CUNI-MTIR at COVID-19 MLIA @ Eval Task 2

Shadi Saleh, Hadi Abdi Khojasteh, Hashem Sellat, Pavel Pecina

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

Abstract. This paper describes the participation of CUNI-MTIR team in the COVID-19 MILA Multilingual Semantic Search task. We participate in the monolingual and the bilingual task and its subtasks. In both cases, we use the English document collection. As for queries, we use the English for the monolingual system, and queries in French, German, Spanish and Swedish for the bilingual task. We follow query-translation approach to reduce the bilingual search task into a monolingual one, and we adopt neural machine translation systems that are deployed for the purpose of this task for the translation process. We also study the effect of the morphological analysis (lemmatisation) of the documents and queries on the recall performance of the retrieval.

1 Introduction

The multi-lingual search task in MLIA Community Effort aims at improving COVID-19 related information access for searchers in multi-lingual settings [3]. We choose in our participation to build a monolingual system where we index the provided English documents and use the English queries for retrieval (monolingual system) then we design five runs in the monolingual settings.

As for the bilingual task, we design five runs where the documents are in English, and the queries are translated into English following the query-translation approach.

2 Data

The data in this task includes documents and queries in multiple languages[4]. Participants could freely choose any pair of languages to build their retrieval systems. For example, a system that takes queries in the French language (query language) and retrieves documents in another language such as English (document language). The English document collection includes $1,452,240$ documents from different resources. Most of them ($1,450,251$ documents) are taken from the Medical Information System (MEDISYS). MEDISYS is a media monitoring system that provides health-related articles for the interest of the public health.[1]

COVID-19 MLIA @ Eval Initiative, http://eval.covid19-mlia.eu/.

[1] https://medisys.newsbrief.eu

2.1 Data Preprocessing

We process the given data before indexing it, and we create two versions of it. In the first version, we convert all the given documents into the TREC format, and we only do lowercasing.

The structure of the given data is defined through a specific annotation. This includes various schema definition of each page such us language, title, main keywords and the body of the text. Each sentence in the main body is identified using the *p* tag. This tag also has few informative attributes such as if the text is a boilerplate or not. The document title is also found out from the *title* and *titleStmt* tags. We keep all of these parts as our previous study showed that a simple cleaning method that removes only HTML tags outperformed more complicated methods that removed boilerplate [9]. Next, we split and unwrap sentences by utilising the Moses toolkit [5] and considering language-specific properties. Finally, we replace all terms that refer to the corona virus disease (such as coronavirus, covid-19, sars-cov-2, 2019-ncov, covid19, sars covid 2, and sars-cov2) into one phrase (corona virus). This is done for both documents, English queries and the translated queries.

2.2 Lemmatisation

The motivation of doing lemmatisation of documents and queries is to reduce search space by mapping multiple morphological variations of a given term into one. This helps retrieving more relevant documents and eventually achieving high recall performance. This experiment is dedicated to subtask 2 (high-recall). To achieve that, we employ trained language specified models by exploiting the prediction of the UDPipe-base baseline system [13] which provides lemmatisation and part-of-speech tagging for an input text in more than 94 languages. We utilise the models for English to process input sentences from the English document collection and from the English and translated queries.

2.3 Indexing

For indexing the document collection and conducting retrieval, we use Terrier [8], which is an open-source framework for information retrieval that includes implementation of various retrieval models and query expansion techniques. We create two indices using Terrier. In the first one, we index the document collection using words forms, and in the second one, we index the lemmatised version of the document collection. The main purpose of having those two different indices is to investigate the effect of the morphological processing towards the recall metric. Table 1 shows statistics of the two indices. Lemmatising the documents reduced the number of vocabularies by around 46000 vocabularies. This is because when lemmatising a text, multiple variations of words might be lemmatised to the one lemma.

Index	#Documents	#Tokens	#Vocab
Forms	1,452,240	1,372,106,395	1,281,067
Lemmatised	1,452,240	1,364,633,452	1,244,686

Table 1. Statistics of the two indices we create (using word forms and using word lemmas), including the number of the indexed documents, tokens and vocabularies

3 Machine Translation of Queries

In the multi-lingual task, we adopt term-based matching models; hence both queries and documents should be represented in on common language. Either documents can be translated into query language (document translation), or queries can be translated into document language (query translation). We investigate the two approaches thoroughly in the medical domain in our previous work [11]. We follow in this work the same approach of building a neural machine translation (NMT) for query translate that reported to give the sate-of-the-art results in the CLEF eHealth IR test collection [10].

Our NMT systems based on the transformer model. The NMT systems translate text from four languages (French, German, Spanish and Swedish) into English. The parallel data that is used to train and tune the systems are taken from the medical domain (10 million sentences for each language pair). In addition to this data, we employ the parallel data that is provided in the Covid-19 MLIA machine translation task. The full description of our MT systems is described in our participation report of the MLIA Machine Translation Task [12].

4 System Description in Round 1

In the following sections, we present the description of our submitted runs in. The following runs are applied in both *Subtask 1 - High Precision* and *Substask 2 - High recall*, the only difference between the two tasks is the index that we retrieve documents from. In Subtask 1, we retrieve documents from the *forms* index, while in Subtask 2 we retrieve documents from the *lemmas* index.

In *Subtask 1*, we required to build the system that retrieves the most relevant documents concerning COVID-19 efficiently. It is a classic ad-hoc multi-lingual search task focused on high precision in retrieving the top-ranked documents. Evaluation measures such as *Precision@K* documents as well as *Normalized Discounted Cumulative Gain* will be used to evaluate.

In *Subtask 2*, the focus is more on the finding as many relevant documents as possible with the least effort and high recall. Given a limited resource, there will be a limit on the maximum number of documents that can be retrieved. Evaluation metric such as *Recall@K* and *Area Under ROC* will be used to compare the systems in this subtask.

For dealing with Subtask 2, we use a lemmatiser for full morphological analysis to accurately identify the lemma for each word. Doing full morphological

analysis produces at most very modest benefits for retrieval. The empirical experiment shows because either form of normalisation tends not to improve English information retrieval performance in aggregate. While it helps quite much for some queries, it equally decreases performance a lot for others. Stemming increases recall while influences precision [7].

The following sections show a description of our runs into both subtask1 (high-precision) and subtask1 (high-recall). The only difference between the two subtasks is that in subtask2, we used the index of the lemmatised documents. The same settings in both subtasks are applied in English-to-English monolingual search system and French, German, Spanish and Swedish to English bilingual search system after translated the queries into those languages.

4.1 Run 1: Dirichlet model

We perform Terrier to estimate a language model for each document, and then rank documents by the smoothed likelihood of the query according to the estimated language model. The derived retrieval model interprets in the term of weighting heuristic and then examined by the Dirichlet priors interpolation method with query expansion and relevance feedback. For the Dirichlet, the score of a term q_i is given by:

$$score(D,Q) = log(1 + \frac{TF}{\mu \frac{f(q_i,C)}{oftokens}}) + log(\frac{\mu}{|D| + \mu})$$

where parameter μ is 2500, D is the input document and f is the frequency of the term q_i in the query Q. This model is formulated, and performance of the weighting has been empirically verified [14].

4.2 Run 2: PL2F Model

The second run is based on the Per-Field Normalisation Weighting (Pl2F) model as it is implemented in Terrier [6].

Term positions are recorded within the compressed inverted index, as well as terms from the titles of documents, and texts as separate fields. Using the PL2F model, the relevance score of a document D for a query Q is given by:

$$score(D,Q) = \sum_{t \in Q} qtw.\frac{1}{tfn+1}(tfn.log_2\frac{tfn}{\lambda}) + (\lambda - tfn).log_2 e + 0.5.log_2(2\pi.tfn)$$

Where λ is the mean and variance of a Poisson distribution, given by $\lambda = F/N$; F is the frequency of the query term t in the whole collection, and N is the number of documents in the whole collection. The query weight qwt is given by qtf/qtf_{max}; qtf is the query term frequency and qtf_{max} is the maximum query term frequency among the query terms. tfn corresponds to the weighted sum of the normalised term frequencies for each used field, known as $Normalisation2F$ [6].

4.3 Run 3: Query Expansion Bo2

Run 3 is the query expansion model based on Bose-Einstein (Bo2) model. First, the term weights of the terms from top-ranked documents are calculated. Then, the top most informative terms are then extracted ad merged with the original query to form an expanded one. This weighting pseudo relevance feedback expansion model is based on the Bose-Einstein distribution [1] and the weight of the term t in the top-ranked documents which is given by:

$$w(t) = tf_x \,.\, log_2 \frac{(1 + P_c)}{P_f} \; + \; log_2(1 \; + \; P_f)$$

where tf_x is the frequency of the query term in the top-returned documents. P_n is given by F/N, where F is the term frequency in the collection, and N is the number of documents in the collection. $Pf_x = (tf_x \,.\, l_x)$ / $token_c$; where l_x is the size in the tokens. f is the term frequency of the query term in the whole collection, and $token_c$ is the total number of tokens in the corpus.

Id	Model	Expanded Query
7	Bo2	serological tests corona virus group disease getty produce contract
	KLD	serological tests corona virus across accept dream group speech
1129	Bo2	hand sanitizer time show currently well summit
	KLD	hand sanitizer time class show organization currently
1135	Bo2	covid lockdown protest affect enforcement opinion sport relief
	KLD	covid lockdown protest affect scanty action violence party

Table 2. Example of expanded queries by Bo2 and KLD Correct models.

4.4 Run 4: Query Expansion KLD Correct

We utilise the automatic query expansion technique for the fourth run. The model select expansion terms from the target corpus and compares the distribution of a term in the relevant documents. It uses the Kullback-Leibler [2] divergence between the probability distribution of terms in the relevant documents and in the complete corpus. In this method, all terms in the pseudo relevant set are treated as candidate expansion terms. Let R, and C represent the (pseudo) relevant documents (PRD) and whole corpus respectively. Terms for the contributions is the largest are selected as following expansion terms:

$$S(t) = p_r(t) \; * \; log \frac{p_r(t)}{p_c(t)}$$

Where $S(t)$ is used as the term weight of a candidate expansion term t. p_r and p_c denotes the unigram probability distribution of the terms in R and C. p is calculated as follows which is the $tf(t, D)$ represents the term frequency of term t in document D:

$$p(t) = \frac{\sum_{D \in R} tf(t,d)}{\sum_{D \in R} \sum_{t' \in D} tf(t',d)}$$

The weights of original query terms are normalized using the mentioned maximum original query term weight. Then added to with weights of the expansion terms to obtain the final weight of the term t in the expanded query:

$$score(t) = \frac{1 + log(tf(t,Q))}{1 + max_{t' \in Q} log(tf(t',Q))} + \frac{S(t)}{max_{t' \in D \in PRD} S(t')}$$

4.5 Run 5: Dirichlet Model for Conversational Queries

In this run, we generate queries from the conversation fields in the topic instead of the query title. Then we run the retrieval using the same settings as in Run1. The main difference between conv and title is that conv fields contain more narrative description of the query and describes more the information need that is represented in the query. Table 3 shows samples of four queries, including query title and the conversational field (conv).

Id	Title	Conversational
1	corona virus origin	what is the origin of corona virus
4	how do people die from the corona virus	what causes death from corona virus
7	serological tests for corona virus	are there serological tests that detect antibodies to corona virus
13	how does corona virus spread	what are the transmission routes of corona virus

Table 3. Samples show the difference between title and conversational fields in the given topics

In both expansion models, we set the number of top-ranked documents (n) to 10 and the number of expansion terms (m) to 5.

5 Conclusion

In this work, we presented our participation in the Covid-19 MLIA Multilingual Semantic Search task and its two subtasks (Subtask 1 - High Precision, and Subtask 2 - High recall). We submitted five runs for each language pair (including the monolingual English settings). Our monolingual runs employed language-based retrieval models, per field normalisation weighting model and two famous query expansion models (Bo2 and KLD correct). The multi-lingual search tasks follow the query-translation approach using NMT models for translation into the document language (English) from French, German, Swedish and Spanish, and then the same models in the monolingual settings were used for retrieval.

References

[1] Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) **20**(4), 357–389 (2002)

[2] Carpineto, C., De Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems (TOIS) **19**(1), 1–27 (2001)

[3] Casacuberta, F., Ceausu, A., Choukri, K., Declerck, T., Deligiannis, M., Di Nunzio, G.M., Domingo, M., Eskevich, M., Ferro, N., García-Martínez, M., Grouin, C., Herranz, M., Jacquet, G., Papavassiliou, V., Piperidis, S., Prokopidis, P., Zweigenbaum, P.: The Covid-19 MLIA @ Eval initiative: Developing multilingual information access systems and resources for Covid-19. `https://bitbucket.org/covid19-mlia/organizers-overall/src/master/report/` (2021)

[4] Di Nunzio, G.M., Eskevich, M., Ferro, N.: The Covid-19 MLIA @ Eval initiative: Overview of the multilingual semantic search task. `https://bitbucket.org/covid19-mlia/organizers-task2/src/master/report/` (2021)

[5] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. pp. 177–180. Association for Computational Linguistics (2007)

[6] Macdonald, C., Plachouras, V., He, B., Lioma, C., Ounis, I.: University of glasgow at webclef 2005: Experiments in per-field normalisation and language specific stemming. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 898–907. Springer (2005)

[7] Manning, C.D., Schütze, H., Raghavan, P.: Introduction to information retrieval. Cambridge university press (2008)

[8] 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)

[9] 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)

[10] 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, Proceedings. pp. 188–195. Lecture Notes in Computer Science, Springer (2019)

[11] Saleh, S., Pecina, P.: Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6849–6860. ACL, Online (2020)

[12] Saleh, S., Sellat, H., Abdi Khojasteh, H., Pecina, P.: CUNI-MTIR at COVID-19 MLIA @ Eval Task 3. https://bitbucket.org/covid19-mlia/ (2021)
[13] Straka, M., Straková, J.: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (August 2017)
[14] Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS) **22**(2), 179–214 (2004)