CUNI team: CLEF eHealth Consumer Health Search Task 2018

Shadi Saleh and Pavel Pecina

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

Abstract. In this paper, we present our participation in CLEF Consumer Health Search Task 2018, mainly, its monolingual and multilingual subtasks: IRTask1 and IRTask4. In IRTask1, we use language-model based retrieval model, vector-space model and Kullback-Leiber divergence query expansion mechanism to build our runs. In IRTask4, we submitted 4 runs for each language of Czech, French and German. We follow query-translation approach in which we employ a Statistical Machine Translation (SMT) system to get a ranked list of translation hypotheses in English. We use this list for two systems: the first one uses 1-best-list translation to construct queries, and the second one uses a hypotheses reranker to select the best translation (in terms of retrieval performance) to construct queries. We also present our term reranking model for query expansion, in which we deploy feature set from different resources (the document collection, Wikipedia articles, translation hypotheses). These features are used to train a logistic regression model that can predict the performance when a candidate term is added to a base query.

Keywords: Multilingual information retrieval, statistical machine translation, hypotheses reranking, term reranking

1 Introduction

Internet searches for medical topics had been increasing recently, and have gotten the attention of information retrieval researchers. Fox [3] reported that about 80% of Internet users in the United States look for medical information online. The main challenge in the medical information retrieval systems that people with different experience express their information need in different way [14]. Laypeople express their medical information need using non-medical terms, while medical experts tend to use advanced medical terms, thus, information retrieval systems need to be stable for such different query variations. The significant increasing of non-English digital content on the World Wide Web has been followed by an increase in looking for this information by internet users. Grefenstette and Nioche [8] presented an estimation of language size in 1996, late 1999 and early

2000 for documents captured from the internet. Their study showed that the English content has grown 800%, German 1500%, and Spanish 1800% in the same period. Furthermore, users started to look for information needs that is represented in documents which are not available in their native languages.

The system that searches for information in a language different from the one of user is called Cross-Lingual (multilingual) Information Retrieval (CLIR) system. It enables users to write queries (information need) represented in a language (lang. A), and returns results from a document collection written in a different language (lang. B). Usually, the baseline system in CLIR is to take *1-best-list* translations which are returned by a statistical machine translation (SMT) system and perform the retrieval as shown in the CLEF eHealth Information Retrieval tasks before [6]. Nikoulina et al. [10] presented an approach to develop Cross-lingual information retrieval (CLIR) system which is based on reranking the hypotheses given from the SMT system. Saleh and Pecina [20] considered Nikoulina's work as a starting point and expanded it by adding a rich set of features for training. They presented approach covered translating queries from Czech, French and German into English and rerank the alternative translations to predict the hypothesis that gives better CLIR performance.

In this paper, we describe our participation at the CLEF 2018 eHealth consumer health search task [23]. We focus in our participation in the multilingual IR Task. We present our machine learning model which reranks the alternative translations given by the machine translation system for better IR results. We also present our new approach to expand translated queries using our machine learning model.

2 Task Description

CLEF eHealth Consumer Health Search Task 2018 [9] is similar to the IR tasks in the previous years (2013–2017). The participants this year are required to retrieve relevant web pages from the provided document collection in response to users' queries. These queries represent information need in the medical domain. The IR task consists of *IRTask1* which is a standard ad-hoc monolingual search task. *IRTask2* is a similar task of the personalised search task in 2017 [16, 7], the retrieved documents are personalised to match user expertise (how likely the user is able to understand the content of the retrieved documents). *IRTask 3* contains query variations for the same information need, and the participants have to design a search system that is steady when the same information need is expressed in different query variations. In the multilingual ad-hoc search task (*IRTask4*), the monolingual English queries were translated by experts into Czech, French and German, and the participants are asked to design a search system to retrieve relevant documents to these queries from the English document collection.

2.1 Document Collection

Document collection in the CLEF 2018 consumer health search task is created using CommonCrawl platform [1]. First, the query set (described in Section 2.2) is submitted to Microsoft Bing APIs, and a list of domains is extracted from the top retrieved results. This list is extended by adding reliable health websites, at the end *clefehealth2018_B* (which we use in this work) contained $1,653$ sites, after excluding non-medical websites such as news websites. After preparing the domain list, these domains are crawled and provided as an indexed collection to the participants. Two indexes are provided, in the first one, documents are stemmed and a stop-word list is used, while no preprocessing is done in the second index. The collection contains $5,560,074$ documents, the stemmed index contains $14,213,903$ vocabularies, while the non-stemmed index contains $15,298,904$ ones.

2.2 Queries

The query set this year includes 50 English queries. This set is a subset of 150 medical queries that were created from HON and TRIP query logs within the Khresmoi project [4]. Table 1 shows the average number of terms in the 50 test queries in all languages. Although the average number of terms in the English queries is 5.64, there are queries that are much longer (e.g. query *199001*), as shown in Table 2. Queries might contain typos since they are constructed from real query logs, as shown in query *175001*, which contains *Emugel* instead of *Emulgel*.

Table 1. The average number of terms in the query test set of the CLEF eHealth 2018 IR task

EN	CS	FR	DE
5.64	5.28	6.08	4.62

Table 2. Query samples from the English query test of the CLEF eHealth 2018 IR task

id	**title**
156001	food allergy test
168001	hiv vaccine phase
175001	Voltaren Emugel 1%
199001	why is there a minimum drinking age and what are the consequences of underage drinking ?
200001	feeling of fullness with hiccups with a feeling of a lump in the back of the throat

[1] http://commoncrawl.org/

3 The training data

The data that we use to train our systems was presented by the CLEF eHealth 2014 Task 3 - Information Retrieval [5] and CLEF eHealth 2015 Task 2: User-Centred Health Information Retrieval [15]. It is almost identical to the collection used in CLEFeHealth 2013 Task 3 - User-Centred Health Information Retrieval, which contained a few additional documents which were excluded from the 2014/2015 collection due to license issues. The document collection includes a total of 1,104,298 web pages in HTML, automatically crawled from various English medical websites such as Genetics Home Reference, ClinicalTrial.gov and Diagnosia. To clean the HTML pages in the collection, we follow the work of Saleh and Pecina [19]. The queries have also been adopted from the CLEF eHealth series and include all the test queries from the IR task of 2013 (50 queries), 2014 (50 queries), and 2015 (66 queries). We joined them to create a more representative and balanced sample for IR experiments. The set of all 166 queries was split into 100 queries for training and 66 queries for testing. The two sets are stratified in terms of distribution of the year of origin, number of relevant/not-relevant documents, and query length (number of words).

4 Methods

4.1 Translation system

For the multilingual task (*IRTask4*), we follow the query translation approach, in which a query is translated into the collection language (English), then the retrieval is conducted. Query translation approach reduces the task into monolingual task (both queries and documents are expressed in the same language). We use Khresmoi statistical machine translation (SMT) system [2], for language pairs: Czech-English, French-English and German-English, to translate the queries into English. Khresmoi SMT system was trained to translate queries, and tuned on parallel and monolingual data taken from the medical domain resources like Wikipedia, UMLS concept descriptions and UMLS metathesaurus. Such domain specific data made Khresmoi perform better when translating sentences in the medical domain like the queries in our case. Generally, feature weights in SMT systems are tuned toward BLEU [17], a method for automatic evaluation of SMT systems correlates with human judgments. It is not necessary to have correlation between the quality of general SMT system and the quality of CLIR performance [18]; therefore Khresmoi SMT system was tuned using MERT [12] towards PER (position-independent word error rate), because it does not penalise word reorder; which is not important for the performance of IR systems.

4.2 Hypotheses reranking

Khresmoi SMT system produces a list of ranked translations in the target language, for each sentence in the source language, this list is called *n-best-list.*

However, this *n-best-list* is ranked based on the translation quality rather than the retrieval performance. Saleh and Pecina [20] presented an approach to rerank an *n-best-list* and predict a translation that gives the best retrieval performance in terms of P@10. The reranker is a generalized linear regression model that uses a set of features which can be divided according to their sources into: 1) **The SMT system**: This includes features that are derived from the verbose output of the Khresmoi SMT system (e.g. phrase translation model, the target language model, the reordering model and word penalty). 2) **Document collection**: This includes IDF scores and features that are based on the blind-relevance feedback approach. 2) **External resources**: Resources like Wikipedia articles and UMLS metathesaurus [22] are employed to create a rich set of features for each query hypothesis. 3) **Retrieval status value (RSV)**: RSV is the score of the retrieval scoring function when constructing a query from a translation hypothesis. It helps to involve more information from the collection in the reranking process by assigning to each hypothesis the score from the retrieval function. This feature is based on the work of Nottelman et al. [11], where they investigated the correlation between RSV and relevance probability. To train the model, we join the training and test sets that we presented in Section 3 in one set, then calculate feature values from each language, and merge them from all seven languages in one training set. The test set is the CLEF eHealth 2018 query set in Czech, French and German.

4.3 Query Expansion

Query expansion is a process that reformulates user's initial queries as an attempt to represent more information to improve retrieval performance eventually. In this section, we present our approach to reformulate user's query in the CLIR task using machine learning model, based on the presented work of Saleh and Pecina [21]. This approach is based on expanding a query by adding candidate terms from an existing pool. This is done by reranking candidate terms using machine learning model towards better IR performance and adding the top ranked terms to the original query. To create a pool of candidate terms for each query, we use two main resources:

- **Translation hypotheses**: This pool is built by merging **n-best-list** translations for each query, after filtering stopwords and terms that already appeared in the **1-best-list** translation.
- **Wikipedia titles**: First, we index English Wikipedia articles (titles and abstracts without any preprocessing) using Terrier [13] and its implementation of Dirichlet language model as an IR model, then we conduct retrieval for each query's **1-best-list** translation from this index, then the top 10 ranked Wikipedia articles are selected and their titles are added to the pool.

To train the model, we use the training data that we presented in Section 3, while for testing, we use the provided queries from the CLEF eHealth 2018 IR task in Czech, French and German languages. After building a pool of candidate terms, we generate the following features for each term:

- **IDF** The inverse document frequency which is calculated from the relevant document collection.
- **Translation pool frequency** This feature represents how many times a term appeared in the translation pool. 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.
- **Wikipedia frequency** The frequency of a term in the top 10 retrieved Wikipedia articles. Retrieval is conducted using the *1-best-list* translation for the query that we want to expand with the candidate term.
- **Retrieval Status Value difference** To calculate this feature, we conduct two retrievals, the first one using the original query (1-best-list translation), and the second one using the original query expanded with the candidate term, then we take the score of the highest ranked document in each retrieval and calculate the difference between them. This feature tells us the contribution of the candidate term to the retrieval status value.
- **Similarity** To calculate the similarity between a candidate term t_m and the query terms, we use a trained model of *word2vec* embeddings on 25 millions articles from PubMed [2]. First, we get the word embeddings for each term in the original query and we sum these 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 t_m vector.
- **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 in the Equation 1.

$$co(t_m, Q) = \sum_{d_j \in C, t_i \in Q} tf(d_j, t_i) tf(d_j, t_m) \tag{1}$$

- **Term frequency** First, we perform retrieval from the collection using a query that is constructed from the *1-best-list* translation, then we calculate the term frequency of a candidate term t_m in the top 10 ranked documents from the retrieval result.
- **Medical term count** This feature represents how many times a term appeared in the UMLS lexicon, as an attempt to give more weight to the medical terms.

Our goal is to design a model that can predict the performance of the retrieval when expanding a query with a term from the terms pool, and add terms that can improve the performance. To train the model, we perform the following steps:

- Generate a pool of candidate terms for each query in the training and test set.

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

- Add one term from the pool to the query that we want to expand (*1-best-list* translation) and perform the retrieval using the baseline system (Dirichlet model)
- Calculate the feature values for each term as we described above.
- For training queries, we evaluate the performance for each expanded query considering P@10 as a main metric, P@10 being the objective function for our model.
- Merge training queries from the 7 languages together to enrich the training set with more instances.
- After preparing the training set, we normalise feature values using standard scaling by removing the mean and scaling them to have unit variance. This is done independently on each feature, then we use the scaler coefficient to standardise the test set. Scaling is important since the range of the feature values varies widely.

The term reranker is a generalised linear regression model which predicts P@10 value for each term when expanding the original query with, we choose the term that has the highest predicted value of P@10.

5 Systems

We submit runs for the monolingual task (IRTask1) and the multilingual task (IRTask4), as we present in the following sections.

5.1 Monolingual system

In the monolingual task, we submit four runs:

- **Run 1** In this system, we use the Terrier's index that is provided by the organisers without applying any data preprocessing. Terrier's implementation of Dirichlet smoothing language model is used as the retrieval model with its default parameters.
- **Run 2** This system also uses the same retrieval model as in Run 1, while as an index, we use Terrier's index that uses Porter-stemming method and English stop-word list.
- **Run 3** This system uses Terrier's implementation of *TF-IDF* model, for the purpose of comparing between a vector-space model and an LM model (the one that is used in Run 1), we use the same index as in Run 1.
- **Run 4** In this run, we use Terrier's implementation of Kullback-Leiber divergence (KLD) [1] for query expansion, with number of top documents is set to 10 and number of terms for expansion is set to 3. These 3 terms are selected as following: first, an initial retrieval is done using the base query and the top 10 documents are chosen as pseudo-relevant documents. Then each term in these documents is scored as shown in Equation 2, where $P_r(t)$ is the probability of term t in the pseudo-relevant documents (these documents are treated as a bag-of-words), and $P_c(t)$ is probability of term t in

the document collection c. Finally the top 3 scored terms are added to the base query and a final retrieval is done using the new expanded query.

$$Score(t) = P_r(t) \cdot log\left(\frac{P_r(t)}{P_c(t)}\right) \tag{2}$$

5.2 Cross-lingual system

- **Run 1** In this run, we translate the queries in the source languages into English and get *1-best-list* translations. Retrieval is conducted using Dirichlet model, and non-stemmed index. The same retrieval settings are used in the following runs.
- **Run 2** This run uses hypotheses reranking approach, in which each query is translated into English and from the 15-best-list translations, the 1-best-list (in terms of IR quality) translation is selected for the retrieval as described in Section 4.2
- **Run 3** First we translate the queries into English and the 1-best-list that is produced by the SMT system is chosen as a base query, then this query is expanded by one term using the term reranking approach that is presented in Section 4.3
- **Run 4** This run is similar to Run1, the only difference is that Google Translate [3] is used to translate the queries into English.

Table 3. Similarity (in percent) of top 10 retrieved documents between the submitted runs in IRTask4

runs	**CS**	**DE**	**FR**
run1-run2	48.80	50.20	55.60
run1-run3	38.00	26.60	35.20
run1-run4	**52.80**	**54.40**	**62.80**
run2-run3	32.20	22.00	33.80
run2-run4	46.20	43.20	43.40
run3-run4	27.80	14.4	28.00

Table 3 shows the percent of similar documents that are retrieved (among the highest 10 ranked ones) by different runs. It is clear from the table that different approaches tend to retrieve different documents, for example, run 3 uses query expansion based approach. Query expansion means that a query will be expanded by more terms to include more information, leading to retrieve different documents, that is the reason why this run has the lowest similarity to the other runs. Both of run 1 and run 4 use 1-best-list translation from two different machine translation systems (Khresmoi and Google Translate respectively) to

[3] translate.google.com

construct the queries. This explains why these two systems share similar documents more than all other systems. Run 2 uses hypotheses reranking approach to select best translation to be used for retrieval, while run1 uses 1-best-list translation as it is selected from the SMT system to construct queries. According to further analysis we performed between the difference between the retrieved documents by these two runs, we found that 23 queries (out of 50) have 100% similarity of the top 10 retrieved documents, and this correlates with what was shown by Saleh and Pecina [20], that an SMT system fails in 50% of the cases to select the best translation to perform the best performance for the retrieval.

6 Conclusion

We presented our participation in CLEF eHealth Consumer Health Search Task 2018 (monolingual and multilingual subtasks). Four runs were submitted to the monolingual task, two runs use a language-model IR with Dirichlet smoothing, they differ in the used index (one uses a stemmed index and one uses an index without stemming). As for the multilingual task, we submitted four runs for each language of Czech, French and German. The first one uses 1-best-list translation from a statistical machine translation system, the second run uses hypotheses translations reranking, the third run is an implementation of query expansion using term reranker model, while the last run uses Google Translate to translate the provided queries into English. Our results analysis shows that similar approaches tend to share more similar retrieved documents than different approaches.

Acknowledgments

This research was supported by the Czech Science Foundation (grant n. P103/12/G084).

References

1. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness, and selective application of query expansion. In: European conference on information retrieval. pp. 127–137. Springer (2004)
2. 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. ACL, Baltimore, USA (2014)
3. Fox, S.: Health Topics: 80% of internet users look for health information online. Tech. rep., Pew Research Center (2011)
4. Goeuriot, L., Hamon, O., Hanbury, A., Jones, G.J., Kelly, L., Robertson, J.: D7.2 Meta-analysis of the first phase of empirical and user-centered evaluations. Tech. rep. (August 2013)

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. pp. 1–22. Springer, Sheffield, UK (2014)
6. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Néváol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2015. In: The 6th Conference and Labs of the Evaluation Forum. pp. 1–15. Springer, Berlin, Germany (2015)
7. Goeuriot, L., Kelly, L., Suominen, H., Nvol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth evaluation lab overview. In: CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS). Springer (2017)
8. Grefenstette, G., Nioche, J.: Estimation of english and non-english language use on the www. In: Content-Based Multimedia Information Access - Volume 1. pp. 237–246. RIAO, Centre de hautes etudes internationales d'informatique documentaire, Paris, France (2000)
9. Jimmy, Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L.: Overview of the CLEF 2018 consumer health search task. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS, Avignon, France (2018)
10. 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. Avignon, France (2012)
11. Nottelmann, H., Fuhr, N.: From retrieval status values to probabilities of relevance for advanced IR applications. Information retrieval 6, 363–388 (2003)
12. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. pp. 160–167. Sapporo, Japan (2003)
13. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proceedings of Workshop on Open Source Information Retrieval. pp. 18–25. ACM, Seattle, WA, USA (2006)
14. Palotti, J.R.M., Hanbury, A., Müller, H., Jr., C.E.K.: How users search and what they search for in the medical domain - understanding laypeople and experts through query logs. Inf. Retr. Journal 19(1-2), 189–224 (2016)
15. Palotti, J.R., 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 (Working Notes). pp. 1–22. Spriner, Berlin, Germany (2015)
16. Palotti, J., Zuccon, G., Jimmy, Pecina, P., Lupu, M., Goeuriot, L., Kelly, L., Hanbury, A.: CLEF 2017 task overview: The IR Task at the eHealth evaluation lab. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR-WS, Dublin, Ireland (2017)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on Association for Computational Linguistics. pp. 311–318. Philadelphia, USA (2002)
18. 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. Artificial Intelligence in Medicine 61(3), 165–185 (2014)

19. Saleh, S., Pecina, P.: CUNI at the ShARe/CLEF eHealth Evaluation Lab 2014. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum. vol. 1180, pp. 226–235. Sheffield, UK (2014)
20. Saleh, S., Pecina, P.: Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. The 7th International Conference of the CLEF Association, CLEF 2016. pp. 54–66. Springer, Évora, Portugal (2016)
21. Saleh, S., Pecina, P.: Task3 patient-centred information retrieval: Team CUNI. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings. vol. 1866. Dublin, Ireland (2017)
22. Schuyler, P.L., Hole, W.T., Tuttle, M.S., Sherertz, D.D.: The UMLS Metathesaurus: representing different views of biomedical concepts. Bulletin of the Medical Library Association 81(2), 217 (1993)
23. Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Nèvèol, A., Ramadier, L., Robert, A., Palotti, J., Jimmy, Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2018. In: CLEF 2018 - 8th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science LNC, Springer, Avignon, France (2018)