Task 3 Patient-Centred Information Retrieval: Team CUNI

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. This paper describes our systems that we submitted to the 2017 CLEF eHealth information retrieval (IR) task. We submitted runs to the monolingual and multilingual tasks. In the monolingual task, we investigate the performance of two IR models: probabilistic model and a model based on language-model. In addition, we experiment query expansion based on blind relevance feedback. In the multilingual task, we submitted runs for all the languages. We employ a Statistical Machine Translation (SMT) system to translate the given queries into English and get the *n-best-list*. Then we use this list of translations for our baseline system by getting 1-best-list to generate queries, we also use *n-best-list* reranker that was developed by us to predict *1-best-list* for better IR performance. Finally, we present our method for query expansion approach based on a machine learning model that predicts a term from a translation pool to be added to the original query.

Keywords: Multilingual information retrieval, Machine Translation, Machine learning

1 Introduction

Internet searches for medical topics had been increasing recently, and have gotten the attention of information retrieval researchers. Fox [2] 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 [12]. Laypeople express their medical information need using non-medical terms, while medical experts express it using specific 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 [5] 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. Further more, users started to look for

information needs 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 the *1-best-list* translation returned by a statistical machine translation (SMT) system and perform the retrieval as shown in the CLEF eHealth Information Retrieval tasks before [3]. However, researchers recently started to investigate looking inside the box of the machine translation system rather than using it as a black box [17, 6] and showed that involving the internal components of the SMT in the retrieval process significantly improved the baseline system.

Nikoulina et al. [8] 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 [16] 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 2017 CLEF eHealth Information Retrieval Task [13, 4]. In the IRTask1, participants were provided with English queries representing medical information need and were asked to provide ranked list of documents from the ClueWeb collection sorted by their relevance. While IRTask4 is a multilingual IR task, the original English queries were translated into seven languages: Czech, French, Hungarian, German, Polish, Spanish and Swedish by medical native speakers. Participants in this task were required to provide a ranked list of relevant documents from the English collection. 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 System description

2.1 Retrieval model

In our experiment we use ClueWeb12 collection indexed and released by the orgnisers of this task. The index was created using Terrier open source engine [11]. We use mainly BM25 as a retrieval model. Documents in this model are ranked for a given query as shown in Equation 1. k_1 and k_3 are tuning parameters, and we leave these parameters as their default values in Terrier. While tf_d is the normalised term frequency in document d, normalised by Equation 2. dl and avg_{dl} are document length and the average of document length in the collection respectively. b is a free parameter, we tune this parameter using the 2016 CLEF

eHealth IR monolingual queries and the provided assessment information, then we set this parameter to 0.6.

$$RSV(d,q) = \sum_{t \in d \bigcap q} \frac{(k_1+1)tf_d}{K+tf_d} * \frac{(k_3+1)*tf_q}{k_3+tf_q} * idf(t) \quad (1)$$

$$tf_d = \frac{tf}{(1+b)+b*\frac{dl}{avg_{dl}}} \quad (2)$$

3 Translation System

We employ Khresmoi statistical machine translation (SMT) system [1], for language pairs: Czech-English, French-English, German-English, Hungarian-English, Polish-English, Spanish-English and Swedish-English, to translate the queries into English. Khresmoi SMT system was trained to translate queries, where most general SMT systems fail, 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 well when translating sentences in the medical domain like the queries in our case. Generally, feature weights in SMT systems are tuned toward BLEU [14] , 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 [15]; therefore Khresmoi SMT system was tuned using MERT [10] 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 Hypothesis reranking

For each input sentence, Khresmoi SMT system returns a list of alternative translations in the target language, we refer to this list as an *n-best-list*. Saleh and Pecina [16] 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**: The collection is employed to derive features like IDF scores and features that are based on the blind-relevance feedback approach. 2) **External resources**: Resources like Wikipedia articles, document collection and UMLS metathesaurus are employed to create a rich set of features for each query hypothesis. 3) **Retrieval status value**: This feature is used to involve the retrieval model in the reranking. It is based on how the Dirichlet model scores the retrieved documents for a given query. This

approach is similar to the work of Nottelman et al. [9], where they investigated the correlation between the RSV and the probability of relevance.

To train the model, we used queries and assessment information from the 2016 CLEF eHealth IR task.

5 Query expansion

5.1 Blind relevance feedback

Query expansion is defined as the procedure of reformulating a user's query for better retrieval efficiency. Blind Relevance Feedback (BRF), also known as Pseudo Relevance Feedback, is the process of automatically expand user's query. It considers the top k documents as relevant to the original query, and then expands the query with terms from these documents. However, the assumption of considering these documents as relevant is risky, because they might not be relevant, thus resulting the original query to be drifted way from its information need. The top k documents are chosen from an initial retrieval that is done using the original query. From these documents we create bag-of-words (BOW) and then we choose from this BOW m terms to be added to the original query. These terms are chosen based on their inverse document frequency from the collection and their frequencies in this BOW. Both k and m need to be tuned based on the used collection and using test queries and assessment information. We use Terrier implementation of BRF and tune k and m using the 2016 CLEF eHealth IR task queries and their assessment information and then based on the results we set $k = 3$ and $m = 10$.

5.2 Term reranking

In this experiment, we present our approach for query expansion in the multilingual task. When we translate the query into English using SMT system, we get *n-best-list* translations. These translations contain different synonyms in the target language for a give term in the source language. The motivation of this experiment is that using more than one of these synonyms, and expanding the original query, could lead to improved retrieval. One of the feature we use in this model is based on the word2vec open source tool developed by [7]. They presented two models: Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram model. These models showed very powerful ability to measure the similarity between words in the collection. We used for our experiment trained model of *word2vec* on 25 millions articles from PubMed using their titles and abstracts, the model available online [1]. To investigate the hypothesis of expanding queries from the translation pool, we use the queries that were provided in CLEF eHealth IR task 2013–2015 by translating them into English and then: 1-) Get *20-best-list* translations for each query. 2-) Create a translation pool as bag-of-words from these translations. 3-) Then we use *1-best-list* translation as an

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

original query, and expand it with one term from the translation pool. 4-) Then we run the retrieval using our baseline setting using the expanded queries. After evaluating the results and collecting the expanded queries that give maximum $P@10$ among all the other expanded queries, we find that the results from expanded queries outperform significantly the results when using only the original queries. To expand the original query with a term from the translation pool, we build regression model that predicts the change of $P@10$ when a term is added to the original query. In order to train the model we present set of features for each term as follows:

- IDF: Inverse document frequency of that term from the indexed collection.
- RSV: First we conduct retrieval using the original query and then we take the RSV of the document that is ranked firstly using our baseline setting, then we add a term to that original query, and conduct the retrieval again, then the feature value is the difference of these two RSVs.
- Similarity: First we use word2vec to get word embeddings for each term in the original query and we sum these embeddings to get vector that represents the entire query. Then we take the embeddings for the candidate term and we calculate the cosine similarity between the query vector and the term vector.

The model is built to predict a term that will give the highest $P@10$ when it is added to the original query, and trained on test queries that are taken from CLEF eHealth IR task 2013–2015.

6 Experiments

This year we submit runs to the Ad-Hoc task in its monolingual and multilingual subtask.

6.1 Monolingual Ad-Hoc search

Run1 This run uses Terrier implementation of BM25 IR model, with normalisation parameter b tuned and set to 0.6.

Run2 For comparison with BM25 model (a probabilistic IR model), we submit this run based on Terrier implementation of Dirichlet Bayesian smoothed model (language-model based IR model).

Run3 In this run, we use Terrier implementation of Blind relevance feedback (Bo1) where k is set to 3 documents and m is set to 10 terms.

6.2 Multilingual task

Run1 In this run, we translate the query variant into English using Khresmoi SMT then we take only the *1-best-list* to generate the topics, then we perform the retrieval using BM25 model.

Run2 First we translate the query into English and take the *15-best-list* translations, then the reranker with all features predicts the translation that gives the highest P@10, the predicted translations are used next to generate the topics and perform the retrieval using BM25 model.

Run3 First we use *1-best-list* to generate queries then we add to each query one term from the translation pool as described in Section 5.2.

Run4 This run uses *1-best-list* English translations to generate queries, then we conduct the retrieval after doing query expansion using Terrier implementation of BRF approach.

7 Conclusion and future work

In this paper we presented our participation in the CLEF eHealth 2017 Task3 Patient-Centred Information Retrieval as the team of Charles university. We submitted runs into the Ad-hoc task including its monolingual and multilingual subtasks. For the monolingual task, we investigated the performance when using probabilistic IR model (BM25) and language-model based IR model, also we submitted run based on BRF approach. We tuned all the parameters for these models using queries and assessment information from the 2016 CLEF eHealth IR task. While for the multilingual task, we employ an SMT system to translate the queries into English and use *1-best-list* to generate queries for our baseline system. We also used our reranker to predict new *1-best-list* for better IR performance. We presented new approach to expand queries with a term from the translation pool using machine learning model.

Acknowledgments

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

References

1. 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, USA (2014)
2. Fox, S.: Health Topics: 80% of internet users look for health information online. Tech. rep., Pew Research Center (2011)
3. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, 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. Springer, Berlin, Germany (2015)
4. 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 (September 2017)

5. 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 '00, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France, France (2000)
6. Magdy, W., Jones, G.: Should MT systems be used as black boxes in CLIR? In: Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., Mudoch, V. (eds.) Advances in Information Retrieval, vol. 6611, pp. 683–686. Springer, Berlin, Germany (2011)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
8. 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)
9. Nottelmann, H., Fuhr, N.: From retrieval status values to probabilities of relevance for advanced IR applications. Information retrieval 6, 363–388 (2003)
10. 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)
11. 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. Seattle, WA, USA (2006)
12. 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)
13. 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)
14. 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)
15. Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlaváčová, 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)
16. 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. Springer, Évora, Portugal (2016)
17. Ture, F., Lin, J., Oard, D.W.: Looking inside the box: Context-sensitive translation for cross-language information retrieval. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1105–1106. Portland, Oregon, USA (2012)