An Extended CLEF eHealth Test Collection for Cross-Lingual Information Retrieval in the Medical Domain

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 test collection for medical cross-lingual information retrieval. It is built on resources used by the CLEF eHealth Evaluation Lab 2013–2015 in the patient-centered information retrieval tasks and improves applicability and reusability of the official data. The document set is identical to the official one used for the task in 2015 and contains about one million English medical webpages. The query set contains 166 items used during the three years of the campaign as test queries, now available in eight languages. The extended test collection provides additional relevance judgements which almost doubled the amount of the officially assessed query-document pairs. This paper describes the content of the extended collection, details of query translation and relevance assessment, and state-of-the-art results obtained on this collection.

Keywords: Cross-lingual Information Retrieval · eHealth · Benchmarking

1 Introduction

Cross-lingual Information Retrieval (CLIR) allows users to search for documents using queries in a language different from the language of the documents. Evaluation of CLIR system is difficult mainly due to limited availability of appropriate benchmarks and their reusability. In this paper, we present an extended version of the test collection used in the CLEF eHealth Evaluation Lab in 2013–2015 [6,7,16] for the patient-centered information retrieval task. This benchmark (available via the LINDAT/CLARIN repository)[1] contains about one million documents (medical webpages in English), 166 queries (generated in English and translated to other languages), and relevance assessments based on pooling the officially submitted results. Our main contribution is providing complete manual translations of the queries into seven languages (Czech, French, German, Hungarian, Polish, Spanish, Swedish) and extending the relevance judgements by assessing highly ranked documents in additional cross-lingual experiments.

[1] http://hdl.handle.net/11234/1-2925.

L. Azzopardi et al. (Eds.): ECIR 2019, LNCS 11438, pp. 188–195, 2019.
https://doi.org/10.1007/978-3-030-15719-7_24

We also include machine translation of the queries into English, propose a new training/test data split and report state-of-the-art CLIR results on this benchmark.

2 Related Work

CLIR has been studied since the 1990's, and several benchmarks have been produced within various evaluation challenges. A brief overview of the major ones follows. **TREC** (Text REtrieval Conference) is an annual event organized by NIST[2]: In 1997, TREC-6 [22] was the first TREC event accommodating a CLIR track. The document collection included three sets of English, French and German documents taken from news agencies. 25 test topics in the same languages were created based on the interest of the participated assessors who performed binary relevance assessment for these queries. The TREC-7 CLIR track used the same document collection as in TREC-6 plus a set of documents and topics (28) in Italian [20]. The TREC-8 CLIR track used the same document collection as in TREC-7 with new set of 28 queries in the same four languages [21]. TREC-9 ran a CLIR track with document collection aggregated from Chinese news agencies and 25 queries in English and Chinese [4]. In the TREC-10 CLIR track, an Arabic newswire document collection was used with a set of 25 topics created by assessors in Arabic and English and afterwards translated into French [5]. In TREC-11 [14], the same Arabic document collection as in TREC-10 was used with newly 25 created English topics then translated into Arabic. **NTCIR** (NII Testbeds and Community for Information access Research) is a project of NII[3]. The first NTCIR workshop (NTCIR-1) was held on 1999 and aimed to improve linguistic research of Asian languages [9]. NTCIR-1 released test collection which included scientific documents in Japanese and English, plus 83 Japanese topics with graded relevance assessment. NTCIR-2 worked with a collection of academic conference papers in Japanese and English and 49 topics in both languages. NTCIR-3 used a document collection of news in Chinese, Japanese and English with 50 topics in Chinese and 30 topics in Japanese and their translations into Chinese, Korean, Japanese and English. The same dataset was used in NTCIR-4 CLIR. The NTCIR-5 CLIR test collection included documents from news agencies in Chinese, Japanese, Korean and English and 50 search topics in all these languages with graded relevance assessment. NTCIR-6 exploited a document collection of newspaper articles. It reused the collection from NTCIR-5, 4 and 3 CLIR tasks and included 50 topics in Chinese, Japanese, Korean and English and additional documents from newspaper articles in Chinese, Japanese and Korean with graded relevance assessment too. NTCIR-7 ACLIA included CLIR as a subtask which included news articles in Chinese, Japanese and Korean, with 100 topics in Japanese and 100 topics in Chinese and 300 English topics and 3-level relevance assessment. NTCIR-8 ACLIA also launched CLIR subtask with documents in Chinese and Japanese with 300 topics in English. **FIRE** (Forum

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

[3] http://ntcir.nii.ac.jp.

for Information Retrieval Evaluation) [13] has been running since 2008 and aims to support research in multilingual information access for Asian languages. In FIRE 2008, a document collection of news articles in English, Hindi and Marathi was used with 50 queries in the same languages. In FIRE 2010, the 2008 document collection was enriched with new documents in Bengali. A set of 50 topics is manually translated into English, Gujarati, Marathi, Tamil and Telugu. FIRE 2011 used the same collection as in 2010, the queries were refined and interactive search was used to improve the relevance assessment. **CLEF** (Cross-Language Evaluation Forum)[4] has organised multiple tasks of multilingual information access. The Ad-hoc track was organised from 2000 to 2009. The document collections in 2000–2007 were collected from news agencies in several European languages and topics were generated in multiple languages to allow CLIR evaluation. In 2008 and 2009, the document collection was created in cooperation with the European Library [2]. The CLEF CL-SR (Cross-Language Speech Retrieval) task was organized annually in 2003–2007 and focused on searching in spoken English news archives using queries in five languages (Czech, English, French, German and Spanish)[17]. **CLEF ShARe/eHealth**[5] has been organized since 2013 aiming at improving access to the medical and health-related documents by laypeople and medical experts in monolingual and cross-lingual settings. In ShARe/CLEF eHealth 2013 Task 3 [6], the English queries were generated by clinical documentation reporters and nurses based on real discharge summaries to mimic the realistic patients' queries. Five queries were used for development purposes and 50 queries for testing. The document collection contained about one million English pages crawled from medical websites. No CLIR task was organized that year. In ShARe/CLEF eHealth 2014 Task 3 [7], the queries were generated in the same fashion as in the previous year. In addition to the monolingual task, a CLIR task was introduced. Five development and 50 test queries were generated in English and then manually translated into Czech, German and French to simulate cross-lingual setting. The document collection was the same as in 2013. In CLEF eHealth 2015 Task 2 [16], the query creation aimed to implement self-diagnosing case. Non-expert student volunteers were shown images of symptoms of specific conditions and asked to create three different queries (in English) for each symptom. 66 queries were then randomly selected and used for testing (plus 5 queries for development). The queries were manually translated into Arabic, Czech, French, German, Farsi and Portuguese. The 2015 collection was a subset of the 2014's collection (a few websites were removed). In CLEF eHealth 2016 Task 3 [10], a new document collection was introduced (ClueWeb12 B13[6]). The collection contained web documents from both medical and non-medical domain in an attempt to give more realistic representation when users look-up information from the web (generic collection). An initial query pool was created from online posts that contain questions about health conditions. Then for each query, six query variations were created by three medical experts

[4] http://www.clef-initiative.eu/.

[5] https://sites.google.com/site/clefehealth/.

[6] http://lemurproject.org/clueweb12/specs.php.

and three people without medical knowledge resulting into the final set of 300 queries representing 50 topics. The queries were translated (by medical experts) into Czech, French, German, Hungarian, Polish, Spanish and Swedish to allow CLIR experiments. CLEF eHealth 2017 IR Task used the same collection and queries as in 2016. However, an additional assessment was performed [15]. CLEF eHealth 2018 Consumer Health Search Task released a document collection created using CommonCrawl platform [8] containing more than five million documents from more than thousand websites. 50 queries were provided in English in the monolingual task (IRTask 1 Ad-hoc search). In IRTask 4 (Multilingual Ad-hoc Search) the same English queries were provided in French, German and Czech.

Table 1. Examples of test queries.

Id	Year	Title
qtest2013.38	2013	*MI and hereditary*
qtest2013.41	2013	*right macular hemorrhage*
qtest2014.1	2014	*Coronary artery disease*
qtest2014.6	2014	*Aortic stenosis*
clef2015.test.1	2015	*many red marks on legs after traveling from US*
clef2015.test.57	2015	*infant labored breathing and tight wheezing cough*

3 Test Collection

The presented test collection is based on the CLEF eHealth resources used in 2013–2015. We adopt the document collection, the original English queries, their translations to other languages (where available), and the relevance assessments. The set of **documents** is identical to the one used in the CLEF eHealth 2015 Task 2: User-Centred Health Information Retrieval [16]. It includes a total of $1,104,298$ web pages in HTML that are automatically crawled from various medical English websites (e.g. Genetics Home Reference, ClinicalTrial.gov, Diagnosia). The average length of a document is 911 words.

The **queries** include all test queries from the IR tasks in 2013 (50 queries), 2014 (50 queries), and 2015 (66 queries). The nature of the queries varies from year to year (see Sect. 2 and Table 1). We mixed them to get more representative and balanced query set, and then split this set into a subset of 100 queries for training (33 queries from 2013 test set, 32 from 2014 and 35 from 2015) and 66 queries for testing (17 queries from 2013 test set, 18 queries from 2014 and 31 from 2015). The two sets are stratified in terms of distribution of the year of origin, number of relevant/not-relevant documents, and the query length (number of words). The query ID tags in the package preserve the original IDs which allows mapping the queries to their original year. All the queries are available in a total of 8 languages (the original English plus human translations into Czech, French, German, Hungarian, Polish, Spanish, and Swedish) to allow monolingual (queries in English) and cross-lingual retrieval (queries in the other languages). The query translations come from two sources: official translations provided by the CLEF eHealth organisers (in 2015 and 2014, queries were officially released in Czech, French and German, while in 2013, only English queries were available) and newly created translations (when the official translations were not existing). The new translations were conducted by medical experts fluent in English and the target language. They followed the same instructions as the official translators [7,19] (i.e. preserve syntax where possible and translate term-by-term otherwise).

In addition to the human translation of the queries from English into the target languages, we included queries machine-translated back to English to allow CLIR experiments without having access to a machine translation system. We employed the phrase-based SMT system that is adapted to translate medical-domain queries described in [3]. For each input query, the system generates a list of 1000 ranked translation hypotheses (*n-best-list*) including internal system information and scores for each one of them (e.g., alignment between source and target language, scores of language model, translation model, reordering model and word penalty).

Table 2. Relevance assessment statistics.

	2013	2014	2015	Extension	Total
Relevant	1,174	3,209	2,515	2,517	9,415
Irrelevant	3,676	3,591	9,576	11,851	28,694

The query-document relevance assessment in the presented test collection was substantially improved. The original assessment of 23,741 query-document pairs (6898 relevant, 16843 irrelevant) was enriched by additional 14,368 judgements (2,571 relevant, 11,851 irrelevant) obtained by domain experts instructed the same way as the official CLEF eHealth assessors. Table 2 shows statistics of the assessment information in the 2013–2015 CLEF eHealth IR tasks and our contribution to the assessment information in the test set. The newly assessed query-document pairs were selected by pooling results of various experiments. These experiments were conducted after the end of CLEF 2015 IR Task, using the queries and the original assessment from 2013–2015 IR tasks. The major pooling experiment is described in [18] – it is the state-of-the-art result obtained using this collection. This approach exploits multiple hypothesis translations (for an input query) produced by the MT system [3] which are reranked using a supervised machine-learning method trained to directly optimise the retrieval quality. The document pool contained unjudged documents from the top 10 retrieved documents for each query translation. Although the assessors were different from the official ones, we attempted to mimic the official assessment procedure to the maximum possible extent. The assessors used the same software (Relevation) [11], the same topic descriptions, and the same instructions. Each topic was assessed by a single assessor by randomly splitting the topics among the assessors, and the pooled documents were judged using three grades (irrelevant, somewhat relevant, highly relevant). To get binary assessment (relevant, irrelevant) from the graded assessment, we followed the CLEF eHealth organisers' approach where somewhat relevant documents are considered to be relevant too. To confirm the assessment quality we performed two dual assessment experiments and measured agreement between: (i) the two new assessors and (ii) the new assessors vs. the official assessment. In both experiments, we randomly selected 2 relevant and 2 irrelevant documents for each topic and asked for additional (independent) relevance judgement. The first experiment (i) showed 86% agreement rate, the second experiment (ii) showed 79% agreement rate (measured as accuracy of binarized relevance), which is generally considered to be sufficient [1]. The dual assessment is included separately in the package. The new relevance assessments are very important for reusability of the presented test collection. Test collections without exhaustive relevance assessment tend to underestimate the evaluation scores (by treating unjudged documents as irrelevant) and enriching the assessment helps to reduce this problem. A major effect also comes from the query-document pairs assessed as not relevant. These are useful for methods employing supervised learning e.g. learning to rank [12] and supervised query expansion [23]. The extended relevance assessment also helped in training and evaluation of the hypotheses reranking model in [18] which predicts the optimal query translation out of 15-best translation hypotheses generated by an SMT system, which lead to the best results achieved using this collection (see Table 3).

Table 3. State-of-the-art results (in terms of common retrieval evaluation measures in %) obtained using the extended CLEF eHealth test collection. See [18] for details.

Language	English	Czech	French	German	Hungarian	Polish	Spanish	Swedish
P@10	50.30	48.03	51.67	46.21	48.48	43.18	50.15	41.36
NDCG@10	55.26	49.51	53.27	47.21	49.88	44.01	52.57	43.16
BPREF	39.94	37.59	37.33	36.46	38.00	38.90	34.61	33.44
MAP	28.31	24.02	25.66	23.09	25.28	21.76	25.11	21.29

4 Conclusion

We presented an extended version of the CLEF eHealth test collection for cross-lingual information retrieval in the medical domain based on the collection used in the CLEF eHealth Evaluation Lab 2013–2015 IR tasks. The extended collection improves reusability of the officially provided resources and allows investigating supervised learning approaches (in both cross-lingual and monolingual IR) on the proposed test set. The test set contains English queries and their manual translations into seven languages to allow cross-lingual retrieval, additional relevance assessment, and a new training/test split of the query set. We also added various data for experimenting with machine translation of queries. The data package containing the official and newly added data is publicly available.

Acknowledgments. The language resources presented in this paper are distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic. This work was supported by the Czech Science Foundation (grant n. P103/12/G084).

References

1. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measure. **20**(1), 37–46 (1960)
2. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007: ad hoc track overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85760-0_2
3. 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 (2014)
4. Gey, F.C., Chen, A.: TREC-9 cross-language information retrieval (English-Chinese) overview. In: Proceedings of the Ninth Text REtrieval Conference (TREC-9), pp. 15–23. NIST, Gaithersburg (2000)
5. Gey, F.C., Oard, D.W.: The TREC-2001 cross-language information retrieval track: searching Arabic using English, French or Arabic queries. In: The Tenth Text REtrieval Conference (TREC 2001), pp. 16–26. NIST, Gaithersburg (2001)
6. Goeuriot, L., et al.: ShARe/CLEF eHealth evaluation lab 2013, task 3: information retrieval to address patients' questions when reading clinical reports. CLEF 2013 Online Working Notes **8138**, pp. 1–16 (2013)
7. Goeuriot, L., et al.: ShARe/CLEF eHealth evaluation lab 2014, task 3: user-centred health information retrieval. In: CLEF Online Working Notes. CEUR Workshop Proceedings, vol. 1180, pp. 43–61. CEUR-WS, Sheffield (2014). http://ceur-ws.org/Vol-1180/. ISSN: 1613-0073
8. Suominen, H., et al.: Overview of the CLEF 2018 consumer health search task. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, pp. 1–15. CEUR-WS, Avignon (2018)
9. Kando, N.: NTCIR Workshop: Japanese- and Chinese-English cross-lingual information retrieval and multi-grade relevance judgments. In: Peters, C. (ed.) CLEF 2000. LNCS, vol. 2069, pp. 24–35. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44645-1_3
10. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2016. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 255–266. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_24

11. Koopman, B., Zuccon, G.: Relevation!: an open source system for information retrieval relevance assessment. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1243–1244. ACM, Gold Coast (2014)
12. Liu, T.Y., Xu, J., Qin, T., Xiong, W., Li, H.: LETOR: benchmark dataset for research on learning to rank for information retrieval. In: Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval, pp. 3–10. ACM, New York (2007)
13. Majumder, P., Pal, D., Bandyopadhyay, A., Mitra, M.: Overview of FIRE 2010. In: Majumder, P., Mitra, M., Bhattacharyya, P., Subramaniam, L.V., Contractor, D., Rosso, P. (eds.) FIRE 2010-2011. LNCS, vol. 7536, pp. 252–257. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40087-2_24
14. Oard, D.W., Gey, F.C.: The TREC 2002 Arabic/English CLIR track. In: The Eleventh Text Retrieval Conference (TREC 2002), pp. 1–15. NIST, Gaithersburg (2002)
15. Palotti, J., Zuccon, G., Jimmy, P.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 Workshop Proceedings, pp. 1–10. CEUR-WS, Dublin (2017)
16. Palotti, J.R.M., et al.: CLEF eHealth evaluation lab 2015, task 2: retrieving information about medical symptoms. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 1391, pp. 1–22. CEUR-WS, Toulouse (2015)
17. Pecina, P., Hoffmannová, P., Jones, G.J.F., Zhang, Y., Oard, D.W.: Overview of the CLEF-2007 cross-language speech retrieval track. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 674–686. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85760-0_86
18. Saleh, S., Pecina, P.: Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 54–66. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_5
19. Urešová, Z., Hajič, J., Pecina, P., Dušek, O.: Multilingual test sets for machine translation of search queries for cross-lingual information retrieval in the medical domain. In: Proceedings of LREC 2014, pp. 3244–3247. ERLA, Reykjavik (2014)
20. Voorhees, E.M., Harman, D.: Overview of the seventh text retrieval conference TREC-7. In: Proceedings of the Seventh Text REtrieval Conference (TREC-7), pp. 1–24. NIST, Gaithersburg (1998)
21. Voorhees, E.M., Harman, D.: Overview of the eighth text retrieval conference (TREC-8). In: Proceedings of the Eighth Text REtrieval Conference (TREC-8), pp. 1–24. NIST, Gaithersburg (2000)
22. Voorhees, E.M., Harman, D.: Overview of the sixth text retrieval conference (TREC-6). Inf. Process. Manage. **36**, 3–35 (2000)
23. Zhang, Z., Wang, Q., Si, L., Gao, J.: Learning for efficient supervised query expansion via two-stage feature selection. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, pp. 265–274. ACM, New York (2016)