CUNI at the CLEF eHealth 2015 Task 2

Shadi Saleh, Feraena Bibyna, and Pavel Pecina

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics, Czech Republic
{saleh,pecina}@ufal.mff.cuni.cz,feraena.b@gmail.com

Abstract. We present our participation as the team of the Charles University in Prague at the CLEF eHealth 2015 Task 2. We investigate performance of different retrieval models, linear interpolation of multiple models, and our own implementation of blind relevance feedback for query expansion. We employ MetaMap as an external resource for annotating the collection and the queries, then conduct retrieval at concept level rather than word level. We use MetaMap for query expansion. We also participate in the multilingual task where queries were given in several languages. We use Khresmoi medical translation system to translate the queries from Czech, French, and German into English. For the other languages we use translation by Google Translate and Bing Translator.

Keywords: multilingual information retrieval, language models, UMLS, blind relevance feedback, linear interpolation

1 Introduction

Can we use the current web search engines to look for medical information? Authors in [17] showed that when users pose queries describing specific symptoms or general health information, the current search engines can not effectively retrieve relevant documents. This can lead to dangerous consequences if users try to apply the results they obtain for self-treatment. The biggest challenge in medical information retrieval is that users do not have enough medical knowledge so they cannot choose the correct medical terms which described their information needs. This often leads to "circumlocution" when the query contains more and vague words instead of less but specific medical terms. Modern information retrieval systems have started to move in the direction of concepts rather than terms which helps to solve the "circumlocution" problem [14].

2 Task description

The goal of CLEF eHealth 2015 Task 2 [6, 10] is to design an IR system which returns a ranked list of medical documents (English web pages) from the provided test collection as a response to patients' queries. The task is defined as a standard TREC-style text IR task[1].

[1] http://trec.nist.gov/

```
<doc>
  <docid>wiki.0842_12_009733</docid>
  <title>
    Testing for Celiac Disease ...
  </title>
  <title_concepts>
    C0683443
    C0007570
    C0521125
    ...
  </title_concepts>
  <text>
    Intestinal biopsy is the gold standard for diagnosing celiac ...
  </text>
  <text_concepts>
   C1704732
   C0036563
   C0423896
   ...
  </text_concepts>
</doc>
```

Fig. 1. An example of an annotated document.

3 Data

3.1 Document Collection

The collection for Task 2 contains about one million documents provided by the Khresmoi project[2]. It contains automatically crawled web pages from popular medical websites. Non-HTML documents (e.g., pdf, rtf, ppt, and doc) which were found in the collection were excluded. The HTML documents were cleaned using the simple HTML-Strip[3] Perl module. Other more advanced tools and statistical approaches for cleaning HTML documents did not bring any improvement in our previous experiments [12]. After cleaning the documents, we used MetaMap [1] to annotate the data with concept identifiers from the UMLS Metathesaurus [15, 2] version 2014AA. The UMLS Metathesaurus is a large vocabulary database containing information about biomedical and health-related concepts, their names and relationship between them. Terms are linked to others by various relationship such as synonymy, hypernymy, hyponymy, lexical variations, and many others. The Metathesaurus is organized by concept, which symbolize a semantic concept or a meaning. Each concept or meaning in the Metathesaurus has a unique and permanent concept identifier (CUI). We utilize MetaMap's highly configurable options in our annotation process. We use the `-I` option so that the concept IDs are shown, and `-y` option to enable word sense disambiguation. The text is broken down into the components that include sentences, phrases, lexical elements and tokens. The disambiguation module then process the variants and output a final mapping. We put this concept annotations into an additional XML field in the document and query files. An example of cleaned and annotated document is given in Figure 1.

[2] `http://khresmoi.eu/`

[3] `http://search.cpan.org/dist/HTML-Strip/Strip.pm`

Table 1. Statistics of the query sets: number of queries, average number of tokens in titles and total number of relevant documents.

query set	queries	title length	relevant documents
CLEF 2014 test set	50	4.30	3,209
CLEF 2015 test set	66	5.03	1,972

3.2 Queries

We use the test queries from the CLEF eHealth 2014 Task 3 [5] to train our system. The experiments are evaluated using the newly created CLEF eHealth 2015 Task 2 test queries. We have annotated both sets of queries by MetaMap the same way as the documents. Some basic statistics associated with the query sets are shown in Table 1.

4 System description

4.1 Retrieval model

We use Terrier [9] to index the collection and to conduct the retrieval, we examine several weighting models and based on comparing P@10 performance for those models, we choose the following:

- **Bayesian smoothing with Dirichlet prior weighting model (DIR)** This retrieval model is based on language modelling. The documents are scored by calculating the product of each term's probability in the query using language model for that document. Term probabilities in a document are estimated by maximum likelihood estimation. This might cause zero probabilities when a query term does not appear in the document. To avoid this problem the estimated probability distribution can be smoothed by various methods [16]. This retrieval model employs Bayesian smoothing with Dirichlet prior which uses different amount of smoothing based on the length of the document, for longer documents the smoothing will be less. The smoothing parameter is set by default to 2500. For more details and comparison with another smoothing methods see [13].
- **Per-field normalisation weighting model (PL2F)** This model extends Poisson model with Laplace after-effect and normalisation 2 (PL2) model. PL2 is based on Divergence from Randomness (DFR) document weighting models. The basic idea behind the DFR models is that the term frequency of a term in a document carries more information when it divergences from its distribution in the collection. In PL2F, each term in the document is assigned to one field and the frequency of that term is normalised and weighted independently of other fields [8].
- **LGD weighting model** In this model, DFR approach is used together with log-logistic distribution, see [3].

4.2 Query Expansion using the UMLS Metathesaurus

In UMLS, a concept can be represented by many different names. In other words, a concept structure represents synonymy relationships, i.e. terms assigned to the same concept are synonymous. For example, concept `C0010054` corresponds to the term `coronary arteriosclerosis`. This term is synonymous with the term `coronary heart disease`, which also has the same CUI in the Metathesaurus. There are 113 other terms that are assigned with this CUI. In our experiment, we utilize this relationship to pick the candidate for query expansion.

As mentioned in the previous section, the queries are annotated with concept identifiers. For every concept in a query, we generate a list of synonymous terms under that concept. We keep the original query terms, and the added candidate terms that are not yet in the query terms. We do not add all the synonymous terms, but only up to five words that have the highest inverse document-frequency in the collection. In one of our runs, we further filter this words by doing an initial document retrieval using the original query, and then only use the synonyms that also appeared in top n relevant documents. We utilize Terrier's query language[4] to experiment on field weighting with the query. Terrier query language has several operators with different functions. In our experiment, we used the ^operator that is used to assign weights to words. `term1^2` means that the weight of `term1` is multiplied by 2. There are three kinds of fields that we utilize: the original query terms, the expanded query terms, and the concept identifiers from the original terms. We assign different weights to different fields. We tune our system using the CLEF 2014 test set to get the best weights configuration. Query language is only available in Terrier for single line query format, so we have to first convert the provided TREC format to single line format. Figure 2 shows some samples of weighted single line queries, where original terms, expansion terms, and concept IDs are given weight 4.1, 0.6, and 0.1 respectively.

4.3 Blind relevance feedback

Blind relevance feedback (BRF) automates user's part of relevance feedback by expanding user's query using extra information from the collection [4]. In BRF, an initial retrieval steps is performed to find a set of n highly-ranked documents. These documents are assumed to be relevant and a set of m terms from these documents is extracted and added to the original query and the final retrieval step is conducted using the expanded query. In our experiments, the term selection is based on IDF scores extracted from the entire collection. We tune both n and m parameters using the CLEF 2014 test set. We found that adding just one term from top 25 documents gives the highest P@10.

4.4 Linear interpolation

In some of our experiments, we perform linear interpolation of scores from multiple retrieval models.The following equation generates a new (interpolated) score

[4] `http://terrier/org/docs/v4.0/querylanguage.html`

```
clef2015.test.1 many^4.1 red^ 4.1 marks^4.1 on^4.1 legs^4.1 after^4.1
    traveling^4.1 from^4.1 us^4.1 color^0.6 poly^0.6 entity^0.6
    traveling^0.6 trailing^0.6 redness^0.6 marking^0.6 crus^0.6 markedly
    ^0.6 status^0.6 qualifier^0.6 regio^0.6 markings^0.6 crural^0.6
    massively^0.6 observable^0.6 subdivision^0.6 colour^0.6 cruris^0.6
    value^0.6 travels^0.6 values^0.6 C0747726^0.1 C1260956^0.1 C0522501
    ^0.1 C1140621^0.1 C0687676^0.1 C0040802^0.1

clef2015.test.2 lump^4.1 with^4.1 blood^4.1 spots^4.1 on^4.1 nose^4.1
    body^0.6 entire^0.6 localised^0.6 naevus^0.6 masses^0.6 angiomas^0.6
     morphologic^0.6 angioma^0.6 lump^0.6 haemangioma^0.6 ectasia^0.6
    structure^0.6 C0577559^0.1 C0343082^0.1 C1278896^0.1

clef2015.test.3 dry^4.1 red^4.1 and^4.1 scaly^4.1 feet^4.1 in^4.1
    children^4.1 feet^0.6 qualifier^0.6 color^0.6 colour^0.6 abnormal
    ^0.6 value^0.6 rubor^0.6 youth^0.6 person^0.6 foot^0.6 desquamation
    ^0.6 finding^0.6 physical^0.6 childhood^0.6 redness^0.6 C0205222^0.1
     C0332575^0.1 C0239639^0.1 C0008059^0.1

clef2015.test.4 itchy^4.1 lumps^4.1 skin^4.1 observable^0.6 d05^0.6
    localised^0.6 specimen^0.6 morphologic^0.6 sample^0.6 tissue^0.6
    lump^0.6 dup^0.6 excl^0.6 masses^0.6 x16^0.6 C0033774^0.1 C0577559
    ^0.1 C0444099^0.1

clef201.test.5 whistling^4.1 noise^4.1 and^4.1 cough^4.1 during^4.1
    sleeping^4.1 ^4.1 children^4.1 sound^0.6 noise^0.6 noises^0.6 signal
    ^0.6 whistlings^0.6 event^0.6 sleeping^0.6 youth^0.6 person^0.6
    asleep^0.6 adverse^0.6 activiaty^0.6 finding^0.6 cough^0.6 data^0.6
    childhood^0.6 C3494463^0.1 C0028263^0.1 C1961131^0.1 C0424522^0.6
    C0008059^0.1
```

Fig. 2. Examples of weighted queries in Run5.

for each document/query pair:

$$Score(D,Q) = \lambda \cdot Score_1(D,Q) + (1-\lambda) \cdot Score_2(D,Q)$$

We tune lambda value using the CLEF 2014 test set to get highest P@10. Figure 3 shows the curves for the CLEF 2014 test set and the CLEF 2015 test set for P@10 when we interpolate Run5 and Run6, lambda is set to 0.71.

4.5 Term weighting

In the multilingual task, we use our term weighting algorithm for languages Czech, French and German to expand queries. First we use Khresmoi translation system to translate the queries into English and return the *n-best-list* translations. These translations form a translations pool. Each term in the translations pool is assigned a weight, which is the score of its translation hypothesis given by the translation model, say $TM(term)$. Then, we use the BRF algorithm as described in Section 4.3 to retrieve the n highly-ranked documents, where the queries are taken only from the best translation. For each term in the translations pool, we calculate the IDF score in the entire collection and the term frequency (TF) in the translations pool. We normalise all of these scores and then each term is weighted as follows:

$$Score(term) = TM(term)^{\lambda_1} \cdot TF(term)^{\lambda_2} \cdot IDF(term)^{(10-\lambda_1-\lambda_2)}$$

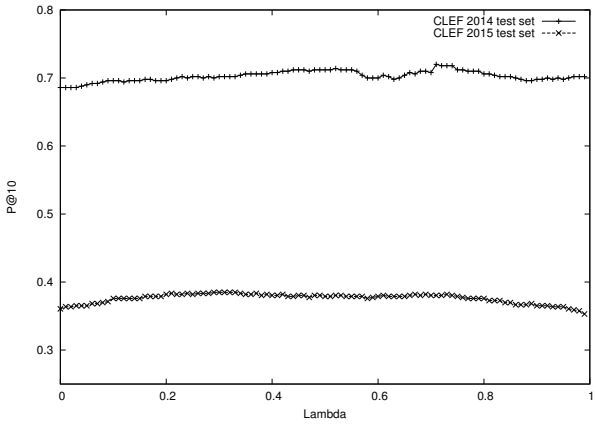

Fig. 3. Tuning the lambda parameter on the CLEF 2014 test set and the scores obtained on the CLEF 2015 test data while interpolating Run5 and Run6.

Where lambda values sum up to 10. After scoring terms in the translations pool, we sort them descending by their score and add top m terms into original query. Lambda values, n and m are trained using the CLEF 2014 test set.

5 Experiments and results

5.1 Monolingual Task

We submitted to this task 10 runs, summarised in Table 2, as follows:

- Run1, Run2, and Run3 employ Terrier's implementation of DIR, LGD, and PL2F retrieval models, respectively. Indexing and retrieval are conducted at term level.
- Run4 is a linear combination of Run1, Run2, and Run3 with parameters set to 0.56, 0.32, and 0.12, respectively. The values were tuned on the CLEF eHealth 2014 test queries.
- Run5 employs query expansion based on the UMLS Metathesaurus. For each UMLS concept ID in each query, we retrieved all entries with that concept ID and then expanded that query with 5 words that have the highest IDF in the collection, excluding words that already occur in the original query. We used original terms, concept id, and expansion terms for the retrieval process using Terrier's implementation of PL2F. Each field is weighted differently, see Section 4.2.
- Run6 uses the same settings at Run5 but employs the LGD retrieval model.
- Run7 interpolates Run5 and Run6 with parameters 0.71 and 0.29, respectively.
- Run8 interpolates Run1 with a system that only uses concept IDs for retrieval (using PL2F model), with parameters 0.98 and 0.01, respectively.

Table 2. Description of the monolingual runs.

run ID	query	doc	model	details
Run1	term	term	DIR	-
Run2	term	term	LGD	-
Run3	term	term	PL2F	-
Run4	-	-	-	lin. interpolation of Run1, Run2, Run3
Run5	term&concept	term	PL2F	UMLS query expansion
Run6	term&concept	term	LGD	UMLS query expansion
Run7	-	-	-	lin. interpolation of Run5 and Run6
Run8	-	-	-	lin. interpolation of Run1 and a concept-only-based PL2F model
Run9	term	term	PL2F	UMLS query expansion filtered by BRF
Run10	term	term	DIR	BRF

Table 3. BRF query expansion on P@10 on selected CLEF 2015 test set queries.

query ID	original query	expanded term	Run1	Run10
clef2015.test.1	many red marks on legs after traveling from us	striaestretch	0.1000	0.1000
clef2015.test.2	lump with blood spots on nose	mixx	0.3000	0.0000
clef2015.test.3	dry red and scaly feet in children	scaleness	0.7000	0.9000
clef2015.test.4	itchy lumps skin	healthdental	0.2000	0.1000
clef2015.test.5	whistling noise and cough during sleeping + children	ringining	0.6000	0.6000

- Run9 is similar to Run5, but the expansion terms are further filtered by relevance feedback, i.e. we only use the terms that also appeared in the initial retrieval.
- Run10 employs our own implementation of blind relevance feedback. We do initial retrieval using Run1, then from top 25 ranked documents. We add one term with the highest IDF in the collection into original query, then we do the retrieval again using Run1, see Section 4.3.

Table 4 shows our system performance on the CLEF 2014 test set. The difference between numbers in italics and bold is not statistically significant using the Wilcoxon test [7]. Run4 which is an interpolation between the first 3 runs gives the best P@10. Run8 also brings some improvement to the baseline system. Run5, which uses the same model with Run3, brings a slight improvement with the query expansion. However, Run6 decreases slightly in P@10 compared to unexpanded Run2. Run7 brings some improvement over the two other runs that it interpolated. In case of Run9, combining our query expansion with relevance feedback decreases the performance compared to Run5.

The results of our submitted runs are shown in Table 5. In terms of P@10 metric, Run1 and Run2 have the same P@10, but Run1 outperforms Run2 in terms of MAP, NDCG@10 and the number of relevant retrieved documents. We have improvement in Run4, which linearly interpolates Run1, Run2 and Run3. As in the training set, Run5 improves the P@10 when compared to Run3,

Table 4. System performance in monolingual task on the CLEF 2014 test set.

run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret	UNJ@10
Run1	**0.7680**	*0.7160*	0.7519	0.7206	*0.3919*	*2588*	14
Run2	*0.7000*	*0.6900*	*0.6872*	*0.6833*	*0.3832*	*2573*	12
Run3	0.7160	*0.6980*	0.7076	*0.6995*	**0.4300**	**2658**	9
Run4	0.7480	**0.7400**	0.7376	**0.7362**	0.4188	2637	1
Run5	0.7280	*0.7000*	0.7250	*0.7057*	0.4134	2628	8
Run6	0.7320	*0.6840*	0.7340	*0.7002*	*0.3849*	*2503*	24
Run7	0.7520	0.7200	0.7390	0.7204	0.4135	2642	11
Run8	0.7640	*0.7200*	**0.7538**	0.7239	*0.3953*	*2588*	14
Run9	0.7160	*0.6940*	0.7077	0.6977	*0.4069*	*2611*	15
Run10	0.7080	*0.6980*	*0.6728*	*0.6756*	*0.3793*	*2588*	**36**

Table 5. System performance in monolingual task on the CLEF 2015 test set.

run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret	UNJ@10
Run1	0.3970	0.3712	0.3352	0.3423	0.2353	**1703**	0
Run2	0.4061	0.3712	0.3399	0.3351	0.2236	1668	0
Run3	0.3818	0.3485	0.3136	0.3138	0.2095	1637	3
Run4	0.4121	0.3742	0.3424	0.3409	**0.2427**	1702	2
Run5	0.3970	0.3530	0.3290	0.3217	0.2046	1607	20
Run6	0.4030	0.3606	0.3439	0.3364	0.2123	1606	48
Run7	**0.4152**	**0.3803**	**0.3513**	**0.3465**	0.2188	1627	10
Run8	0.4061	0.3621	0.3385	0.3383	0.2369	1703	0
Run9	0.3970	0.3530	0.3287	0.3215	0.2045	1607	19
Run10	0.3273	0.3000	0.2604	0.2597	0.1919	1695	**121**

and Run6 does not bring any improvement over Run2. Run7, which is the best performing run, brings improvement over Run5 and Run6. Run3 has 3 unjudged documents in the first 10 ranked documents among 66 queries, so the shown metrics may not be accurate. BRF in Run10 does not bring overall improvement in P@10, but it does in some queries. BRF still does not guarantee to choose the best term to expand the query with (see Table 3).

5.2 Multilingual Task

In this task, we are given parallel queries in Arabic, Czech, Farsi, French, German, Italian and Portuguese and the goal is to design a retrieval system to find relevant documents to these queries from the English collection. For queries in Czech, French and German, we submitted 10 runs as follows (see also Table 7):

- Run1, Run2 and Run3 runs, we translate the queries using Khresmoi [11] then use Terrier's implementation of DIR, PL2F and LGD retrieval models respectively.
- Run4 interpolates Run1, Run2, and Run3 with parameters (0.57, 0.40, 0.03) respectively, tuned on the CLEF eHealth 2014 test set.

Table 6. Multilingual run description for AR, FA, IT and PT.

run ID	MT system	query	doc	model	details
Run1	Google	term	term	DIR	-
Run2	Google	term	term	PL2F	-
Run3	Google	term	term	LGD	-
Run4	Google	-	-	-	lin. interpolation of Run1, Run2, Run3
Run5	Bing	term	term	PL2F	-
Run6	Bing	term	term	DIR	-
Run7	Bing	-	-	LGD	-
Run8	-	-	-	-	lin. interpolation of Run5, Run6, Run7
Run9	Google	term	term	DIR	BRF
Run10	Google	term	term	DIR	lin. interpolation of Run1 and Run6

Table 7. Multilingual run description for CS, FR, and DE.

run ID	MT system	query	doc	model	details
Run1	Khresmoi	term	term	DIR	-
Run2	Khresmoi	term	term	PL2F	-
Run3	Khresmoi	term	term	LGD	-
Run4	-	-	-	-	lin. interpolation of Run1, Run2, Run3
Run5	Khresmoi	term	term	-	term weighting
Run6	Khresmoi	term	term	BRF	n documents and m terms
Run7	Google	term	term	DIR	-
Run8	Bing	term	term	DIR	-
Run9	Google	term	term	PL2F	-
Run10	-	-	-	-	lin. interpolation of Run7 and Run8

– Run5 uses Khresmoi to translate the queries into English, queries are expanded using term weighting algorithm as described in Section 4.5. After model tuning, we use as parameters $\lambda_1 = 4$, $\lambda_2 = 2$, $n = 1$, and $m = 30$.
– In Run6, we also use Khresmoi, then we use BRF to expand queries by adding one term from the top 25 documents, which has the highest IDF in the collection into original query. Both the initial and final retrieval steps use the DIR model.
– Run7, In this run we translate the queries using Google Translate, and use Terrier's implementation of DIR model
– Run8 is similar to Run7, but queries are translated using Bing translator.
– Run9, we use Google Translate to translate the queries into English then use Terrier's implementation of PL2F.
– Run10 interpolates Run7 and Run8, for Czech we use the parameters 0.92 and 0.08 respectively, for French 0.90, and 0.10 and for German 0.7, and 0.3.

For AR, FA, IT and PT, we submitted the following runs (see also Table 6):

– Run1, Run2 and Run3, we translate the queries using Google Translate and then use Terrier's implementation of DIR, PL2F, and LGD respectively.
– Run4 interpolates Run1, Run2 and Run3 with the parameters 0.56, 0.32, and 0.12 respectively.

Table 8. Sample of Google and Bing translations for query clef2015.test.2

Google	
FA	Highlights of the red spots on the nose
FR	bulge with blood stains on his nose
DE	tumor with bloody spots on the nose
IT	clot with blood stains on his nose
Bing	
FA	The mass highlighted with red spots on nose
FR	bump with blood stains on the nose
DE	Tumor with bloody points on the nose
IT	lump with bloodstains on the nose

Table 9. System performance in multilingual task against test set.

run ID	Arabic	Czech	Farsi	French	German	Italian	Portuguese	Monolingual
Run1	0.2727	**0.3318**	0.3258	0.3121	0.2859	0.3712	0.3500	0.3712
Run2	0.2621	0.2727	0.3227	0.3061	0.2562	0.3424	**0.3576**	0.3712
Run3	0.2591	0.3030	0.3242	0.3273	0.2766	0.3394	0.3364	0.3485
Run4	0.2621	0.3030	0.3273	0.3182	0.2828	**0.3727**	0.3485	0.3742
Run5	**0.3030**	0.2909	0.3045	0.2879	0.2703	0.3318	0.3182	0.3530
Run6	0.2894	0.3000	0.2864	0.2803	0.2437	0.3606	0.3333	0.3606
Run7	0.2924	**0.3318**	0.2803	**0.3682**	0.3545	0.3318	0.2985	**0.3803**
Run8	0.2879	0.2924	0.3000	0.3182	0.2985	0.3515	0.3318	0.3621
Run9	0.2439	0.2864	0.2727	0.3333	0.2924	0.3106	0.2924	0.3530
Run10	0.2924	0.3288	**0.3333**	**0.3682**	**0.3561**	**0.3727**	0.3379	0.3000

- Run5, Run6 and Run7, queries are translated using Bing translator, then retrieval is conducted using PL2F, DIR, and LGD respectively.
- Run8 interpolates Run5, Run6 and Run7 with the parameters 0.57, 0.40, and 0.03.
- Run9 uses Google Translate and BRF to expand queries using 25 document and 1 term and DIR model for both initial and final retrieval.
- Run10 interpolates Run1 and Run6 with the parameters 0.84 and 0.16.

The results for multilingual submission are shown in Table 9. The table shows P@10 values for all languages and the last column shows the result of our monolingual task for comparison. Linear interpolation between runs use Google Translate and Bing translator together with DIR model improved the results for Farsi, French, German and Italian. The Baseline in Italian is identical to monolingual one, other Italian runs are very close to monolingual runs and sometimes higher, anyway we have many unjudged documents in our Italian submission so results may differ after full assessment. Example of how Google and Bing translate one query in different languages is shown in Table 8.

5.3 Conclusion and future work

In this paper, we have described our participation in the CLEF eHealth 2015 Task 2. We used the Terrier IR platform to index the collection and conduct retrieval using different retrieval models and our own implementation of blind relevance feedback. We used MetaMap to annotate the documents in the collection and the queries with concepts and then built systems on concept levels, which improved the performance measured by P@10. Linear interpolation of runs conducted using different approaches, it improved P@10 when interpolating models use different retrieval models. Although query expansions did not bring improvement over the baseline, we got improvement for some individual queries. In some cases it also improved the performance compared to a system with the same retrieval model that only used the original query terms. We also submitted runs to multilingual task. We translated queries in the given languages into English and performed several experiments. The most promising results were obtained by linear interpolation of runs using different translation system for languages Farsi, French, German and Italian which use DIR model. We also used information from translation variants provided by the Khresmoi translation system (n-best lists). This did not bring overall improvement but it did for some individual queries. We believe that more thorough investigation should be done on terms selection algorithms to refine query expansion based approaches, so it will lead to better performance.

Acknowledgments

This research was supported by the Charles University Grant Agency (grant n. 920913), Czech Science Foundation (grant n. P103/12/G084), and the EU H2020 project KConnect (contract n. 644753).

References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proc AMIA Symp pp. 17–21 (2001)
2. Bodenreider, O.: The unified medical language system (umls): Integrating biomedical terminology (2004)
3. Clinchant, S., Gaussier, E.: Bridging language modeling and divergence from randomness models: A log-logistic model for ir. In: Azzopardi, L., Kazai, G., Robertson, S., Rger, S., Shokouhi, M., Song, D., Yilmaz, E. (eds.) Advances in Information Retrieval Theory, Lecture Notes in Computer Science, vol. 5766, pp. 54–65. Springer Berlin Heidelberg (2009)
4. Efthimiadis, E.N.: Query expansion. Annual review of information science and technology 31, 121–187 (1996)
5. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Mueller, H.: Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In: Proceedings of CLEF 2014 (2014)

6. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Nvol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2015. In: CLEF 2015 - 6th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS), Springer (September 2015)
7. Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 329–338. SIGIR '93, ACM, New York, NY, USA (1993)
8. Macdonald, C., Plachouras, V., He, B., Lioma, C., Ounis, I.: University of glasgow at webclef 2005: Experiments in per-field normalisation and language specific stemming. In: Peters, C., Gey, F., Gonzalo, J., Mller, H., Jones, G., Kluck, M., Magnini, B., de Rijke, M. (eds.) Accessing Multilingual Information Repositories, Lecture Notes in Computer Science, vol. 4022, pp. 898–907. Springer Berlin Heidelberg (2006)
9. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006) (2006)
10. Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G.J., Lupu, M., Pecina, P.: Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In: CLEF 2015 Online Working Notes. CEUR-WS (2015)
11. Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlaváčová, J., Jones, G., Kelly, L., Leveling, J., Mareček, D., Novák, M., Popel, M., Rosa, R., Tamchyna, A., Urešová, Z.: Adaptation of machine translation for multilingual information retrieval in the medical domain. Artificial Intelligence in Medicine (2014)
12. Saleh, S., Pecina, P.: Cuni at the share/clef ehealth evaluation lab 2014. Proceedings of the ShARe/CLEF eHealth Evaluation Lab 1180, 226–235 (2014)
13. Smucker, M.D., Allan, J.: An investigation of dirichlet prior smoothings performance advantage. Tech. rep., Tech. Rep. IR-445, University of Massachusetts (2005)
14. Stanton, I., Ieong, S., Mishra, N.: Circumlocution in diagnostic medical queries. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 133–142. ACM (2014)
15. U.S. National Library of Medicine: UMLS reference manual (2009), metathesaurus. Bethesda, MD, USA
16. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS) 22(2), 179–214 (2004)
17. Zuccon, G., Koopman, B., Palotti, J.: Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In: Advances in Information Retrieval (ECIR 2015). pp. 562–567. Springer (2015)