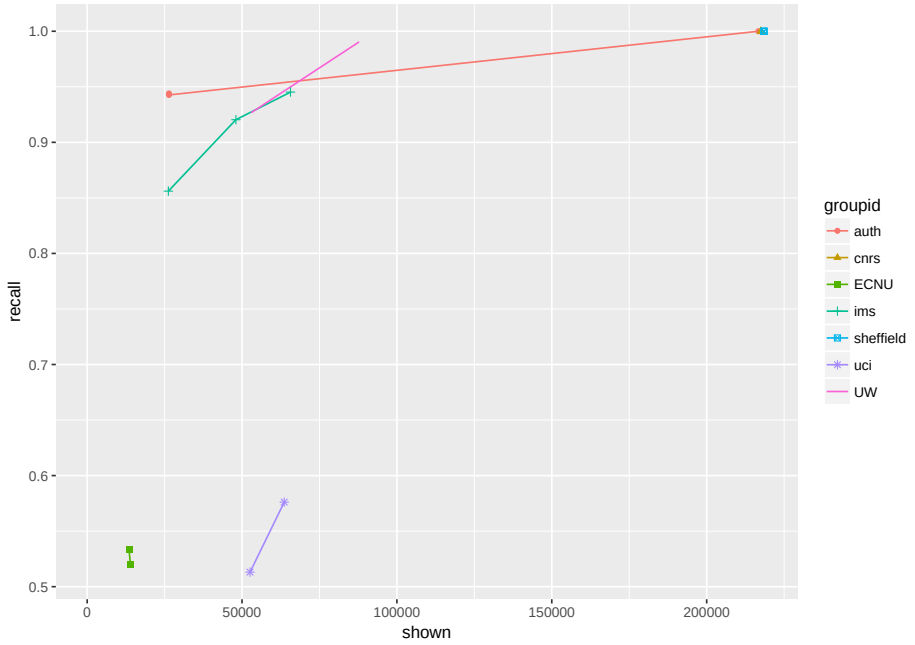

Fig. 4: Recall per number of documents shown of all the CLEF 2018 runs subtask 2. Runs have been grouped per participant.

designation both for its lexical and semantic framing. This terminological record is composed of four general fields, which individually refer to formal features, semantics, corpus and references. Each field in turn is divided in specific subfields describing the term according to linguistic and notional criterions.

Focusing on the first two subfields, the section named "formal feature" contains lexical and morphosyntactic information such as genre, tonic accent, spelling, etymology (derivation and composition), orthographic variant, acronyms/expansions and related forms. From the semantic viewpoint, the subfield "semantics" contains the definition of the term, its semic analysis, cases of phraseology (collocations and colligations) and all the possible semantic variants.

For example for topic CD011602, the information need provided is:

> Ultrasonography for diagnosis of alcoholic cirrhosis in people with alcoholic liver disease.

We initially proceeded with the extraction of technical terms (both single-word and multi-word terms) such as *ultrasonography, diagnosis, alcoholic cirrhosis, cirrhosis, alcoholic liver disease, liver, disease* and then we started to formulate terminological records for each of them. The subfield named "formal feature" was useful for the human readable reformulation, whereas 'semantics" subfield provided the information necessary for the keywords reformulation.

Table 1: Number of documents and relevant documents (true and estimated) per topic. Experiment with threshold t = 500.

topic	num_docs	docs_read	num_rel	sampled	est_rel	range_min	range_max	found_rel
CD008122	1911	987	272	76	150	54	246	239
CD008587	9152	993	79	104	722	255	1189	63
CD008759	932	932	60	65	73	27	119	60
CD008892	1499	987	69	74	118	43	193	64
CD009175	5644	992	65	105	445	158	732	55
CD009263	78803	995	124	639	6221	2187	10255	49
CD009694	161	161	16	10	12	6	18	16
CD010213	15198	993	599	114	1199	423	1976	348
CD010296	4602	991	53	88	363	129	597	52
CD010502	2985	990	229	91	235	84	386	144
CD010657	1859	989	139	83	146	53	240	133
CD010680	8405	993	26	100	663	235	1092	25
CD010864	2505	989	44	84	197	71	324	37
CD011053	2235	989	12	90	176	63	288	12
CD011126	6000	994	13	156	473	168	779	13
CD011420	251	251	42	10	19	9	30	42
CD011431	1182	984	297	62	93	34	151	279
CD011515	7244	992	127	95	571	202	940	104
CD011602	6157	994	8	129	486	172	799	8
CD011686	9443	994	55	208	745	263	1227	54
CD011912	1406	989	36	76	111	40	181	36
CD011926	4050	994	40	96	319	114	525	39
CD012009	536	536	37	29	42	16	67	37
CD012010	6830	994	290	99	539	191	887	146
CD012083	322	322	11	16	25	11	39	11
CD012165	10222	993	308	117	807	285	1328	229
CD012179	9832	995	304	119	776	274	1277	205
CD012216	217	217	11	16	17	8	26	11
CD012281	9876	994	23	157	779	275	1283	17
CD012599	8048	994	575	103	635	225	1045	279

First variant: keywords reformulation In particular, semic analysis turns out to be the most useful process for the keyword reformulation and it aims to decompose the meaning of the term analyzed. This process consists of breaking down the sememe (i.e. the meaning) of a word in all its sense components, e.g. the semes. So for exemple, for the term *cirrhosis* the process of decomposition of meaning produced the following list of keywords: /chronic disease/ /liver/ /degeneration/ /cells/ /human body/ /inflammation/ /fibrous/ /thickening/ /tissue/ /alcoholism/ /hepatitis/.

We repeat this kind of analysis of each technical term in the information need and considering the above mentioned exemple for topic CD011602, the keyword reformulation is the following:

Table 2: Number of documents and relevant documents (true and estimated) per topic. Experiment with threshold t = 1000.

topic	num_docs	docs_read	num_rel	sampled	est_rel	range_min	range_max	found_rel
CD008122	1911	1911	272	104	150	54	246	272
CD008587	9152	1987	79	190	722	255	1189	75
CD008759	932	932	60	65	73	27	119	60
CD008892	1499	1499	69	90	118	43	193	69
CD009175	5644	1986	65	185	445	158	732	59
CD009263	78803	1994	124	737	6221	2187	10255	50
CD009694	161	161	16	10	12	6	18	16
CD010213	15198	1989	599	209	1199	423	1976	486
CD010296	4602	1981	53	163	363	129	597	53
CD010502	2985	1984	229	136	235	84	386	187
CD010657	1859	1859	139	122	146	53	240	139
CD010680	8405	1990	26	188	663	235	1092	26
CD010864	2505	1971	44	132	197	71	324	39
CD011053	2235	1973	12	126	176	63	288	12
CD011126	6000	1991	13	234	473	168	779	13
CD011420	251	251	42	10	19	9	30	42
CD011431	1182	1182	297	65	93	34	151	297
CD011515	7244	1986	127	169	571	202	940	115
CD011602	6157	1988	8	202	486	172	799	8
CD011686	9443	1990	55	289	745	263	1227	55
CD011912	1406	1406	36	86	111	40	181	36
CD011926	4050	1987	40	160	319	114	525	39
CD012009	536	536	37	29	42	16	67	37
CD012010	6830	1986	290	178	539	191	887	168
CD012083	322	322	11	16	25	11	39	11
CD012165	10222	1988	308	200	807	285	1328	272
CD012179	9832	1991	304	197	776	274	1277	264
CD012216	217	217	11	16	17	8	26	11
CD012281	9876	1992	23	233	779	275	1283	19
CD012599	8048	1988	575	180	635	225	1045	444

/technique/ /echoes/ /ultrasound pulses/ /ultrasound/ /pulse/ /delineate/ /areas/ /different density/ /body/ /human being/ /cells/ /examination/ /evaluation/ /diagnostic/ /diagnosing/ /diagnose/ /alcohol/ /chronic/ /disease/ /cirrhosis of the liver/ /liver/ degeneration/ /cells/ /inflammation/ /fibrous/ /thickening/ /tissue/ /alcoholism/ /hepatitis/ /patient/ /large lobed glandulare organ/ /abdomen/ vertebrates/ /metabolic processes/ /disorder/ /structure/ /function/ / symptoms/ /affect/ /location/ /physical injury/.

Second variant: human readable reformulation The second type of query was written with the aim of reformulating the information need in a humanly readable sentence. Thanks to terminological records, we have been able to replace

Table 3: Number of documents and relevant documents (true and estimated) per topic. Experiment with threshold t = 1500.

topic	num_docs	docs_read	num_rel	sampled	est_rel	range_min	range_max	found_rel
CD008122	1911	1911	272	104	150	54	246	272
CD008587	9152	2984	79	274	722	255	1189	77
CD008759	932	932	60	65	73	27	119	60
CD008892	1499	1499	69	90	118	43	193	69
CD009175	5644	2977	65	238	445	158	732	61
CD009263	79786	2993	124	1811	6298	2214	10383	56
CD009694	161	161	16	10	12	6	18	16
CD010213	15198	2984	599	283	1199	423	1976	523
CD010296	4602	2967	53	210	363	129	597	53
CD010502	2985	2954	229	170	235	84	386	229
CD010657	1859	1859	139	122	146	53	240	139
CD010680	8405	2983	26	270	663	235	1092	26
CD010864	2505	2505	44	142	197	71	324	44
CD011053	2235	2235	12	161	176	63	288	12
CD011126	6000	2979	13	293	473	168	779	13
CD011420	251	251	42	10	19	9	30	42
CD011431	1182	1182	297	65	93	34	151	297
CD011515	7244	2976	127	244	571	202	940	117
CD011602	6157	2982	8	268	486	172	799	8
CD011686	9443	2979	55	370	745	263	1227	55
CD011912	1406	1406	36	86	111	40	181	36
CD011926	4050	2970	40	199	319	114	525	39
CD012009	536	536	37	29	42	16	67	37
CD012010	6830	2977	290	242	539	191	887	177
CD012083	322	322	11	16	25	11	39	11
CD012165	10222	2981	308	282	807	285	1328	292
CD012179	9832	2984	304	269	776	274	1277	278
CD012216	217	217	11	16	17	8	26	11
CD012281	9876	2985	23	321	779	275	1283	20
CD012599	8048	2978	575	246	635	225	1045	503

original terms with validly attested synonyms and use orthographic alternatives as variants of the medical terms provided in the original information need as well as to systematically replace acronyms with their expansions and expansions with their acronyms. Considering the previous topic CD011602, we obtained the following readable reformulation:

> Diagnostic accuracy of medical ultrasound, known as diagnostic sonography or ultrasonography, for the detection of alcoholic liver disease (ALD) as the liver manifestations of alcohol overconsumption, including fatty liver, alcoholic hepatitis, and chronic hepatitis with liver fibrosis or cirrhosis.

Table 4: Recall at documents shown: official vs unofficial results

	t = 500			t = 1000			t = 1500		
topic	original	readable	keyword	original	readable	keyword	original	readable	keyword
CD008122	0.879	0.882	0.879	1.000	1.000	1.000	1.000	1.000	1.000
CD008587	0.797	0.797	0.785	0.949	0.962	0.962	0.975	0.975	0.975
CD008759	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD008892	0.928	0.928	0.928	1.000	1.000	1.000	1.000	1.000	1.000
CD009175	0.846	0.846	0.846	0.923	0.923	0.923	0.938	0.938	0.938
CD009263	0.403	0.395	0.355	0.411	0.427	0.427	0.452	0.452	0.476
CD009694	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD010213	0.581	0.589	0.611	0.811	0.816	0.820	0.873	0.871	0.871
CD010296	0.981	0.981	0.981	1.000	1.000	1.000	1.000	1.000	1.000
CD010502	0.633	0.633	0.633	0.817	0.996	1.000	1.000	1.000	1.000
CD010657	0.964	0.964	0.971	1.000	1.000	1.000	1.000	1.000	1.000
CD010680	0.962	0.962	0.962	1.000	1.000	1.000	1.000	1.000	1.000
CD010864	0.841	0.841	0.841	0.886	0.909	0.909	1.000	1.000	1.000
CD011053	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD011126	1.000	1.000	0.846	1.000	1.000	1.000	1.000	1.000	1.000
CD011420	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD011431	0.939	0.946	0.939	1.000	1.000	1.000	1.000	1.000	1.000
CD011515	0.819	0.811	0.819	0.906	0.906	0.906	0.921	0.921	0.921
CD011602	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD011686	0.982	0.982	0.982	1.000	1.000	1.000	1.000	1.000	1.000
CD011912	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD011926	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975
CD012009	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD012010	0.507	0.507	0.507	0.579	0.579	0.583	0.614	0.617	0.614
CD012083	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD012165	0.747	0.737	0.727	0.886	0.880	0.880	0.948	0.942	0.945
CD012179	0.674	0.681	0.678	0.868	0.875	0.878	0.914	0.921	0.921
CD012216	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CD012281	0.739	0.739	0.783	0.826	0.826	0.826	0.870	0.870	0.826
CD012599	0.485	0.492	0.483	0.772	0.767	0.765	0.877	0.875	0.878
	0.856	0.856	0.851	0.920	0.928	0.928	0.945	0.945	0.945

5 Discussion

We are currently evaluating the results of these reformulation topic by topic, Table 4, and studying the impact, from a linguistic point of view, of a query reformulation in the top 10 retrieved documents, Table 5.

In this phase of the analysis, we noted that there are some topics for which the two reformulations ("keywords" and/or "readable") retrieved, in the first 10 positions, more relevant documents than the original query. Table 6 shows these topics and the number of documents retrieved depending on the type of reformulation. We then proceed with the manual analysis of such topics by reading the abstracts of the relevant documents retrieved from the two variants and we started to analyse from a linguistic viewpoint which terms contained in the two reformulations allowed the retrieval of such relevant documents.

As a first and approximate analysis, we noted that the terms that were most frequently used in the two reformulations are those related to the diagnostic and evaluative sphere such as *diagnosis* and related forms as *diagnostic, diagnose* and *diagnosing* as well as *evaluation, examination, test* and *detection*. Furthermore, even the replacement of the full multi-word terms with the acronym such as DMSA for *Dimercaptosuccinic Acid Scan*, VUR for *Vesicoureteral Reflux* and UTI for *Urinary Tract Infection*, has turned out to be a good approach because

Table 5: Precision at 10 documents for each topic for the official runs.

topic	original	readable	keyword
CD008122	0.800	0.100	0.200
CD008587	0.300	0.000	0.000
CD008759	0.700	0.000	0.000
CD008892	0.700	0.600	0.300
CD009175	0.400	0.200	0.100
CD009263	0.500	0.000	0.000
CD009694	0.500	0.600	0.300
CD010213	0.500	0.400	0.000
CD010296	0.600	0.000	0.000
CD010502	0.600	0.600	0.500
CD010657	0.400	0.600	0.100
CD010680	0.100	0.300	0.000
CD010864	0.000	0.000	0.000
CD011053	0.400	0.200	0.100
CD011126	0.100	0.000	0.000
CD011420	0.600	0.900	0.800
CD011431	0.400	0.000	0.800
CD011515	0.100	0.100	0.000
CD011602	0.100	0.100	0.100
CD011686	0.100	0.700	0.000
CD011912	0.400	0.400	0.200
CD011926	0.400	0.700	0.600
CD012009	0.000	0.200	0.400
CD012010	1.000	1.000	0.100
CD012083	0.300	0.000	0.300
CD012165	0.200	0.100	0.300
CD012179	0.600	0.300	0.000
CD012216	0.100	0.000	0.100
CD012281	0.100	0.100	0.000
CD012599	0.400	0.400	0.000

reduced lexical forms are one of the typical feature of medical language and abbreviations are used in order to rapidly transmit health information.

6 Ongoing and Future Work

In this work, we presented a continuous active learning approach that uses a fixed stopping strategy to simulate the maximum amount of documents that a physician is willing to review, and a sampling strategy that is used to estimate the number of relevant documents in the collection. We are currently performing a failure analysis to understand the possibile reasons of a recall below 90% and identify the linguistic aspects of a query rewriting approach that may help to improve the performance of an interactive system.

Table 6: Topics and number of relevant documents retrieved

topic	readable	keywords
CD009694	1	0
CD010657	5	1
CD010680	2	0
CD011420	7	7
CD011431	0	8
CD011686	7	0
CD011926	5	4
CD012009	2	4
CD012165	1	3

References

1. Evangelos Kanoulas, Rene Spijker, Dan Li, and Leif Azzopardi, editors. *CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview. CLEF 2018 Evaluation Labs and Workshop: Online Working Notes*, CEUR Workshop Proceedings. CEUR-WS.org, 2018.
2. Giorgio Maria Di Nunzio. A new decision to take for cost-sensitive naïve bayes classifiers. *Inf. Process. Manage.*, 50(5):653–674, 2014.
3. Giorgio Maria Di Nunzio. Interactive text categorisation: The geometry of likelihood spaces. *Studies in Computational Intelligence*, 668:13–34, 2017.
4. Giorgio Maria Di Nunzio. A study of an automatic stopping strategy for technologically assisted medical reviews. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 672–677, 2018.
5. Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. An interactive two-dimensional approach to query aspects rewriting in systematic reviews. IMS unipd at CLEF ehealth task 2. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.
6. Giorgio Maria Di Nunzio, Maria Maistro, and Federica Vezzani. A gamified approach to naïve bayes classification: A case study for newswires and systematic medical reviews. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1139–1146, 2018.
7. Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. Gamification for machine learning: The classification game. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval co-located with 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), Pisa, Italy, July 21, 2016.*, pages 45–52, 2016.
8. Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. The university of padua (IMS) at TREC 2016 total recall track. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, 2016.
9. Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

10. Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Evangelos Kanoulas, Leif Azzopardi, Rene Spijker, Dan Li, Aurélie Névéol, Lionel Ramadier, Aude Robert, Joao Palotti, Jimmy, and Guido Zuccon, editors. *Overview of the CLEF eHealth Evaluation Lab 2018. CLEF 2018 - 8th Conference and Labs of the Evaluation Forum*, volume Lecture Notes in Computer Science (LNCS). Springer, September 2018.
11. Federica Vezzani, Giorgio Maria Di Nunzio, and Geneviève Henrot. Trimed: A multilingual terminological database. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, 2018.