Term Selection for Query Expansion in Medical Cross-Lingual Information Retrieval

Shadi Saleh(✉) and Pavel Pecina

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

Abstract. We present a method for automatic query expansion for cross-lingual information retrieval in the medical domain. The method employs machine translation of source-language queries into a document language and linear regression to predict the retrieval performance for each translated query when expanded with a candidate term. Candidate terms (in the document language) come from multiple sources: query translation hypotheses obtained from the machine translation system, Wikipedia articles and PubMed abstracts. Query expansion is applied only when the model predicts a score for a candidate term that exceeds a tuned threshold which allows to expand queries with strongly related terms only. Our experiments are conducted using the CLEF eHealth 2013–2015 test collection and show significant improvements in both cross-lingual and monolingual settings.

1 Introduction

In Cross-lingual Information Retrieval (CLIR), search queries are formulated in a language which differs from the language of documents. Machine Translation (MT) of queries into the document language is a common method which reduces this task into monolingual retrieval [19]. In this work, we tackle the *vocabulary mismatch problem* which occurs when MT fails to select the most effective query translation option and subsequently, a term-matching IR system fails to retrieve relevant documents because the terms in the translated query and terms in the relevant documents do not match.

The proposed method is based on a simple linear regression model that predicts the retrieval performance for each candidate expansion term when combined with a query translated by a Statistical Machine Translation (SMT) system. The model features are obtained from the SMT system, external document sources (Wikipedia, PubMed) and information from the document collection. The model is used to score each term from a candidate pool and those scored above a (pre-trained) threshold are automatically added to the translated query. As a result, the queries are expanded with strong candidates only. If no strong candidates are available, the queries remain unchanged. This prevents performance drop caused by adding irrelevant terms to the query.

L. Azzopardi et al. (Eds.): ECIR 2019, LNCS 11437, pp. 507–522, 2019.
https://doi.org/10.1007/978-3-030-15712-8_33

The work presented in this paper is focused on cross-lingual retrieval in the domain of medicine and health. The experiments were conducted using the CLEF eHealth 2013–2015 IR collection. The method, however, is domain-independent and can be used in monolingual retrieval too (after excluding the cross-lingual features). Our results demonstrate a significant improvement over the baseline system which exploits plain query translation using a domain-adapted SMT system. In the monolingual setting, our model significantly outperforms both the monolingual baseline system (no expansion) and the standard Kullback-Leiber divergence (KLD) method for automatic query expansion.

2 Related Work

2.1 Query Expansion

Web search user queries tend to be short. The average web search query length, as reported by Gabrilovich et al. [9], is about 2.5 terms. The information represented in these terms might be too brief and/or vague. This is considered to be a challenge for IR systems that follow the term-matching approach, since they fail to find relevant documents which do not contain the terms specified in the query. Query expansion (QE) can be done automatically, or by interaction with users (e.g. selecting one or more terms to be added to the query), which is known as interactive query expansion [12]. In this study, we will focus on automatic query expansion.

Blind Relevance Feedback (BRF) is one of the most popular techniques for QE, also known as pseudo-relevance feedback [33]. First, an initial retrieval is conducted using the base query and top m ranked documents are selected as a source for term candidates. Then each term in these documents is scored using some approaches like a combination of its frequency (TF) in these documents and its inverse document frequency (IDF) in the collection. Finally, the highest scored m terms are added to the base query and a final retrieval is done. However, there is a risk when following this approach because one or more of these m documents might be irrelevant; thus, adding terms from these documents might drift the information away from the intended one. QE can have significant improvement on one of the main evaluation metrics (such as MAP (mean average precision), precision at 10 documents, or recall) and degrades the others; thus, the use of QE should consider the context of the IR application when using query expansion [13]. Pal et al. [26] employed WordNet to weight a candidate term and measure its usefulness for expansion. They leveraged the similarity score of the top retrieved documents using BRF assumption, and excluded terms from WordNet which do not appear in these documents. They calculated different similarity scores between the query term and the candidate term based on term distribution in the document collection. Then they linearly combined these scores to select the weights of the expansion terms. This approach brought an improvement over the use of base queries on multiple TREC collections. Ermakova and Mothe [8] used local context analysis by choosing terms which surround query terms from documents that are retrieved from the initial retrieval. They assumed that

document terms which appear closely to query terms are more likely to be good candidates for expansion. They experimented their approach on TREC Ad-Hoc track datasets from three years (1997–1999)[1] and the WT10G dataset [5]. Cao et al. [3] showed that when QE is based only on term distribution, it can not distinguish good terms, which will improve the IR performance, and bad terms which will harm it. They presented a classification model that is integrated into a BRF method. It uses features from the collection to predict the usefulness of the expansion terms and select only the good ones.

2.2 Query Expansion in Cross-Lingual Information Retrieval

Cross-lingual Information Retrieval (CLIR) enables users to search in a collection that is different than their language. In order to conduct a retrieval that is based on term-matching, both documents and queries should be represented in one language [24]. Query-translation is the most common approach in CLIR, wherein user queries are translated into the document language and then a monolingual retrieval is conducted. Popular machine translation techniques struggle translating short queries because of the lack of linguistic information that is required to solve ambiguity, which eventually causes information loss in the translated queries [32]. Query expansion in CLIR helps to solve this issue by adding relevant terms to the translated queries. Chandra and Dwivedi [4] used Google Translate[2] to translate queries from Hindi into English in the FIRE 2008 dataset. Then, they did an initial retrieval using the translated queries and created a set of candidate terms. They applied different methods for term selection. They found that adding the term which has the lowest frequency in the top 3 ranked documents gave the best result.

2.3 Query Expansion in Medical Information Retrieval

Query expansion in the medical domain is considered to be a more difficult task. Approaches which work on the general domain might not work perfectly when applied in the medical domain. Nikoulina et al. [21] reported that simply merging the top 5 scored translation hypotheses (as a special QE approach) to create queries in the CLIR task outperformed the baseline system in the general domain data. However, the same approach did not work when tested on the medical domain [34]. Kullback-Leiber Divergence (KLD) for query expansion (explained in Sect. 3.4) failed to outperform the baseline system (using initial queries) during the CLEF 2011 medical retrieval task [16]. Choi and Choi [6] used Google Translate to translate the queries into English (from Czech, French and German) during their participation in the CLEF eHealth 2014 CLIR task [10]. Then, they annotated each query with medical concepts using MetaMap [2], and the top scored concepts were added to the original query. Finally, they weighted the original query and the expanded query with 0.9 and 0.1 respectively. Query

[1] https://trec.nist.gov.
[2] http://translate.google.com.

expansion approach outperformed their baseline system by 18% for Czech, 4% for German and 4% for French. Liu and Nie [18] participated in the monolingual task of CLEF eHealth 2015 [11], and presented a system which expanded queries with UMLS [15] concepts and terms extracted from Wikipedia articles. Authors claim that Wikipedia abstracts are similar to the way that users pose queries (more generic), while the titles of Wikipedia articles contain medical terms. However, using only Wikipedia to expand the queries did not help. Only a system that combined Wikipedia with MetaMap [2] improved the baseline system. Employing MeSH (Medical Subject Heading)[3] for QE was investigated thoroughly. Wright et al. [39] presented a simple method that expands queries with five synonyms from MeSH. Nunzio and Moldovan [23] expanded a query with one MeSH term that is related to the base query, when there was more than one MeSH candidate term, they created multiple expanded queries, then for each expanded query, they conducted retrieval and merged the retrieved documents by different approaches like averaging document scores or summing them.

Word embeddings became a well-known technique in representing terms in high dimensional vectors. This allows estimating semantic and syntactic similarities between terms. Term vectors can be generated using famous models like word2vec [20] and Glove [31]. The main idea is to expand the query with terms that are semantically related and appear in a position close to the query terms [22,40,41]. Multiple researchers confirmed that embeddings models that are trained on medical data like PubMed articles are not significantly better than those which are trained on general domain data, such as news [42].

3 Experimental Setting

3.1 Test Collection

The training and evaluation data used in our work is described in [36]. It is adopted from the IR tasks of the CLEF eHealth Lab series 2013–2015 [10,27,38]. The **document collection** is taken from the IR task in 2015 eHealth Task 2: User-Centred Health Information Retrieval [11], and consists of about 1.1 million web pages (documents) crawled from various medical-domain websites. We cleaned the documents using HTML-Strip Perl module[4]. We did not perform any preprocessing (stemming, lemmatisation) since it showed degrading in our previous experiments. The **query set** contains 166 items used during the three years of the CLEF eHealth IR tasks as test queries. The queries were originally created in English and then manually translated into seven languages (Czech, French, German, Spanish, Swedish, Polish, and Hungarian) to allow cross-lingual experiments. As proposed in [36], we used 100 queries for training the model parameters (feature weights, term selection threshold, IR model parameters) and 66 queries for testing (measuring retrieval performance). See Table 1 for query examples and [36] for additional details.

[3] https://www.nlm.nih.gov/mesh.
[4] http://search.cpan.org/dist/HTML-Strip/Strip.pm.

Table 1. Query samples from the extended CLEF eHealth test collection.

Query id	Query title
2013.02	*Facial cuts and scar tissue*
2013.41	*Right macular hemorrhage*
2013.30	*Metabolic acidosis*
2014.04	*Anoxic brain injury*
2014.21	*Renal failure*
2014.17	*Chronic duodenal ulcer*
2015.08	*Cloudy cornea and vision problem*
2015.59	*Heavy and squeaky breath*
2015.48	*Cannot stop moving my eyes medical condition*

3.2 Machine Translation of Queries

In our experiments, we employ a statistical machine translation (SMT) system for query translation. This system was developed under the Khresmoi project [7]. It is based on Moses [17], a state-of-the-art phrase-based system, trained on a combination of in-domain (EMEA, PatTR, COPPA, UMLS, etc.) and general-domain (e.g., EuroParl, JRC Acquis and News Commentary corpus) resources. The system employs several special features [28] that allow for optimal translation of medical search queries. For an input text, an SMT system produces a list of translation hypotheses ranked by their translation quality, the best one is referred to as *1-best* translation. In this research, we employ seven SMT models to translate queries from Czech, German, French, Hungarian, Polish, Spanish and Swedish into English.

3.3 Baseline Retrieval System

Our baseline CLIR system is designed as follows: The non-English queries are translated into English (using *1-best* translations produced by the SMT system described above) which reduces the CLIR task into a monolingual IR task. For indexing and retrieval, we use Terrier, an open source search engine [25], and its implementation of the language model with Bayesian smoothing with Dirichlet prior [37]. The default value of the smoothing parameter μ is set to 2500 (this has been proven to work well in our previous work [35]). For comparison purposes, we also report results of a monolingual system which employs the English (reference) translations of the queries. It sets a theoretical maximum which a CLIR system can reach when a query translation is completely correct.

Retrieval results are evaluated by the standard *trec_eval* tool[5] using two evaluation metrics: precision at top 10 documents (*P@10*) which is used as the main evaluation measure in our work, and preference-based measure *BPREF* which

[5] http://trec.nist.gov/trec_eval.

considers if the judged relevant documents are ranked above the judged irrelevant ones. Statistical significance tests are performed using the paired Wilcoxon signed-rank test [14], with α set to 0.05.

3.4 KLD Query Expansion

To compare our query expansion method with other approaches, we report the results of Kullback-Leiber Divergence (KLD) for query expansion as it is implemented in Terrier [1]. In KLD, the top n ranked documents (pseudo-relevant documents) are retrieved using the base query, then each term in these documents is scored by the equation below, where $P_r(t)$ is the probability of term t in the pseudo-relevant documents, and $P_c(t)$ is the probability of term t in the document collection c. Finally the top m scored terms are added to the base query and a final retrieval is done using the new expanded query.

$$Score(t) = P_r(t) \cdot log\left(\frac{P_r(t)}{P_c(t)}\right)$$

We set n and m to 7 and 2 respectively by grid-search tuning (using the monolingual English training queries) as shown in Fig. 3.

4 Term Selection Model

The proposed CLIR query expansion method is performed in four steps. First, a set of candidate terms (candidate pool) are collected from various sources. Second, each term from the candidate pool is assigned a vector of features describing its potential to identify relevant documents. Third, the features are combined in a regression model to score each candidate term. Finally, terms with scores exceeding a given threshold are selected to expand the query. Figure 1 shows the architecture of our presented model in detail. The following sections explain the term selection process.

4.1 Candidate Pool

Three sources of candidate terms are considered in our experiments:

Machine Translation (MT). For each source query, we collect all the terms from the 100 highest-scored translation hypotheses as produced by the SMT system. The motivation behind this is based on the fact that the translation hypotheses might contain alternative translations of query terms (synonyms/other related terms).

English Wikipedia (Wiki). The base query (*1-best* translation) is used to retrieve articles from an indexed Wikipedia collection. Only titles and abstracts are used to build the index following the same settings as in our baseline model. The titles of the top 10 ranked retrieved articles are added to the expansion pool.

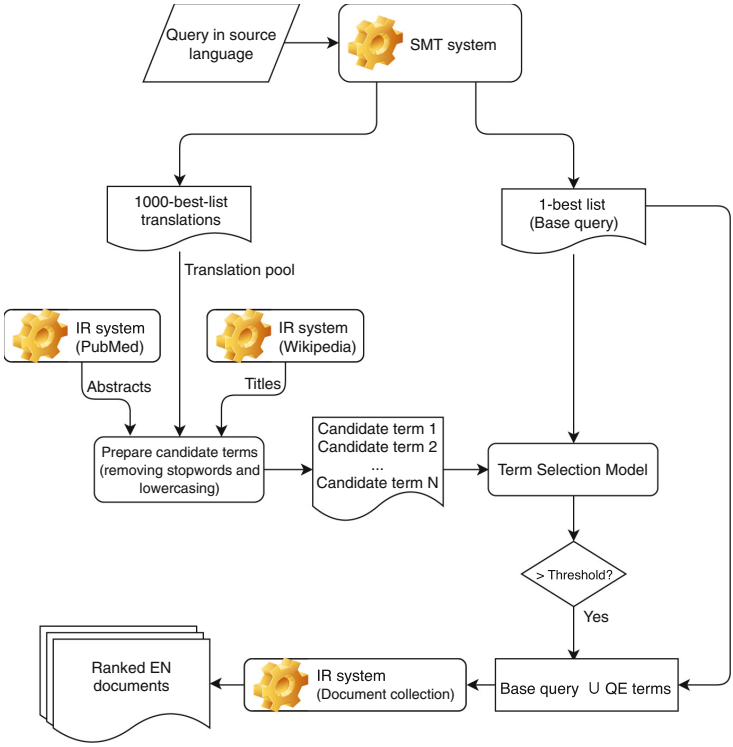

Fig. 1. System architecture overview.

The use of Wikipedia titles is to tackle the challenge when users pose a query in the medical domain using non-medical terms by describing the symptoms of a specific disease. Disease names usually appear in the title and their symptoms are described in the abstract [18].

PubMed. We also enrich the candidate pool with terms from the PubMed articles [30] following the settings as the Wikipedia articles. PubMed articles (both abstracts and titles) are indexed, then the top 10 ranked articles are retrieved using the *1-best* translation as a base query and added to the candidate pool.

4.2 Feature Set

Each term from the candidate pool is described by a set of features designed to reflect the term's usefulness for expansion:

IDF, which is calculated in the document collection.

Translation pool frequency, i.e. the frequency of the term in the 100 highest-scored translation hypotheses as produced by the SMT system. When a term

appears in multiple hypotheses, this means that the probability of being a relevant translation to one of the terms in the original query is high. This feature is excluded in our monolingual QE model.

Wikipedia frequency, i.e. the frequency of the term in the top 10 Wikipedia articles retrieved from the Wikipedia index using the *1-best* translation as a base query.

Retrieval Status Value (RSV), which is the difference of the RSV value (the score of the Dirichlet retrieval model) of the highest-ranked document retrieved using the base query, and the RSV value of the highest-ranked document retrieved using the base query expanded by the candidate term. This feature tells us the contribution of the candidate term to the RSV score.

Query similarity, i.e. an average similarity between a candidate term t_m and the query term obtained using a pre-trained model of *word2vec* embeddings on 25 millions articles from PubMed[6]. First, we get the word embeddings for each term in the original query and we sum those embeddings to get a vector that represents the entire query. Then we take the embeddings for t_m, and calculate the cosine similarity between the query vector and the t_m vector. It is important to point out here that choosing terms that are similar to each term of the query caused significant drift in the information need, for example: *mother* was suggested as a similar term to *baby*, and *white* as a similar term to *black*.

Co-occurrence frequency, the co-occurrences of a candidate term t_m and the query terms $t_i \in Q$ indicates how likely t_m is related to the original query Q, we sum up the co-occurrence frequency for each term in query Q and the candidate term t_m in all documents d_j in the collection C, as shown below:

$$co(t_m, Q) = \sum_{d_j \in C, t_i \in Q} tf(d_j, t_i) tf(d_j, t_m)$$

Term frequency, first, we perform retrieval from the collection using a query that is constructed from the *1-best* translation, then we calculate the term frequency of a candidate term t_m in the top 10 ranked documents from the retrieval result.

UMLS frequency, this feature represents how many times a term appeared in the UMLS lexicon [15], as an attempt to give more weight to the medical terms.

4.3 Regression Model

The term selection model is based on linear regression. Training instances are candidate terms for the training queries after translating those queries from all seven languages into English. Each term t from a candidate pool of a given query is assigned a value computed as the difference of P@10 obtained by the baseline query (*1-best-list* translation) and P@10 obtained by the expanded baseline

[6] https://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/DATASET/.

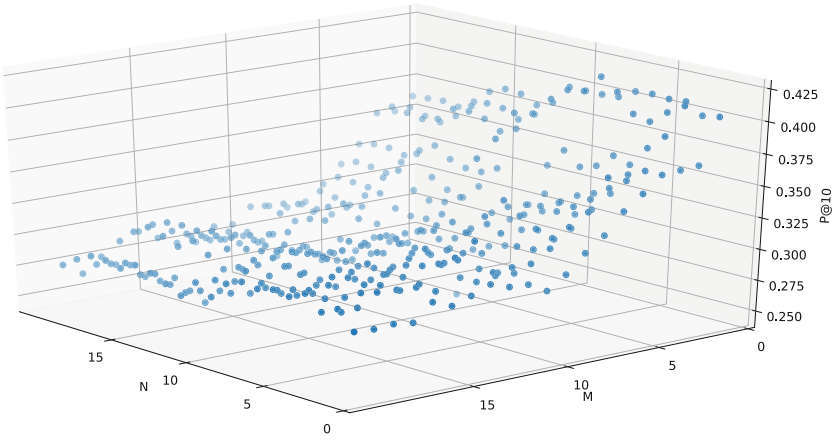

Fig. 2. Tuning KLD parameters, number of documents (N) and number of expansion terms (M) on the monolingual queries.

query with the term t. Expansion terms increasing P@10 for the given query are assigned positive values, terms decreasing P@10 are assigned negative values, and terms without any effect on the retrieval performance for that query are assigned zero. The purpose is to expand the queries with terms that can improve the performance, rather than terms that harm the performance. The feature vectors are centered and reduced. This is done independently on each feature on the training set, then we use the scaler coefficient to standardise the test set. We consider P@ difference as the objective function, and we use the proposed feature set to train the model. Linear Regression (LR) models the relationship between the dependent variable (P@10 in our case) and the regressors x (term feature values).

We use ordinary least squares linear regression as it is implemented in *scikit* package [29]. There might be one or more good candidate terms for expansion. To select these terms, we set a threshold value for the predicted score. The threshold value is tuned on the training set for all languages as shown in Fig. 3. All terms which have a score equal or higher than the threshold are added to the base query. This allows us to avoid expanding queries with irrelevant terms.

5 Experiments and Results

Results of all experiments for the seven languages are presented in percentages in Table 2 (in terms of P@10) and Table 3 (in terms of BPREF). For each language, the underlined score denotes the best result, and the scores in bold refer to results which are not significantly different (given the Wilcoxon signed-rank test) from the best (underlined) score. TS refers to the proposed QE technique based on term selection, and the text in the brackets denote the candidate term sources: machine translation (MT) hypotheses, Wikipedia titles (Wiki),

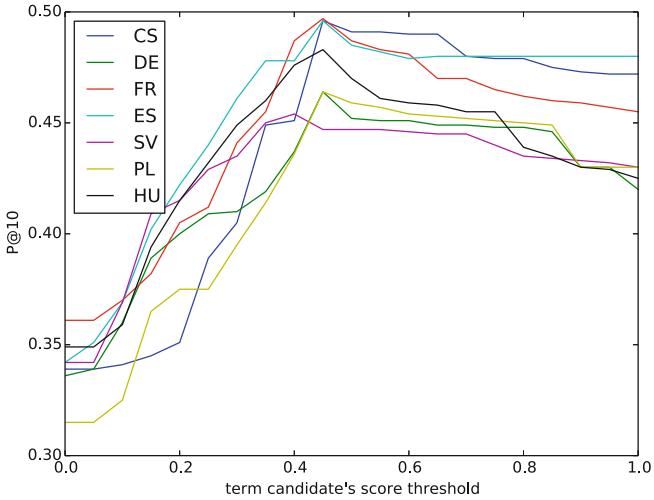

Fig. 3. Tuning threshold for term candidate selection based on their predicted scores.

and PubMed articles (PubMed). The monolingual experiment (exploiting the reference English queries) sets a theoretical upper-boundary for the results of the CLIR experiments. It is 53.03 in terms of P@10 and 39.94 in terms of BPREF. (These values hold for all the languages since the reference translations of the source queries are the same). Monolingual+KLD refers to the result of the KLD-based query expansion applied to the reference translations of the queries. In terms of P@10, the result went down substantially. This can be explained because either the indexed documents are not good enough as a source of candidate expansion terms, or because there is no criteria to prevent

Table 2. Experiment results in terms of P@10 (percentage)

System/query language	CS	FR	DE	ES	HU	PL	SV
Monolingual	53.03	53.03	53.03	53.03	53.03	53.03	53.03
+KLD	48.18	48.18	48.18	48.18	48.18	48.18	48.18
+TS(PubMed)	55.76	55.76	55.76	55.76	55.76	55.76	55.76
Baseline	47.27	48.03	44.24	46.97	45.91	42.12	40.00
+KLD	39.85	45.76	38.33	42.12	42.12	39.24	36.36
+TS(MT)	47.42	48.03	43.03	46.82	46.21	**42.42**	**41.52**
+TS(Wiki)	44.85	44.70	43.03	43.18	**47.12**	41.06	39.70
+TS(PubMed)	50.15	47.12	43.33	45.30	43.48	37.58	36.52
+TS(MT ∪ Wiki)	**52.58**	**49.55**	**47.12**	**48.33**	**47.88**	**42.42**	**41.52**
+TS(MT ∪ PubMed)	50.30	**48.79**	**45.45**	**48.03**	42.73	38.48	34.85
+TS(MT ∪ Wiki ∪ PubMed)	**52.12**	**48.94**	**45.45**	**47.42**	**47.58**	**43.18**	**41.21**

Table 3. Experiment results in terms of BPREF (percentage)

System/query language	CS	FR	DE	ES	HU	PL	SV
Monolingual	39.94	39.94	39.94	39.94	39.94	39.94	39.94
+KLD	41.22	41.22	41.22	41.22	41.22	41.22	41.22
+TS(PubMed)	41.41	41.41	41.41	41.41	41.41	41.41	41.41
Baseline	36.79	35.65	35.38	37.24	37.08	33.77	20.94
+KLD	36.21	**38.34**	34.84	**39.64**	36.59	**34.33**	32.11
+TS(MT)	36.80	35.49	35.64	37.05	37.03	**33.92**	**33.38**
+TS(Wiki)	36.82	36.10	36.09	36.17	**38.77**	**33.82**	**34.23**
+TS(PubMed)	**39.16**	**38.14**	**39.15**	**39.47**	36.87	33.51	**33.78**
+TS(MT ∪ Wiki)	**40.49**	**38.82**	**40.86**	37.93	36.95	**33.92**	**33.38**
+TS(MT ∪ PubMed)	38.90	**37.63**	36.09	**38.87**	36.57	**34.16**	**33.67**
+TS(MT ∪ Wiki ∪ PubMed)	**40.21**	**37.15**	36.02	37.93	37.70	**33.86**	32.98

Table 4. Precision (percentage) of selected terms manually checked by a medical expert (first raw) and with respect to the terms that appeared in the reference English queries (second raw)

Measure/query language	CS	FR	DE	ES	HU	PL	SV
Precision w.r.t. manual judgments	87.60	89.33	90.84	87.50	96.43	90.91	87.50
Precision w.r.t. reference translations	21.49	14.04	13.74	25.00	21.43	36.36	12.50

expanding some queries with low scored term candidates. The proposed term selection (TS) method applied to the monolingual retrieval (using PubMed only as a source of candidate terms) seems to be much more promising. The P@10 score is as high as 55.76. This system improved the results for 13 queries and degraded 4 queries. The rest of the queries did not change due to the low scores of candidate terms as predicted by the model. In terms of BPREF, both KLD and TS bring a small improvement which is not statistically significant. P@10 scores of the CLIR baseline systems (exploiting *1-best* translation) range between 40.00 and 48.03 depending on the query language. The KLD-based expansion in CLIR brings the scores even lower (36.36–45.76) which is in line with the monolingual expansion experiments. Though, for some queries (10 on average), P@10 improved, and results for more queries (20 on average) degraded. The proposed term selection experiments show consistent improvement over the baseline. The best system uses terms from MT and Wiki for expansion. Samples of queries that are improved by this system are shown in Table 5. The CLIR system improved 21 queries in Czech, 18 in French, 14 for German and 11 in Spanish, 10 queries in Hungarian, 2 queries in Polish, and 3 queries in Swedish. While it degraded 11 queries in Czech, 12 in French, 11 in German, 4 in Spanish, 5 queries in Hungarian, 2 queries in Polish, and 2 queries in Swedish. The performance of the rest of the queries did not change. The average result in Czech is very close to

Table 5. Examples of queries from different systems including Mono (*ref*), Baseline (*base*), and expansion terms to the baseline query (*QE*). The scores in parentheses refer to query P@10 scores in percentages

Query: 2015.18 (Czech)	**Query: 2013.11 (German)**
ref: poor gait and balance with shaking (50.00)	*ref:* chest pain and liver transplantation (50.00)
base: bad posture and balance with tremor (60.00)	*base:* breast pain and liver transplantation (10.00)
QE: poor balanced shaking (70.00)	*QE:* chest hepatic graft thoracic (40.00)
Query: 2014.21 (French)	**Query: 2014.11 (Spanish)**
ref: white patchiness in mouth (10.00)	*ref:* Diabetes type 1 and heart problems (40.00)
base: renal impairment (00.00)	*base:* type 1 diabetes and heart problems (40.00)
QE: kidney disease function dysfunction failure insufficiency deficiency poor (30.00)	*QE:* cardiac disease (60.00)

Table 6. Examples of queries degraded in the QE approach (*QE*) with respect to Mono (*ref*), Baseline (*base*), the scores in parentheses refer to query P@10 scores in percentages

Query: 2013.41 (Czech)	**Query: 2015.65 (German)**
ref: right macular hemorrhage (60.00)	*ref:* weird brown patches on skin (10.00)
base: amacular bleeding right (70.00)	*base:* strange brown spots on the skin (40.00)
QE: hemorrhage haemorrhage side blood (30.00)	*QE:* spot patches cutaneous patch (10.00)
Query: 2013.41 (French)	**Query: 2014.31 (Spanish)**
ref: right macular hemorrhage (60.00)	*ref:* Acute renal failure (80.00)
base: macular hemorrhage right eye (80.00)	*base:* acute renal failure (80.00)
QE: eyes haemorrhage hemorrhagic bleeding (50.00)	*QE:* kidney disease (60.00)

the monolingual performance. Table 6 shows examples of queries that degraded in the TS(MT ∪ Wiki) system.

For further analysis of the expansion quality, we report in Table 4 the percentage of relevant expansion terms calculated by the two methods. In the first method, we provided a medical doctor with query titles, their narratives (to understand the topic for each query) and the expansion terms as suggested by the TS(MT ∪ Wiki) system. We asked them to identify the expanded terms whether they are relevant to the topic or not. The second method is an automatic evaluation that is done by checking if the expansion terms exist in the reference queries. For example in the Czech system, 78.51% of the expansion terms did not appear in the reference query; however, we could not tell if they are relevant or not. In contrast, when checked by a medical doctor, it appeared that only 12.4% of them are irrelevant to the topic.

6 Conclusion

In this work, we have addressed the problem of automatic query expansion for cross-lingual information retrieval as an attempt to improve the information represented in user queries. The presented model is based on machine translation of queries from a source language to a document language and machine learning to predict relevant expansion terms from a rich source of term candidates. The feature set is based on information derived from the collection, external resources (Wikipedia and PubMed articles), and word-embeddings. Fine-tuning the threshold value of the term predicted score helps to expand queries only when there is a good candidate. This prevents expanding queries when candidate terms are irrelevant to the topic. Our evaluation has shown that our approach helps significantly improving the baseline system in cross-lingual and mono-lingual settings.

Acknowledgments. This work was supported by the Czech Science Foundation (grant n. 19-26934X).

References

1. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness, and selective application of query expansion. In: McDonald, S., Tait, J. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 127–137. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24752-4_10
2. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of AMIA Symposium, pp. 17–21 (2001)
3. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 243–250. ACM, New York (2008)
4. Chandra, G., Dwivedi, S.K.: Query expansion based on term selection for Hindi-English cross lingual IR. J. King Saud Univ. Comput. Inf. Sci. (2017)
5. Chiang, W.T.M., Hagenbuchner, M., Tsoi, A.C.: The wt10g dataset and the evolution of the web. In: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, WWW 2005, pp. 938–939. ACM, New York (2005)
6. Choi, S., Choi, J.: Exploring effective information retrieval technique for the medical web documents: Snumedinfo at clefehealth2014 task 3. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, vol. 1180, pp. 167–175. CEUR-WS.org, Sheffield (2014)
7. Dušek, O., Hajič, J., Hlaváčová, J., Novák, M., Pecina, P., Rosa, R., et al.: Machine translation of medical texts in the Khresmoi project. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 221–228, Baltimore (2014)
8. Ermakova, L., Mothe, J.: Query expansion by local context analysis. In: Conference francophone en Recherche d'Information et Applications (CORIA 2016), pp. 235–250. CORIA-CIFED, Toulouse (2016)
9. Gabrilovich, E., Broder, A., Fontoura, M., Joshi, A., Josifovski, V., Riedel, L., Zhang, T.: Classifying search queries using the web as a source of knowledge. ACM Trans. Web **3**(2), 5 (2009)

10. Goeuriot, L., et al.: ShARe/CLEF eHealth evaluation lab 2014, Task 3: user-centred health information retrieval. In: Proceedings of CLEF 2014, pp. 43–61. CEUR-WS.org, Sheffield (2014)
11. Goeuriot, L., et al.: Overview of the CLEF eHealth evaluation lab 2015. In: Mothe, J., et al. (eds.) CLEF 2015. LNCS, vol. 9283, pp. 429–443. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_44
12. Harman, D.: Towards interactive query expansion. In: Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 321–331. SIGIR 1988, ACM, New York (1988)
13. Harman, D.: Information retrieval. In: Relevance Feedback and Other Query Modification Techniques, pp. 241–263. Prentice-Hall Inc., Upper Saddle River (1992)
14. 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. ACM, Pittsburgh (1993)
15. Humphreys, B.L., Lindberg, D.A.B., Schoolman, H.M., Barnett, G.O.: The unified medical language system. J. Am. Med. Inform. Assoc. **5**(1), 1–11 (1998)
16. Kalpathy-Cramer, J., Muller, H., Bedrick, S., Eggel, I., De Herrera, A., Tsikrika, T.: Overview of the clef 2011 medical image classification and retrieval tasks. In: CLEF 2011 - Working Notes for CLEF 2011 Conference, vol. 1177. CEUR-WS (2011)
17. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions, pp. 177–180, Stroudsburg (2007)
18. Liu, X., Nie, J.: Bridging layperson's queries with medical concepts - GRIUM @CLEF2015 eHealth Task 2. In: Working Notes of CLEF 2015 Conference and Labs of the Evaluation forum, vol. 1391. CEUR-WS.org, Toulouse (2015)
19. McCarley, J.S.: Should we translate the documents or the queries in cross-language information retrieval? In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 208–214, College Park (1999)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS 2013, vol. 2, pp. 3111–3119. Curran Associates Inc., Red Hook (2013)
21. Nikoulina, V., Kovachev, B., Lagos, N., Monz, C.: Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 109–119, Stroudsburg (2012)
22. Nogueira, R., Cho, K.: Task-oriented query reformulation with reinforcement learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 574–583 (2017)
23. Nunzio, G.M.D., Moldovan, A.: A study on query expansion with mesh terms and elasticsearch. IMS unipd at CLEF ehealth task 3. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, 10–14 September 2018. CEUR-WS, Avignon (2018)
24. Oard, D.W.: A comparative study of query and document translation for cross-language information retrieval. In: Farwell, D., Gerber, L., Hovy, E. (eds.) AMTA 1998. LNCS (LNAI), vol. 1529, pp. 472–483. Springer, Heidelberg (1998). https://doi.org/10.1007/3-540-49478-2_42

25. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 517–519. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31865-1_37
26. Pal, D., Mitra, M., Datta, K.: Improving query expansion using wordnet. J. Assoc. Inf. Sci. Technol. **65**(12), 2469–2478 (2014)
27. Palotti, J.R., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G.J., Lu pu, M., Pecina, P.: CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving information about medical symptoms. In: CLEF (Working Notes), pp. 1–22. Springer, Heidelberg (2015)
28. Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlavářová, J., Jones, G.J., et al.: Adaptation of machine translation for multilingual information retrieval in the medical domain. Artif. Intell. Med. **61**(3), 165–185 (2014)
29. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
30. Peng, Y., Wei, C.H., Lu, Z.: Improving chemical disease relation extraction with rich features and weakly labeled data. J. Cheminformatics **8**(1), 53 (2016)
31. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
32. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-based cross-language information retrieval: problems, methods, and research findings. Inform. Retrieval **4**(3–4), 209–230 (2001)
33. Rocchio, J.J.: Relevance feedback in information retrieval. The SMART Retrieval Syst. Exp. Autom. Doc. Process. 313–323 (1971)
34. Saleh, S., Pecina, P.: Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In: Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 54–66. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_5
35. Saleh, S., Pecina, P.: Task3 patient-centred information retrieval: Team CUNI. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum. CEUR-WS.org, Evora (2016)
36. Saleh, S., Pecina, P.: An Extended CLEF eHealth Test Collection for Cross-lingual Information Retrieval in the medical domain. In: Advances in Information Retrieval - 41th European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings. Lecture Notes in Computer Science, Springer (2019)
37. Smucker, M.D., Allan, J.: An investigation of Dirichlet prior smoothing's performance advantage. University of Massachusetts, Technical report (2005)
38. Suominen, H., et al.: Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 212–231. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40802-1_24
39. Wright, T.B., Ball, D., Hersh, W.: Query expansion using mesh terms for dataset retrieval: OHSU at the biocaddie 2016 dataset retrieval challenge. J. Biol. Databases Curation 2017, Database (2017)
40. Zamani, H., Croft, W.B.: Embedding-based query language models. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, ICTIR 2016, pp. 147–156. ACM, New York (2016)

41. Zamani, H., Croft, W.B.: Relevance-based word embedding. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 505–514. SIGIR 2017. ACM, New York (2017)
42. Zuccon, G., Koopman, B., Bruza, P., Azzopardi, L.: Integrating and evaluating neural word embeddings in information retrieval. In: Proceedings of the 20th Australasian Document Computing Symposium, p. 12. Stroudsburg (2015)