Task3 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. In this paper we present our participation as the team of the Charles University at Task3 Patient-Centred Information Retrieval. In the monolingual task and its subtasks, we submitted two runs: one is based on language model approach and the second one is based on vector space model. For the multilingual task, Khresmoi translator, a Statistical Machine Translation (SMT) system, is used to translate the queries into English and get the *n-best-list*. For the baseline system, we take *1-best-list* translation and use it for the retrieval, while for other runs, we use a machine learning model to rerank the *n-best-list* translations and predict the translation that gives the best CLIR performance in terms of P@10. We present set of features to train the model, these features are generated from the SMT verbose output, different resources like UMLS Metathesaurus, MetaMap, document collection and from the Wikipedia articles. Experiments on previous CLEF eHealth IR tasks test set show significant improvement brought by the reranker over the baseline system.

Keywords: multilingual information retrieval, Machine Translation, Machine learning

1 Introduction

The increasing of internet user searches for medical topics recently gets the attention of the researchers in the field of information retrieval. The main challenge in the medical information retrieval systems that people with different experience, express their information need in different way [8]. 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.

In this paper, we describe our submitted systems to Task 3: Patient-centred information retrieval [2, 13], taking a part in its three subtasks IRTask1 ad-hoc search, IRTask2 query variations and the multilingual search task IRTask3. Also we present our machine learning model which reranks the alternative translations given by the machine translation system for better IR results. The baseline

system in the multilingual task is to take the *1-best-list* translation returned by the statistical machine translation (SMT) system and perform the retrieval as shown in the CLEF eHealth Information Retrieval tasks before. However, researches recently started to investigate looking inside the box of the machine translation system rather than using it as a black box [12, 3] and showed that involving the internal components of the SMT in the retrieval process significantly improved the baseline system.

Nikoulina et al. [4] 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 [11] consider Nikoulina's work as a starting point and expanded it by adding 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 2016 CLEF eHealth Information Retrieval Task and its subtasks. This year, *ClueWeb 12 B13*[1] collection is used, and queries are extracted from posts which were published in health web forums (askDocs[2]). In IRTask 1, we have to retrieve set of documents for each query separately, while in IRTask 2 each group of queries are treated like an information need and we have to design an information retrieval system that is stable although the information need is represented in different query variations. We focus mainly on IRTask 3, the multilingual search. In this task, we are given parallel queries in: Czech, French, Hungarian, German, Polish and Swedish. We are required to build a retrieval system that uses these queries to conduct the retrieval from the given collection.

2 System description

2.1 Retrieval model

We use Terrier, an open source information retrieval system [7], with the same index that was provided by the organizers of this task. The main retrieval model which is used in this paper is Terrier's implementation of Bayesian smoothing with Dirichlet prior weighting retrieval model. This retrieval model is based on language modeling approach. Documents are ranked by calculating the product of each term's probability in the query using the language model for that document. Bayesian smoothing has one smoothing parameter (μ) that is based on the length of the documents in the collection. To tune the μ parameter, we use CLEF 2013–2015 queries that were provided in Czech, French and German. Then, we translate these queries into English and take the *1-best-list* to perform the retrieval and evaluate the results considering P@10 (The percent of relevant documents in the highest 10 retrieved documents); for each μ between 1 and 5000. Figure 1 shows that $\mu = 2500$, which is the default value in Terrier, is a

[1] `http://lemurproject.org/clueweb12`
[2] `https://www.reddit.com/r/AskDocs`

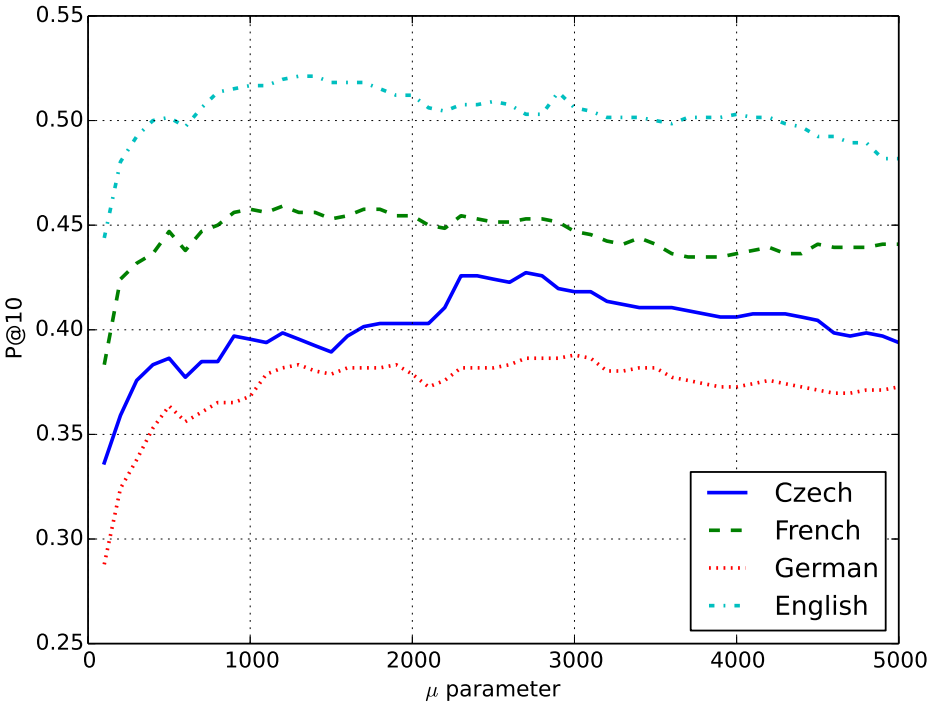

Fig. 1. Tuning μ parameter

reasonable choice for all languages, so we keep it as a smoothing parameter for Dirichlet model in all of our experiments.

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, the case 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 [9] , 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 [10]; therefore Khresmoi SMT system was tuned using MERT [6] towards PER (position-independent word error rate) because it does not penalise word reorder; which is not important for the performance of IR systems.

Table 1. Number of Out-Of-Vocabulary terms remain untranslated in all of test queries

Czech	French	German	Hungarian	Polish	Spanish	Swedish
91	43	101	200	185	36	154

In the test queries, Khresmoi SMT system could not recognize some words because they are unseen in its training data, we call them Out Of Vocabulary (OOV) words. Table 1 shows statistics about the number of OOVs in whole test queries for each language, we do not process these OOVs and just copy them into translated queries as they are.

4 Hypothesis reranking

For each input sentence, Khresmoi SMT system returns list of alternative translations in the target language, we refer to this list as *n-best-list*. Saleh and Pecina [11] 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 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. [5], where they investigated the correlation between the RSV and the probability of relevance.

5 Experiments

5.1 Monolingual task

Ad-Hoc search

Run1 This run uses Terrier implementation of Dirichlet smoothed language model, the smoothing parameter μ is setup by default to 2500, we do not do any preprocessing for the queries nor for the collection.

Run2 For comparison with language model based IR model, we submit this run based on vector space model (TF_IDF) as it is implemented in Terrier.

Query variations

Run1 and Run2 These two runs are similar to Run1 and Run2 in the ad-Hoc search but for each information need, we take the 1000 highest ranked documents that are returned by all of its query variations.

5.2 Multilingual task

Ad-Hoc search

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 Dirichlet 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 Dirichlet model.

Run3 This run is similar to previous run (Run2), but the reranker uses SMT features and the rank features.

Query variations

Run1 For each information need, we translate all of its query variations into English and then concatenate the *1-best-list* translations after removing the duplicated terms, then we perform the retrieval using the resulted query.

Run2 In this run we first translate each query variant and take *15-best-list* translations, then we merge these translations from all query variations that belong to the same information need together, after generating the feature values from these translations the reranker which uses all features predicts the translations that gives the highest P@10 to be used for the retrieval.

Run3 First we use the reranker with all features to get the best translation for each query variant, then we perform the retrieval using each variant separately , after that for each information need, we take the highest 1000 ranked documents among all the documents that are returned by its query variations.

Table 2 shows samples from the test queries, where the reference query (*ref*) is the original query which is provided in the monolingual task, while the baseline query (*base*) is the *1-best-list* translation which is returned by Khresmoi SMT and used in our baseline system. Last query (*Reranker*) is the hypothesis that is predicted by the reranker which uses all features. The evaluation and results of this approach will be presented in the overview paper [13].

Table 2. Examples of query translations including reference translation (*ref*), translation by the baseline system (*base*) and by our reranker selection (*Reranker*). The language labels denote the source language.

Query: 101003 (ES)
ref: inguinal hernia success rate
base: success rate of inguinal hernia
Reranker: percentage success inguinal hernia

Query: 147004 (CS)
ref: viral throat infection symptoms
base: viral infection in neck manifestations
Reranker: viral infection of throat symptoms

Query: 114003 (CS)
ref: rolled ankle healing time
base: podvrtnut ankle healing time
Reranker: podvrtnut ankle healing period

Query: 105004 (DE)
ref: water intoxication symptoms
base: wasservergiftungsymptome
Reranker: wasservergiftungsymptome

Query: 126003 (FR)
ref: cardiovascular issues
base: cardiovascular disorders
Reranker: cardio - vascular disorders

Query: 102003 (HU)
ref: skin tag remove
base: skin appendage removal
Reranker: skin removal

Query: 107001 (PL)
ref: irregular period coughing up blood
base: irregular menses up blood
Reranker: irregular bleeding period

Query: 103002 (SV)
ref: high iron headache
base: high iron headache
Reranker: high iron balance headaches

5.3 Conclusion and future work

In this paper we presented our participation in the CLEF eHealth 2016 Task3 Patient-Centred Information Retrieval as a team of Charles university. For the monolingual task, we investigated both of the language model-based IR model and the vector space model, while for the query variations task, we used the score returned by the retrieval model for all variations to get the highest ranked documents for each information need.

In the multilingual task, we used Khresmoi SMT to translate the queries into English and perform the retrieval using the *1-best-list* in the baseline. For other runs, we used a machine-learning reranking model, which is trained to predict the hypothesis that produces better CLIR performance in terms of P@10. To train the model, we used different sources of features based on the SMT system, translation pool, the collection, MetaMap, UMLS Metathesaurus, Dirichlet retrieval model and the Wikipedia articles. For the query variations in the multilingual task, we used the reranker with all presented features to predict the best hypothesis that belongs to the same information need. In the future, we plan to investigate new features and train special model to predict the best query variation for each information need, also we aim to solve the Out-Of-Vocabulary problem which hurts the retrieval performance.

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. Kelly, L., Goeuriot, L., Suominen, H., Nvol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. In: CLEF 2016 - 7th Conference and Labs of the Evaluation Forum. Springer (September 2016)
3. 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)
4. 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)
5. Nottelmann, H., Fuhr, N.: From retrieval status values to probabilities of relevance for advanced IR applications. Information retrieval 6, 363–388 (2003)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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)
11. 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)
12. 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)
13. Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., Mueller, H., Budaher, J., Deacon, A.: The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval. In: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (September 2016)