CHARLES UNIVERSITY IN PRAGUE
FACULTY OF MATHEMATICS AND PHYSICS
INSTITUTE OF FORMAL AND APPLIED LINGUISTICS

Cross-lingual information retrieval systems

Shadi Saleh

Ph.D. Thesis Proposal

Supervised by
RNDr. Pavel Pecina, Ph.D
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Prague, 2017

Cross-lingual information retrieval systems: Methods and Challenges

Shadi Saleh
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
saleh@ufal.mff.cuni.cz

Abstract

In this work, we will explore different approaches used in Cross-Lingual Information Retrieval (CLIR) systems. Mainly, CLIR systems which use statistical machine translation (SMT) systems to translate queries into collection language. This will include using SMT systems as a black box or as a white box, also the SMT systems that are tuned towards better CLIR performance. After that, we will present our approach to rerank the alternative translations using machine learning regression model. This includes also introducing our set of features which we used to train the model. After that, we adapt this reranker for new languages. We also present our query expansion approach using word-embeddings model that is trained on medical data. Finally we reinvestigate translating the document collection into query language, then we present our future work.

1 Introduction

The term *information retrieval* (IR) refers to the operations that an IR engine does to get information from a large collection of documents as a response to a given query. When users ask an IR engine for information needs, they express their needs as a query. This query could be a direct question, sequence of words (form a correct sentence or not) or could be an expression using logical operators. Figure 1 shows the general structure of an IR engine.

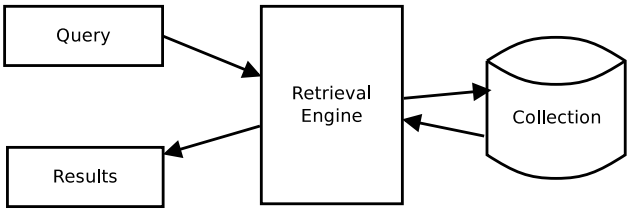

Figure 1: IR engine structure

2 Cross Language Information Retrieval

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 [2000] 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 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). The first CLIR system, which supported multi-lingual search, was the international road research documentation that used aligned terms between keywords in English, French and German [1]. In this proposal, we will study the development of CLIR systems. This include review of the related work and the challenges of the task.

[1] http://www.oecd.org

Then, we will present our experiments and findings in designing our state-of-art CLIR system. Finally, we will present our future work to improve our current CLIR system.

3 Related work

In CLIR, documents and queries are written in two different languages. To conduct a term-matching based retrieval, they should be represented in one language; therefore, either the queries should be translated into collection language or the collection should be translated into the query language.

3.1 Translate queries or documents?

Different studies and experiments have been conducted to answer this question. Oard [1998] investigated the performance of the CLIR when translating documents, queries, or both. For this, they used Logos translation system [2] in order to translate between German and English languages. As for collection and test queries, they used *TREC-6 CLIR SDA/NZZ* which contains about $251,840$ German documents and 22 English queries, and relevance judgments for these queries. They compared two systems: The first system translates English queries into German (the collection language), and the second system translates the documents from German into English (query language). The study showed that translating the documents into query language outperformed translating the queries. This can be explained because the documents are longer than the queries, leading to more contextual and linguistic information; which are usually necessary to reduce the ambiguity that happens when one term in a source language has more than one translation in a target language. Reducing the ambiguity gives better machine translation performance and this in turn leads to better retrieval performance. What also confirmed this explanation, that the longer queries outperformed the short ones. McCarley [1999], for instance, exploited query-translation and document-translation systems in parallel and combined their outputs by averaging document scores obtained by both. Fujii and Ishikawa [2000] employed two-step method where the query-translation approach was used to retrieve a limited number of documents which were then translated into the query language and reranked according to relevance to the original query.

Nonetheless, comparing the performance of different machine translation systems could affect the results of the retrieval and as a result affects our judgment on which approach is better. To avoid this, McCarley [1999] built two translation systems (English to French and French to English) that are similar in performance as much as possible by training them using the same training parallel data. Then they built three systems: 1) A system that translates only queries. 2) A system that translates only documents. 3) A system that calculates the mean average of the document scores from both previous systems - they called this system *hybrid system*. Their experiments on *TREC-6* and *TREC-7* CLIR tracks data showed that the hybrid system outperformed the other two systems. Later in this paper, we will focus on CLIR systems which are based on the use of machine translation systems to translate the queries into collection language. Also we will describe how we employ word-embeddings in expanding user queries and also, we will investigate the document translation approach.

3.2 Dictionary-based CLIR systems

In this approach, bilingual or multilingual dictionaries are used to translate each word of a given query written in the source language into a word in the target language. Pirkola et al. [2001] spotted the main disadvantages behind dictionary-based CLIR systems which are:

- Untranslatable words due to out-of-vocabulary (OOV) problem.
- Processing inflected words.
- Identify phrases and collocations to be translated correctly.
- Lexical ambiguity in source and target languages.

Dictionary-based approach can be supported by information that is extracted from the collection, Bosca et al. [2014] used multilingual semantic and

[2] http://logos-os.dfki.de/

domain-based information from the collection during the indexing in order to map query fragments into concepts.

Gao et al. [2001] used an approach that enhanced the query translation by identifying the phrases using statistical model, then translating the phrases using set of phrase translation patterns and probabilities of the translated phrases using target language model, then translating the remained words as words. Such an approach led to some improvements of CLIR performance.

3.3 Corpus-based CLIR system

This approach of CLIR systems uses aligned corpus as a resource. Where the documents in the collection written in different languages are aligned together, and user queries are translated based on terms from this parallel collection. Talvensaari [2008] showed in their work that the main three factors which affect the performance of corpus-based CLIR systems are:

- Topical nearness between the corpus and the translated queries.
- Quality of the alignment of two documents written in different languages.
- Size of the corpus, where the more aligned documents we have, the more reliable translation knowledge is.

The author also showed that topical nearness is the most important factor among them. However, since the documents should be available in all supported languages, we are keeping this approach out of our research scope.

3.4 SMT system as a black box

Translating the queries using statistical machine translation (SMT) system has shown recently potential improvement against other methods in CLIR. Beside the state-of-art that has been shown when using MT systems, there are many MT systems available for free, these two points make the development of CLIR systems using SMT much easier and promising.

Usually in CLIR , an SMT system is considered to be a black box and separated system from the CLIR. It takes a sentence that is written in a source language (query) as input, then it returns the best translation in a target language (the language of the document collection) for that sentence. Finally, this best translation is used for the retrieval. Hull and Grefenstette [1996] studied the main challenges of building a CLIR system. They found that the main sources of noise and errors in CLIR systems are the translation ambiguity and the missing terminology in the target language when translating the queries into collection language. Also they compared the monolingual queries that were provided in English and the automatically translated ones, and they found that there is big difference in quality between them. This confirms the claims that further investigation should be put to improve the translation quality and disambiguating the translated query terms. Users usually use only 2 terms in average to formulate query and 48.4% of users formulate only one query for each search session [Spink et al., 2001]. This leads to two problems: 1) Translating short sentences (Queries) is difficult for SMT systems because queries usually are not completely grammatically written by users. 2) Queries expressed with 2 terms are insufficient to describe user's information needs even if the translation part goes well.

Improving the quality of MT systems for better CLIR performance might sound feasible. However, the correlation between MT system quality and the performance of CLIR system has been studied before. Pecina et al. [2014] investigated the effect of adapting MT system to improve CLIR system. The system was tested on CLEF eHealth 2013 data set and it supported Czech-English, German-English and French-English pairs. Even the MT systems improved by average of 55% and significantly outperformed the well known public MT systems like Google Translate [3] and Bing Translator [4], but for the CLIR systems only French-English outperformed the baseline system. This means that improving translation quality does not guarantee to improve the performance of CLIR system.

Fujii et al. [2009] investigated the correlation between the translation quality and the retrieval quality in the cross-lingual question answering task, which

[3] http://translate.google.com

[4] https://www.bing.com/translator/

is comparable to the CLIR task. They created search topics from the patent applications that were rejected (because they were available in digital form). For relevance information, the citations, which were used for rejection reason, were considered to be relevant documents (patents). Then these search topics were translated by humans into English. Each participant was required to translate the topics into English using their own MT system. BLEU was used to evaluate the translations, and MAP was used to evaluate the retrieval (MAP and BLEU are explained in Section 4.2). The system that got the highest human evaluation in terms of translation quality, got the lowest MAP value in terms of retrieval quality. This means that the best translation quality in human perspective does not necessary lead to best retrieval quality.

The domain of the collection that CLIR uses for retrieval and the domain of the data that was used to train MT system should be similar as much as possible for better results. Anyway, we conducted experiments to investigate this hypothesis. Results show that general domain SMT systems like Google Translate and Bing Translator outperformed specific-domain MT system like Khresmoi[5] when using these MT systems for CLIR in the medical domain, more details will be showed later in this proposal. Also the overview of CLEF 2009 [Ferro and Peters, 2010], showed that using Google Translate to translate queries improved the CLIR results from 55% of monolingual baseline in 2008 to more than 90% in 2009 for French and German languages. However, using a generic MT system for CLIR has several drawbacks:

1) The system is not aware of the fact that the input is not a complete grammatical sentence but attempts to translate it as such.

2) The MT system (usually statistical) is often able to produce much richer output, including multiple translation hypotheses, provided with various scores from the decoding process, which is ignored.

3) The MT system produces translations in the traditional human-readable form although this is not necessary for the retrieval. If MT is more tightly integrated with IR it can construct the output as a more complex structure (e.g., with translation alternatives or stemmed words).

In order to improve the translated queries, different approaches tried to expand or lexically process the query after translating it into target language. Choi and Choi [2014], who placed first in the multilingual CLEF eHealth 2014 Task 3 [Goeuriot et al., 2014], translated the queries into English (from Czech, French and German) using Google Translate. Then they annotated each query with medical concepts using MetaMap [Aronson, 2001], after that, the top scored concepts are added to the original query after removing those terms which do not appear in the discharge summary of that query. Finally, they weighted the original query with 0.9 and the expansion query is weighted 0.1. Weighting queries was done using Indri query language [Strohman et al., 2005]. Their baseline system used the translated queries only, while the system that uses the expanded queries as mentioned above outperformed the baseline system by 18% for Czech, 4% for German and 4% for French language.

A similar approach was used by Liu and Nie [2015] in the monolingual task of CLEF eHealth 2015 [Goeuriot et al., 2015], who expanded the queries not only through the UMLS concepts but also by terms extracted from Wikipedia articles. The main motivation by using Wikipedia was that the layperson poses the medical query usually using ordinary terms (without using medical terms). This makes it difficult for MetaMap to find relevant concepts, also MetaMap, according to the authors, covers $213,844$ out of 3 million concepts, so using Wikipedia might help to increase the coverage of the medical concepts. Authors claim that Wikipedia text is similar to the way that users pose queries (more generic), while the titles of Wikipedia articles contain medical terms. They used only the abstracts of the articles since they contain less noise. However, using only Wikipedia to expand the queries did not help. Only a system combined Wikipedia approach with MetaMap improved the baseline system (Original queries).

3.5 Looking inside the box of MT systems

MT systems and CLIR systems look at the quality of a sentence in totally different perspectives. MT systems are tuned to produce the best translation in the perspective of humans, where during the evalua-

[5] http://khresmoi.eu/

tion, human annotators are asked to choose the best translation which is the most syntactically and semantically correct [Bojar et al., 2013]. However, the best translation from MT point of view is not necessary to be the best one from CLIR's one. Sokolov et al. [2014] presented an approach based on directly optimising an SMT decoder to immediately output the best translation for CLIR rather than output *n-best-list* hypotheses then rerank them. Optimising process was done by tuning the SMT model weights towards the retrieval objective. They tuned the SMT by using decomposable proxy inside it, which estimates the quality of the retrieval. Such approach enables the decoder to score the hypotheses considering the optimal weights for retrieval objective. Magdy and Jones [2011] modified an MT system to preprocess the queries before translating them. The preprocessing includes case folding, stemming and stopword removal. Their results showed that when translating the preprocessed queries are not statistically different when translating the queries without preprocessing but it showed that MT systems can translate the preprocessed queries up to five times faster than the ordinary MT system. Hieber and Riezler [2015] integrated the step of scoring the retrieved documents into query translation process. They employed a joint model of translation and retrieval which used a decoder based on IR features to score the documents. IR features force the SMT decoder to prefer relevant documents with high probability, which allows the use of SMT decoder directly in the retrieval. Their approach significantly outperformed a baseline system that follows direct translation. Also they confirmed the importance of using language model feature to select better translation.

3.6 Reranking SMT translations

Recently, researchers started to investigate alternative translations reranking approach to build a CLIR system. Reranking is implemented by taking the alternative translations that are produced by an SMT system, rerank them and take the translation that gives the best performance for CLIR in descending way. Different approaches have been presented to estimate the quality of a translation from retrieval's point of view. Tuning the SMT system to produce the best translation needs an access to its internal components, which is not always possible. Therefore; reranking approaches are proposed to use an existing SMT system that produces *n-best-list* and then rerank this list of alternative translations. Darwish and Oard [2003] investigated the effect of usage and selection of alternative translations for a given term to be used in the retrieval. Their experiments showed that combining all of alternative translations for a given term and using them for retrieval outperformed the selection of one best translation. Also they claimed that the use of manual disambiguation to select the best translation can not help us to get significant improvement. Finally, they presented new approach to better use all possible translations. This was done by creating structured queries that use boolean operators to combine alternative translations. Also they found that Boolean structured queries outperformed unstructured queries that simply concatenate all the possible translations.

The limitation of using the best translations from MT systems leads us to the conclusion that using the translation probabilities to select the best term for CLIR is not enough as an evidence. This was the motivation of the study by Xu et al. [2001]. They presented probabilistic CLIR model that uses statistics from a corpus whose language is similar to the query's one. Corpus statistics helps to indicate the importance of the query terms (motivated by TF.IDF model). Their probabilistic model combines bilingual word lists and parallel corpora together in a way outperformed using only one of them. This is done by linearly combining translation probabilities with these two sources using equal weights. This approach was tested on TREC5C, TREC9X and TREC5S and reported to achieve 90% of the monolingual system in Chinese documents and 85% in Spanish ones. Levow et al. [2005] designed CLIR system in which the queries were provided in English and the collection was in French and Mandarin Chinese. For testing, they used CLEF 2000 French documents, TREC-2002 CLIR Arabic test collection, TREC-5 Chinese documents and CLEF 2000 German subset. They presented three approaches to use query alternative translations in retrieval:

1) Combining the alternative translations with their original weights, they called this approach unbalanced translation.

2) Balanced translation: Each query term transla-

tion is mapped to a weight that is based on scores calculated from the documents. To weight a query term, they use the term frequencies in retrieved documents and document frequencies.

3) Structured query translation: The translations of query terms which occur in the documents are assigned weights based on term frequency and document frequency. Structured query approach significantly outperformed the other two methods in all languages.

The morphological analysis turned to help solving the ambiguity that happens in the translation step. Levow et al. [2005] presented the effect of morphological processing for different languages on the translation process and the retrieval quality at the end. For Arabic, they applied stemming using light stemmer and morphological analyzer. For German, they processed compounded words. As it is known, German uses compounded words by concatenating terms together in one word without using white spaces. E.g *Berlin Kenner* means a person who knows Berlin. It is written in one word *Berlinkenner*. This is considered to be a challenge for translation, so Levow et al. [2005] used dictionary-based greedy approach to decompound the compounded terms before translation. Also they applied word segmentation algorithm for Chinese and simple rule-based stemming for French.

Ture et al. [2012] were among the firsts who exploited multiple query translations provided by SMT (often called n*-best-list*). They improved the previous work on probabilistic structured queries [Darwish and Oard, 2003], where query terms were represented by a probability distribution over its translations, by estimating the term translation probabilities from the n-best-lists.

Nikoulina et al. [2012] developed a model to rerank the *k-best* translations which are provided by an SMT system for German and French languages.

The first method they presented addressed translation quality and the second one addressed retrieval quality. For their machine learning algorithm, they employed Margin Infused Relaxed Algorithm (MIRA), an online learning algorithm for multiclass categorisation problem [Crammer and Singer, 2003], and for training they used queries that are provided by CLEF tracks. They used features based on dependency structure of the translated query and the original one, also they used features based on Part Of Speech Tags between the query and the alternative translations. For each query translation hypothesis, first, a list of features is generated and then the model weights are trained towards MAP, so the model is expected to predict the hypothesis that gives the best retrieval quality. The model improved the CLIR performance between 1% and 2.5% on the CLEF AdHocTEL 2009 task (French to German) [Macdonald et al., 2006]. Also they showed that simply concatenating *5-best-list* hypotheses from the SMT output ,as a special way of query expansion, improved the baseline as well.

The work presented by Ture and Boschee [2014] converted the problem of reranking into classification problem. They built a set of binary classifiers to produce *query-specific* weights of various different features to select optimal translations from the n-best-lists. In order to train their system they used surface features, parsing-based features, features consist of statistics of query and its translation and collection based features. They showed that using full combination of evidences for each query outperformed partial combination, and as a result, outperformed the baseline system.

However, all of the above mentioned approaches are based on machine learning approaches, which require large data for training. To overcome the challenge of data unavailability in CLIR, Schamoni and Riezler [2015] extracted relevance information from links between the Wikipedia articles that are written in different languages and the citations of the patents. From Wikipedia they used for training 4.1M of English and German parallel sentences that were provided by WMT [6] and 1.8M sentences from the NTCIR-7 JP-EN PatentMT sub-task. They showed that combining systems that are very dissimilar outperformed combining systems that give best performance by more than 10 points of MAP and NDCG.

3.7 CLIR tracks

Cross Languages Evaluation Forum (CLEF) is considered to be one of very first campaign in the field of CLIR, it started in 1990s. One of its initiative is CLEF eHealth tracks which has started for the first time in 2013. It aims to improve the retrieval that is

[6] http://statmt.org/wmt11/translation-task.html

conducted by users on the medical domain who are searching for topics related to health issues. In 2014, the task introduced a CLIR sub-task in which participants were given translated queries together with the monolingual English queries. More details about the collection, queries and evaluation process in this task will be provided later in Section 4.1. We participated in the eHealth IR task including its monolingual and cross-lingual subtasks in 2014, 2015 and 2016.

While Text REtrieval Conference (TREC) tracks [7] were started in 1994, it includes multiple tracks for different purpose (creating test collections, evaluation tools, etc.), and mainly focuses on information retrieval researches. TREC also organised cross-lingual tracks appeared in TREC6, TREC7, TREC8 and TREC9 which was the last one in 2002.

3.8 Word Embeddings approaches

Latent Semantic Analysis (LSA) is considered to be the first approach that discovers the relations between words in a semantic space. The main hypothesis that LSA depends on is that similar words appear in the same parts of text (paragraph). To reduce the dimensions of the text, LSA uses singular value decomposition (SVD). SVD is based on converting matrix (eventually two dimensions) into a product of three different matrices. For example, in the case of information retrieval, we can build a matrix that contains documents as columns and terms as rows, and each cell in the matrix could tell if the term exists in that document or not. The decomposition of this matrix using SVD will give us a matrix contains concepts, a matrix represents the strength of these concepts and a matrix represents terms as concepts (moving into semantic space). The first challenge that will come to mind is loading the entire collection into memory will be impossible for a big corpus. This issue was solved by the work of Řehůřek and Sojka [2010]. They presented a novel framework (gensim) that topically models the documents using wide set of algorithms including LSA and LDA algorithms. A state-of-art approach was presented by Mikolov et al. [2013]. They presented two models: Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram model. CBOW predicts words given a context. The range of the context is called windows size (c). While Skip-gram model predicts a context of size (c) for a given word. After representing words as vectors, the model is evaluated using algebraic operations to answer both semantic and syntactic questions. Skip-gram outperformed CBOW model on the semantic questions set but CBOW outperformed it on the syntactic set. These two algorithms and a vector representations of words from the training croups is presented as a neural-network based open source tool called *word2vec*.

Roy et al. [2016] presented a method that represents both documents and queries as a set of word vectors. After that, the similarities between a given query and documents can be calculated using any well-know similarity function like cosine similarity. The vector-based similarity is then combined with text-based similarity to rank the documents for a given query. For their experiments, the used TREC6, TREC7, TREC8 and TREC Robust dataset, and Lucene for indexing and retrieving from the collection. Their experiments showed that the hybrid model outperformed the text-based LM model on Robust and TREC-8 collection when using *K=100* clusters.However, the experiments also showed that when using one cluster (single-point representation) for representing the documents, results were similar to the text-based model. This might be occurred because using 100 clusters to represent the collection is very small. Kuzi et al. [2016] used word-embeddings to expand the query with terms from the collection. First, they trained word2vec on the entire collection (WSJ, AP, Robust, WT10G and GOV2) which contains about 28 millions documents. A candidate term is scored using its semantic similarity with a given query by calculating cosine similarity between that term and the query centroid. Results showed that the expanded query outperformed the original query in terms of MAP and P@5. However, word2vec does not have to be trained on the same collection that we use for the retrieval as was shown by the work of Zamani and Croft [2016]. In which they trained the model on a collection that is different from the IR collection. Then they used the model to expand the queries with candidates terms that are chosen by the model. Their method showed to be an effective method for query expansion.

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

Kim et al. [2016] used word-embeddings to calculate the similarity between documents and a given query. First, they used inverse document frequency to weight query terms. Then query terms were mapped to the most similar terms in a document based on word-embeddings. Finally For document scoring, they used cosine similarity between query terms and document terms. They trained *word2vec* model on 25 millions articles from PubMed using their titles and abstracts and made the model available online [8].

4 System description

In this section we will describe our set of experiments that we have done so far. Our experiments are conducted on the CLEF eHealth IR tasks data and queries set. First, we use the monolingual English queries to design state-of-art monolingual retrieval system. Then we investigate the retrieval performance on the concept level by annotating the collection and queries with concepts. Then we present our CLIR system that uses regression model to select the best translation to be used for the retrieval and how we adapt the model for new languages. Finally, we present our plan for the future work.

4.1 Data

We use in our experiments data taken from CLEF 2015 eHealth Task 2: User-Centred Health Information Retrieval [Goeuriot et al., 2015], the document collection is given by the Khresmoi project[9]. It contains web pages (HTML documents) automatically crawled from medical-domain websites. The collection consists of 1,096,879 documents containing the total of 1,111,711,884 tokens (after filtering). The average length of a document is 6,316 tokens.

4.1.1 Document processing

The documents in the collection are provided as raw web pages including all the HTML markup and eventually also CSS style definitions and Javascript code which should be removed before indexing. The collection in CLEF eHealth 2014 IR task was identical to the collection in the year of 2015. In 2014 during our participation in that task, we employed three data cleaning methods and evaluated their effect on the retrieval quality measured on the training queries.

First, we simply removed all markup, style definitions, and script code by the Perl module HTML-Strip[10] (but keep meta keywords and meta description tags). This reduces the total size of the collection from 41,628 MB to 6,821 MB, which is about 16% of the original size. The average document length is 911 tokens (words and punctuation marks). Although the size reduction is very substantial, the resulting documents still contained a lot of noise (such as web page menus, navigation bars, various headers and footers), which is likely not to be relevant to the main content of the page. This noise is often called boilerplate. We used two methods to remove it: Boilerpipe [Kohlschütter et al., 2010] reduced the total number of tokens in the collection by additional 58% (the average document length is 383 tokens) and JusText [Pomikalek, 2001] by 55% (the average document length is 409 tokens). Surprisingly, the most effective method is the simple HTML-Strip tool. The two other methods are probably too aggressive and remove some relevant material important for IR. In all the following experiments, the collection is cleaned by HTML-Strip.

4.1.2 Queries

In the rest of experiments, we use multilingual queries from CLEF 2013–2015 eHealth information retrieval tasks. The English queries were constructed by medical professionals. Then, the queries were translated into Czech, French and German by native speaking medical professionals in these languages. Table 1 shows statistics about the query set. Queries from the years of 2013 and 2014 were generated from discharge summaries which include medical specifications for the patient cases involving diagnostic procedures and the treatment results. Thus, queries in these years are more likely to include medical terms written by medical experts. On the other hand, queries in 2015 were created in different way, non-medical expert people created the queries by describing their symptoms using non-medical terms, e.g. to describe the *jaundice*, laypeople might use words like *white part of eye turned*

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

[9] http://khresmoi.eu/

[10] http://search.cpan.org/dist/HTML-Strip/Strip.pm

query set	size		#rel	#¬rel
2013 ShARE/CLEF eHealth, Task 3	50	4.0	23.0	70.5
2014 ShARE/CLEF eHealth, Task 3	50	4.2	64.2	71.9
2015 CLEF eHealth, Task 2	66	4.5	35.4	145.1

Table 1: Query sets statistics: size, average length, number of relevant and irrelevant docs per query

query id	title
2013.02	Facial cuts and scar tissue
2013.41	right macular hemorrhage
2013.30	metabolic acidosis
2014.04	Anoxic brain injury
2014.21	renal failure
2014.17	chronic duodenal ulcer
2015.08	cloudy cornea and vision problem
2015.59	heavy and squeaky breath
2015.48	cannot stop moving my eyes medical condition

Table 2: CLEF 2013–2015 eHealth query samples

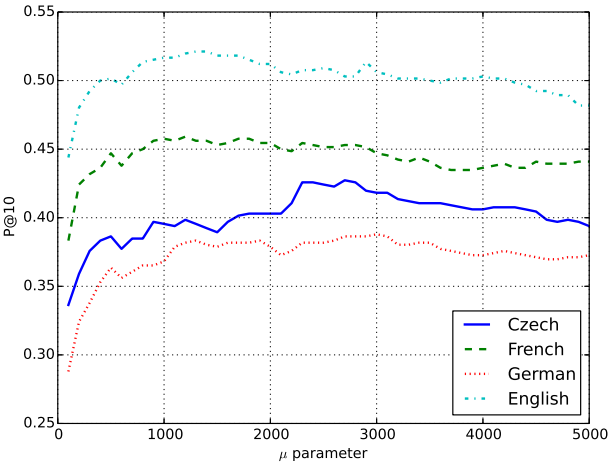

Figure 2: Tuning μ parameter

green. In order to prevent our system from biasing to one type of queries, we create two sets of queries, one for training and one for testing, which include queries from all years. First we created pool contains all queries, then we randomly split the queries set into 100 queries for training and 66 queries for testing.

Table 2 shows samples of queries from CLEF 2013, 2014 and 2015 eHealth tasks. More information about the collection and the queries can be found in Goeuriot et al. [2015], Goeuriot et al. [2014] and Suominen et al. [2013].

4.2 Retrieval system

We use Terrier, an open source information retrieval system that was developed by Ounis et al. [2006], to index the collection and to conduct the retrieval. Based on different experiments in which we compare different models and parameters, we decide to use Terrier's implementation of Bayesian smoothing with Dirichlet prior weighting retrieval model (we will refer to it by Dirichlet). This retrieval model is based on language modeling approach. The documents are scored by calculating the product of each term's probability in the query using the language model for that document. Term probabilities in a document are estimated by maximum likelihood estimation which could be zero when a query term does not appear in the document. Many smoothing methods are used to avoid zero probabilities as shown by Zhai and Lafferty [2004]. Dirichlet retrieval model employs Bayesian smoothing with Dirichlet prior by using different amount of smoothing based on the length of the document. Smoothing parameter for longer documents will be smaller. In this model we apply the Equation 1 to smooth the probability of term w in document d.

$$p_u(w|d) = \frac{c(w;d) + \mu p(w|C)}{\sum_{w' \in v} c(w';d) + \mu} \qquad (1)$$

Dirichlet model has one parameter (μ), by default it is setup to 2500 in Terrier. Since the optimal value for μ depends on the collection (length of documents), we tune this value on the collection toward highest P@10, after taking the monolingual English queries and *1-best-list* hypotheses from all languages. Figure 2 shows tuning μ for all languages, this experiment supports the decision to setup the parameter to 2500, even this value is not optimal for individual language (English), where $\mu = 1400$ gives $P@10 = 0.5212$ comparing to $\mu = 2500$ gives $P@10 = 0.5091$, but taking $\mu = 2500$ gives more stable results and highest averaged P@10 for all languages. For more details and comparison with another smoothing methods, see the work that was done by Smucker and Allan [2005].

For system evaluations we mainly consider precision at 10 (P@10), which corresponds to the percent of relevant documents at the first 10 retrieved documents. Also we report in our experiments the normalised discounted cumulative gain at 10 (nDCG@10) [Järvelin and Kekäläinen, 2002] and MAP [Pecina et al., 2014].

We evaluate our systems using trec evaluation tool [11], which assumes the non assessed documents as non relevant documents, for this reason and since it is very likely to retrieved non assessed documents; we use another metric related to the relevance information which is CVR@10. It is the percent of assessed documents in the top 10 ranked retrieved documents. CVR@10 gives us better glance about the performance of the systems if we had full assessment information.

Relevance information, which was provided by the CLEF eHealth orgnisers, covered about 80% of the top 10 ranked documents that are retrieved by our baseline system for all languages. In order to have these documents fully assessed, we sent our system runs to medical experts and asked them to assess [12] all the top 10 retrieved documents, so we reached almost 100% coverage of the top 10 ranked retrieved documents for all languages.

For the significance test between different systems we use Wilcoxon test with $\alpha = 0.05$ [Hull, 1993], which is used to compare two independent samples with paired data taken from the same distribution.

4.3 Concept-level retrieval

CLIR systems that are based on terms-matching approaches, suffer from the problem when a term has different synonyms. Retrieval system will fail when one concept appears in the translated query as a different synonym from the same concept in the collection. This will hurt the retrieval results, and in order to tackle this problem, we aim in this experiment to represent the collection and the queries using the same conceptual space.

After cleaning the documents, we use MetaMap to annotate the data with concept identifiers from the UMLS Metathesaurus [Humphreys et al., 1998; Bodenreider, 2004] version 2014AA. Annotation process is considered to be a heavy one, where annotating about one million documents using 200 CPUs in our university cluster required us about one week to get all the collection annotated. The UMLS Metathesaurus is a large vocabulary database containing information about biomedical and health-related concepts, their names and relationship between them. Terms are linked to others by various relationships such as synonymy, hypernymy, hyponymy, lexical variations, and many others. The Metathesaurus is organized by concept, which symbolizes a semantic concept or a meaning. Each concept or meaning in the Metathesaurus has a unique and permanent concept identifier (CUI). We utilize MetaMap's highly configurable options in our annotation process. We use the `-I` option so that the concept IDs are shown, and `-y` option to enable word sense disambiguation. The text is broken down into the components that include sentences, phrases, lexical elements and tokens. The disambiguation modules then process the variants and output a final mapping. We put this concept annotations into an additional XML field in the document and query files. An example of cleaned and annotated document is given in Figure 3.

The main challenges of this method are:

1) Annotating text with concepts, as we did in the medical domain text, is not easy for a text that is taken from a generic domain, since there is not similar tool to MetaMap to do so.

2) Even when it comes to the medical domain,

[11] `http://trec.nist.gov/trec_eval/`

[12] Either a document is relevant to a query, irrelevant or somewhat relevant

```
<doc>
  <docid>wiki.0842_12_009733</docid>
  <title>
    Testing for Celiac Disease ...
  </title>
  <title_concepts>
    C0683443
    C0007570
    C0521125
    ...
  </title_concepts>
  <text>
    Intestinal biopsy is the gold
    standard for diagnosing celiac ...
  </text>
  <text_concepts>
   C1704732
   C0036563
   C0423896
   ...
  </text_concepts>
</doc>
```

Figure 3: An example of an annotated document

sometimes laypeople pose a query that does not contain medical terms. In such a case, MetaMap probably will fail to annotate the query with relevant concepts. This will cause to information losing in the text.

The retrieval on the concept level led us to very bad results, less than 50% of the baseline system. Our explanation for this result is that MetaMap could not annotate the document and the query with similar concepts. This caused information loose.

4.4 Translation system

The SMT system employed in our experiments was developed within the Khresmoi[13] project [Dušek et al., 2014] as a part of a large-scale multi-lingual multi-modal search and access system for biomedical information and documents. The SMT system is built on Moses, a state-of-the-art phrase-based SMT system which was introduced by Koehn et al. [2007], and adapted to translate texts from the medical domain. It is available for three language pairs (Czech–English, French–English and German–English) and supports translation of standard sentences and search queries. In a phrase-based SMT, the output translation is constructed from possible translations of subsequences of consecutive words (phrases) in the input sentence during encoding. The best translation is searched for by maximizing the probability of the output given the input formulated as a log-linear combination of several feature functions. They include the following scores:

- Direct and inverse phrase translation probability.
- Direct and inverse lexical weighting which estimates the probability of a phrase pair.
- Phrase penalty which penalises the length of the phrase.
- Word penalty penalises translations of phrases which differ in length with the source segment.
- Word distortion is the amount of word reordering between the translation and the source segment.
- Language model estimates the probability of translation, this model is trained using monolingual data.
- Sentence score which is the final combination of the above mentioned features.

For an input sentence, Moses can return list of hypotheses sorted by their final scores, we will refer to this list by *n-best-list*.

The models of the Khresmoi SMT system were trained on a combination of general-domain data (e.g., EuroParl, JRC Acquis, or News Commentary corpus) and medical-domain data (e.g., EMEA, PatTR, COPPA, or UMLS), more details about data set are presented by Pecina et al. [2014]. The query translation system was designed to translate short and rather ungrammatical sequences of terms typical for search queries. The feature weights were not optimized towards the traditional translation quality usually measured by BLEU (Bilingual Evaluation Understudy [Papineni et al., 2002]) but towards PER (Position-independent word Error Rate [Tillmann et al., 1997]), which does not penalise word order and was shown to be more adequate for tuning SMT for search queries [Pecina et al., 2014].

4.5 Monolingual system

It is generally useful to compare CLIR results to monolingual results obtained by using manual translations of the queries into the document language. This also sets a "soft upper bound" of the cross-lingual results. The "monolingual" P@10 score is 47.10% for the training queries and 50.30% for the

[13] http://www.khresmoi.eu/

test queries. In the cross-lingual experiments we would like to get as close to this value as possible for all languages. The complete monolingual results on the test set are shown in Table 3 (row denoted as *Mono*). This system uses Dirichlet retrieval model on the cleaned collection, and it returns top $k = 1000$ ranked documents.

4.6 Baseline

Our baseline is the system which accepts the single best translation as provided by the Khresmoi SMT system. Results of the baseline systems are presented in Table 3 (row *Baseline*). On the training queries, the P@10 values of those systems are 41.90 for German to, 45.30 for French, and 46.00 for Czech. On the test queries, the P@10 values range from 42.42 for German to 47.73 for French, with Czech in between with 45.61. We should emphasize that the baseline is quite strong. Compared to generic translation systems, the Khresmoi system is specifically adapted to the medical domain and tuned to translate queries for CLIR [Pecina et al., 2014]. Therefore, the relative performance w.r.t. the monolingual results is as high as 84%–94% (depending on the source language).

4.7 Oracle experiments

The main hypothesis in this research is that an SMT system produces n-best-lists that is not reranked perfectly for CLIR. To confirm this hypothesis and to show that a perfect reranking of SMT n-best-lists can improve CLIR quality, we performed the following experiments: For each query in the training data we selected the translation hypothesis with the highest P@10 and averaged those values to get the maximum (oracle) score of P@10 achievable if the reranking method always selects the best translation. On the training data, the oracle score would be 55.10 for Czech, 58.90 for French, and 52.70 for German. This result is very encouraging and confirms that there is enough potential space for improvement. The baseline scores could be improved by 11.67 on average.

A deeper analysis of this observation is illustrated in Figure 4. The two plots visualize distribution of the best translations (highest P@10) in the 20-best-lists for all training queries (per language). The first plot shows histograms of the top ranks with the best translations. Here, for about 45% of the queries, the best translations are ranked as first. For the remaining 55% queries, the first best translations are ranked lower. Those are the cases, which can be improved by better ranking. The second plot displays the histogram for all hypotheses with the highest P@10 (not just the top ones). For each query there are multiple translations which can be selected to achieve optimal performance.

4.8 n-best list merging

Nikoulina et al. [2012] presented a method combining n-best-list translations by trivial concatenation of 5 top translations as produced by SMT. This approach completely failed on our data (all languages) and did not improve the baseline for any value of n from 0 to 20 (on the training data and the test data). Figure 5 shows how concatenating translations to create queries degrades the performance. Results of the 5-best-list concatenation on the test data are shown in Table 3 (row *5-best*).

4.9 Document translation experiments

The question whether we should translate the document collection or the queries was investigated before as we showed previously in this research. However, the mentioned experiments were outdated. Since the statistical machine translation systems showed significant improvement last few years, we decided to reinvestigate the question again. Firstly, we preprocessed the data collection by tokenising the text, lowercasing it and splitting its sentences so every sentence contains 50 words maximum (this is done due to the maximum length limitation in the Moses decoder). Secondly, we used around 800 CPUs at out department's cluster to translate the documents from English into four languages (Spanish, Hungarian, Polish and Swedish). It took us one week to finish the experiment. After that, we indexed the translated documents and built four indexes, one for each language. Finally, we used the human translated queries and conducted the retrieval using the same setup (highest 1000 ranked documents using Dirichlet LM IR model). Table 5 shows the document translation (DT) results on the test set. DT system does not outperform any system when we use query translation (QT) approach (our baseline which uses *1-best-list* from the SMT system).

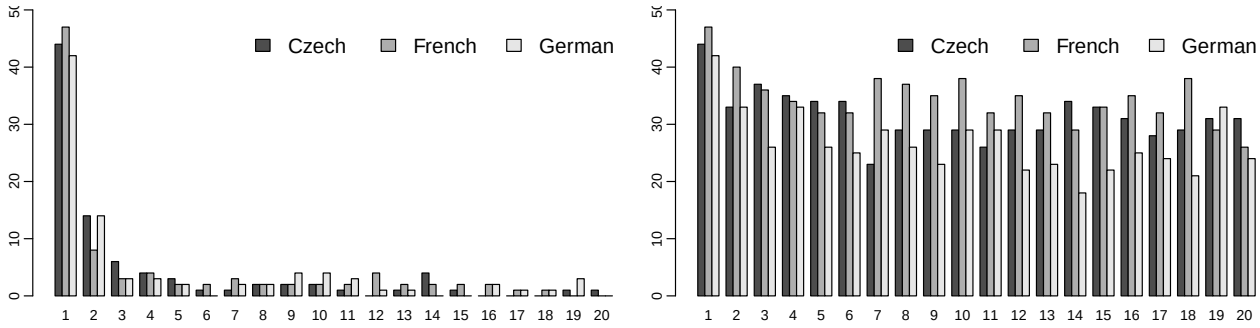

Figure 4: Histograms of ranks of translation hypotheses with the highest P@10 for each training query: the first such ranks only (left), all such ranks (right).

Table 3: Complete results of the final evaluation on the test set queries

	Czech			French			German		
system	P@10	NDCG@10	MAP	P@10	NDCG@10	MAP	P@10	NDCG@10	MAP
Mono	50.30	29.97	99.85	50.30	29.97	99.85	50.30	29.97	99.85
Baseline	45.61	38.57	23.58	47.73	41.11	25.72	42.42	36.47	22.74
5-best	38.94	33.01	22.30	41.06	37.20	23.05	30.45	30.16	17.28
SMT	44.70	37.92	24.77	48.79	42.85	25.81	42.73	37.88	22.65
+RANK	48.64	**41.63**	**25.73**	48.48	43.83	26.07	44.55	**40.76**	24.09
++IDF	48.03	41.06	25.22	48.64	43.83	26.10	44.39	40.71	**24.11**
++BRF	47.27	40.52	24.99	49.70	44.12	26.64	43.64	39.81	23.76
++TP	45.76	39.92	23.74	48.48	43.88	26.26	44.39	40.41	24.07
++WIKI	48.64	**41.63**	**25.73**	49.24	44.00	26.36	43.64	39.81	23.76
++UMLS	48.64	**41.63**	**25.73**	49.09	44.10	26.09	44.55	**40.76**	24.09
++RSV	48.64	**41.63**	25.66	48.94	43.84	25.95	43.03	39.20	23.55
ALL	**50.15**	40.72	**25.73**	**51.06**	**46.49**	**27.86**	**45.30**	39.47	23.71
QE	36.67	34.12	20.27	38.03	35.19	20.97	33.79	32.02	18.87
Google	50.91	39.98	26.93	49.70	43.88	26.36	49.39	42.77	26.87
Bing	47.88	40.51	25.22	48.64	42.75	26.43	46.52	41.69	25.04

Coverage rate (CVG) ,which is the percent of assessed documents in the highest 10 retrieved documents, is much lower in the DT documents. In average, the coverage in the DT experiments in all languages is around 50%, which means that we have 5 unassessed documents in each query out of 10 documents. These documents are treated as irrelevant documents by the evaluation tool (trec evaluation tool). Assessing these documents might improve the results, if we can ask annotators to finish the assessment for the DT runs, then we will have more strong say whether QT is better than DT or not. However, we started few days ago translating the collection into these four languages in addition to Czech, French and German, after new SMT systems have been released recently by a team in our department. So the document translation experiments are still ongoing.

4.10 Our method

Our method employs SMT to translate queries (in Czech, German, French) into the language of documents (English). However, we do not rely on the SMT decoder to select the best translation variant. Instead, we obtain multiple top-scored hypotheses (n-best-list) and rerank then w.r.t. the retrieval objective. The highest-ranked hypothesis is then used to query the document collection.

Formally, for each query q_i, its each translation hypothesis $q_{i,j}$ is represented by a vector of features

Table 4: Documents translation (QT) and query translation (QT), evaluation on the training set queries

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Mono	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90
QT	41.90	22.02	78.80	40.10	20.67	74.40	**37.30**	**19.55**	72.80	39.60	20.07	73.80
DT	37.70	20.56	63.90	29.20	14.47	52.30	25.70	13.35	48.50	29.30	14.93	51.30

Table 5: Documents translation (DT) and query translation (QT), evaluation on the test set queries

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Mono	50.30	29.97	99.85	50.30	29.97	99.85	50.30	29.97	99.85	47.10	25.90	99.90
QT	44.09	24.72	86.67	40.76	22.31	70.61	36.82	19.92	70.76	36.67	20.60	76.21
DT	38.94	22.22	65.15	25.30	13.13	41.67	19.55	11.05	35.76	27.73	14.45	55.61

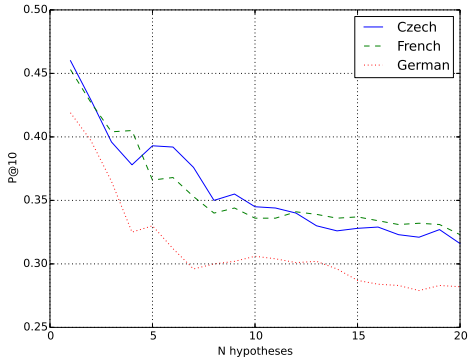

Figure 5: Baseline performance when concatenating N hypotheses

(predictors). For training queries, each hypothesis is assigned a score (response) equal to $1 - (O_j - P_{i,j})$, where $P_{i,j}$ is P@10 score of top 10 documents retrieved by the translation hypothesis $q_{i,j}$ and O_j is the maximum (oracle) P@10 of all the translation hypotheses of the query q_i. The response values are in the range of $\langle 0, 1 \rangle$, where 1 indicates a good query translation and 0 a bad translation.

The reranker is trained by fitting a generalized linear regression model (GLM) with logit as the link function (ensuring the response to be in the $\langle 0, 1 \rangle$ interval) [McCullagh and Nelder, 1989]. For testing, translation hypotheses of the test queries are scored by this model and the highest-scored hypothesis is selected as the translation. We employed the GLM implementation in R[14] which optimizes the model parameters by the iteratively reweighted least squares algorithm. We need to mention that we conducted experiments to convert the problem into binary classifier, in which a classifier was trained to select the best translation (positive case) among the bad ones (negative cases), but this approach could not improve the results.

The features are extracted from various different sources and include:

SMT The main set of features are the eight scores from the SMT models plus the final translation score (see Section 4.4).

RANK Two features extracted from the original ranking – the rank itself and a binary feature indicating the top-ranked hypothesis.

IDF To distinguish translations containing informative terms, each hypothesis is scored by the sum and average of inverse document frequency of the terms.

BRF Motivated by the blind-relevance feedback approach for query expansion, a single best translation provided by SMT for each query is used to retrieve the 10 highest-ranked documents and each hypothesis is scored by the sum and average of term frequencies extracted from the retrieved documents.

[14] https://www.r-project.org/

TP Hypotheses of each query (n-best-list) are merged and each is scored by the sum and average of term frequencies extracted from the merged n-best-list.

WIKI Each hypothesis is scored by the sum and average of term frequencies extracted from abstracts of 10 Wikipedia articles retrieved as a response to the single best query translation provided by SMT (using our own index of abstracts of all English Wikipedia articles and the Terrier search engine).

UMLS Two features based on the UMLS Metathesaurus [Schuyler et al., 1993]: the number of UMLS concepts identified in each hypothesis by MetaMap [Aronson and Lang, 2010] (with word sense disambiguation and part-of-speech tagging on); the number of unigrams and bigrams which match entries in the UMLS Metathesaurus.

RSV Retrieval Status Value, a score assigned to the highest-ranked document by the retrieval system in the response to the query translation hypothesis.

4.10.1 Reranking

For each query (both in the training and test sets) we considered up to 15 best translation hypotheses (excluding duplicities). Queries with oracle P@10=0 were excluded from training. The training data then included 1,249 items for Czech, 1,181 for German, and 1,246 for French. We merged these data into one single training data set and trained a single language-independent model which proved to be a better solution than to train a specific model for each language. The training set included a total of 3,676 items of query translation hypotheses of the 100 original queries (each translated from Czech, German and French).

Prior training, the data was normalized to have sample mean equal to 0 and sample variance equal to 1. The test data was normalized using the same coefficients (those obtained on the training data).

The training data was first used in a leave-one-query-out-cross-validation fashion to tune the hyper-parameters (such as the type of the learning algorithm, the n-best-list size, and parameters of all the features). Then, all the training data was used to train a single model which was then applied to the 15-best-lists of the 66 test queries for each language.

In the remainder of this section, we first present some complementary experiments for comparison and then the main results of our method. We comment on the main evaluation measure (P@10) but the main results (Table 3) also includes scores of other measures (NDCG@10, MAP).

The reranking method described in Section 4.10 was tested with several combinations of features. The complete results are displayed in the middle section of Table 3. The figures in bold denote the best scores for each language and evaluation metric. All of those are statistically significantly better than the respective baselines (tested by Wilcoxon signed-rank test, α=0.05). For comparison, we also provide results of systems based on translation by two on-line translation tools: Google Translate and Bing Translator[15] (rows Google and Bing, respectively).

The system based only on the SMT features did not bring any substantial improvement over the baseline (row SMT) for any of the languages. P@10 improved by more than 1 point for French only. For Czech, the score decreased and for German, the difference was negligible. However, none of these differences was statistically significant. The transitional way of SMT tuning towards translation quality seems sufficient if no additional features are available. However, adding the explicit features derived from the SMT rankings helped a lot (row +RANK), especially for Czech and German, where the increase of the NDCG@10 scores was statistically significant.

The effect of the other features was studied independently by adding those features to the model with the SMT+RANK features. However, in terms of P@10, none of them brought any notable improvement. Although the BRF, WIKI, and UMLS features improved the results for French, they were not statistically significant even in comparison with the baseline.

The baseline, however, was outperformed by a statistically significant difference by systems combining all the features (row ALL). P@10 increased by 3.58 on average (which is a relative improvement

[15] https://www.bing.com/translator/

of 7.90%) In comparison with the monolingual results, the ALL system performed at 101.51% for French, 99.70% for Czech, and 90.05% for German. For French the system even outperformed the one based on translations by Google Translate.

In Figure 6, we present detailed comparison of the baseline results and the results of the best system (ALL). For each query in the test set, the plot displays the difference of P@10 obtained by the best system and the baseline system. Positive values denote improvement which was observed for a total of 9 queries in Czech, 15 queries in German, and 8 queries in French. Negative values denote degradation which was observed in 2 cases for Czech, 4 cases for German, and 3 cases for French. A good example of a query whose translation was improved is 2015.11 (reference translation: *white patchiness in mouth*). The Czech baseline translation *white coating mouth* improved to *white coating in oral cavity* (P@10 increased from 10.00 to 80.00) and the French baseline *white spots in the mouth* improved to *white patches in the mouth* (P@10 increased from 10.00 to 70.00). More examples are given in Table 6.

4.11 Adapting reranker for new languages

Later, after we finished our experiments with Czech, French and German, developing new SMT models in the Khresmoi project was completed including translation from Spanish, Hungarian, Polish and Swedish into English. We investigate adaptation of our method to allow reranking of query translations for four new languages (Spanish, Hungarian, Polish, Swedish). The baseline approach, where a single model is trained for each source language on query translations from that language, is compared with a model co-trained on translations from the three original languages. First, we asked medical experts to translate the monolingual English queries also into Spanish, Hungarian, Polish and Swedish. However, a complete relevance assessment (of top 10 documents in all the experiments) is available only for the three original languages. Results for the new languages are not completely assessed, where the ratio of unjudged documents among the top 10 retrieved documents is around 25%.

These translations are used to create one training file contains up to 1500 instances. Then we use these instances separately to build a language-specific models for each language. We experiment with three systems: 1) A system which only uses features derived from the SMT system (denoted as SMT). 2) A system which combines the SMT features and features that are based on the original ranking of the translations (denoted as SMT+Rank). 3) A system exploits all features that we presented before (denoted as ALL).

Results of the LOOCV evaluation of language-specific models on the training set are shown in Table 8. The italics font refer to results statistically significantly different from the baseline system. All systems are able to significantly outperform the baseline system for the Spanish language only, and the system which uses all features (ALL), gives the best result. When testing the model against the test set, see Table 7, we do not observe any improvement in any language.

The features we presented are source-language-independent, which makes merging data from different languages possible in order to train the machine learning model and expand the training data. For each of the new languages (Spanish, Hungarian, Polish, and Swedish), we merge the translations from that specific language with the available training data for the the original language (Czech, German, French) to create a richer training data set. Results for all systems that use the merged data in the LOOCV experiments are shown in Table 9. For the Spanish and Hungarian languages, the system combining all features (ALL) significantly outperforms the baseline system. A small and not statistically significant improvement is observed for Swedish by the system based on all the features and for Polish by the system based on the SMT features only (SMT).

Table 3 shows the results for the models that are trained on expanded data set against the test set. Systems exploiting all the features (ALL) in Spanish and Hungarian outperforms the baseline system significantly. We observe in the test results by our best system (ALL) 11 queries improved in Spanish, 8 in Hungarian, 5 in Polish and 7 in Swedish. Also there are degradations of 5 queries in Spanish, 3 in Hungarian, 5 in Polish and 3 in Swedish. The impact of untranslated terms appears mostly in the Polish language.

For example, query *2015.37: łuszcząca skin* has P@10 = 00.00, (reference translation: *scaly skin*).

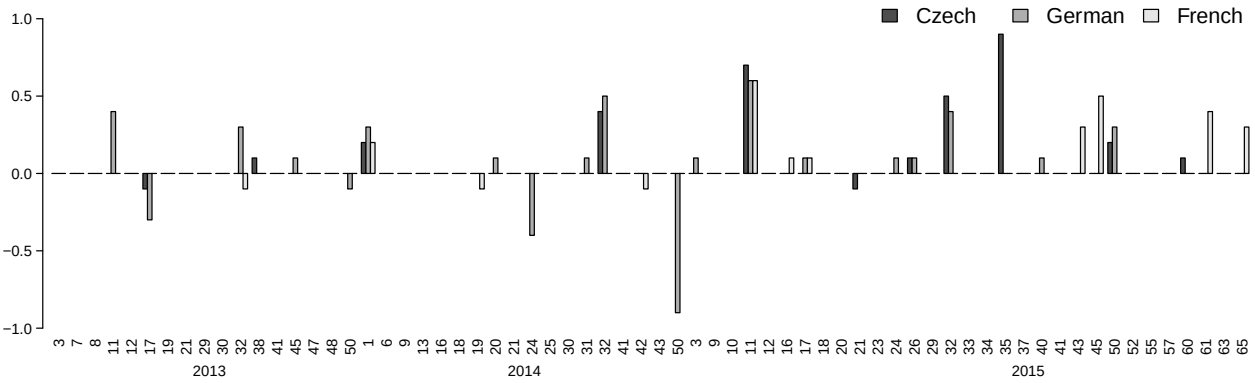

Figure 6: Per-query results on the test set. The bars represent absolute difference of P@10 of the best system (ALL) and the baseline system for each query and each language.

Table 6: Examples of translations of training queries including reference (*ref*), oracle (*ora*), baseline (*base*) and best (*best*) translations (system using all features). The scores in parentheses refer to query P@10 scores.

Query: 2013.02 (German)
ref: facial cuts and scar tissue (30.00)
ora: cut face scar tissue (80.00)
base: cut face scar tissue (80.00)
best: face cuts and scar tissue (80.00)
Query: 2013.42 (French)
ref: copd (70.00)
ora: disease copd (90.00)
base: copd (70.00)
best: disease copd (90.00)
Query: 2014.5 (German)
ref: bleeding after hip operation (60.00)
ora: bleeding after hip surgery (80.00)
base: bleeding after hip surgery (80.00)
best: hemorrhage after hip operation (50.00)
Query: 2015.53 (Czech)
ref: swollen legs (10.00)
ora: leg swelling (80.00)
base: swollen lower limb (40.00)
best: swollen lower limb (40.00)

It contains the untranslated term *łuszcząca*, which means *scaly* in English. The monolingual query has P@10 = 99.00, the difference in performance is caused by the untranslated (out-of-vocabulary, OOV) words only.

A similar situation appears in query *2015.35* (Monolingual English query: *lot of irritation with contact lenses*), its P@10 = 00.00. The translated query is *significant irritation szkłami kontaktowymi*. It contains two untranslated terms: *szkłami* (lenses) and *kontaktowymi* (contact). These two untranslated words destroy the query.

Query *2015.29* in Spanish has P@10 = 30.00 in the baseline, its translation is *red patch on the skin and dry pus*. The (ALL) system improves it to P@10 = 90.00 and selects the translation *red patch on the skin and dry pus blister*.

Another example of improvement is observed in query *2013.32*, the baseline translation is *dyspnoea* with P@10 = 60.00, the selected translation is *shortness of breath* with P@10 = 90.00. The reference translation is *SOB* with P@10 = 50.00, and this is one case in which the best system outperforms not only the baseline system but also the monolingual one.

We find in the translated Spanish queries that we have a total of 10 terms which the SMT system was unable to translate (OOV) which harm the performance. There are also 20 OOVs in the Hungarian queries, while in the Swedish and Polish the case is worse, where there are 40 OOVs in Swedish and 54 OOVs in Polish queries. Usually the SMT system can not translate some terms because they did not appear in the parallel data. So for our case, having this OOVs translated into English correctly definitely improves the retrieval results. The prob-

Table 7: Final evaluation results of language-specific models on the test set

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Monolingual	50.30	29.97	99.85	50.30	29.97	99.85	50.30	29.97	99.85	47.10	25.90	99.90
Baseline	**44.09**	**24.72**	86.67	40.76	22.31	70.61	36.82	19.92	70.76	36.67	**20.60**	76.21
SMT	43.18	23.96	86.97	**42.58**	**22.98**	90.45	36.06	19.24	85.30	**37.12**	19.69	89.24
SMT+Rank	42.88	23.90	87.12	40.76	22.31	89.70	**38.33**	**20.57**	91.52	36.52	20.16	90.91
ALL	43.33	23.71	88.48	40.00	21.80	88.64	37.73	20.16	90.00	36.21	20.49	88.03

Table 8: Cross-validation of language-specific models on the training set

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Monolingual	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90
Oracle	52.10	25.86	86.50	51.80	24.55	79.40	47.40	22.16	77.50	50.10	23.23	78.70
Baseline	41.90	22.02	78.80	40.10	20.67	74.40	**37.30**	**19.55**	72.80	39.60	20.07	73.80
SMT	*44.30*	22.65	89.10	40.40	20.39	87.30	34.90	17.51	79.00	38.90	19.00	87.80
SMT+Rank	*43.00*	22.46	87.80	40.50	20.63	88.30	35.70	18.18	81.40	38.80	19.83	86.90
ALL	***45.30***	**23.91**	90.30	**41.30**	**21.59**	89.40	35.50	18.38	82.10	**39.80**	**20.07**	87.00

lem of OOVs in the CLIR can be a subject of further investigation in the future. In these experiments, we presented our approach to adapt our SMT query translation reranker in the cross-lingual information retrieval task to new languages. The new languages suffer from low assessment coverage in their baseline systems, leading to low quality data to train the reranker model with. Our approach tackled this problem by exploiting data which were taken from languages whose retrieval systems were fully assessed (Czech, French and German) by medical experts who we asked to do so. Data were merged from the fully assessed languages together with the data from one new language in order to train the model. Firstly, we created one training set for each of the new languages to build a source-language-specific model. This approach could not bring significant improvement to the baseline system. Then, we used the expanded data to train reranker models for the new languages. This approach significantly outperformed the baseline systems for Spanish and Hungarian. However, it could not significantly improve the baseline in Polish and Swedish, because the SMT system produces high number of OOVs in these languages comparing to Spanish and Hungarian.

4.11.1 Query expansion

Query expansion (QE) is a well-known method in IR that is used to expand the original user's query with related terms from the collection. The main motivation here is that when users pose a query to look-up their information need, they are usually not aware of all correct or related terms. So, we aim to select additional terms that are added to their query before conducting the retrieval from the collection. Before applying our method we prepare two things: 1) *word2vec* that is trained on PubMed articles as provided by Kim et al. [2016]. We trained the model on our collection which contains about one million documents, but using the model that is trained on PubMed collection showed to be better than our version, since the PubMed data is bigger than CLEF eHealth data collection. 2) A flat list of medical terms and their synonyms as provided by Medline-Plus dictionary [16].

To perform QE, we apply the following:

- We get the *1-best-list* translation for each query and consider it to be the original query.
- Then we get the embeddings for each term (as a vector) using *word2vec*: $Q =$

[16] https://medlineplus.gov

Table 9: Cross-validation of language-independent models on the training set

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Monolingual	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90	47.10	25.90	99.90
Oracle	52.10	25.86	86.50	51.80	24.55	79.40	47.40	22.16	77.50	50.10	23.23	78.70
Baseline	41.90	0.2202	78.80	40.10	20.67	74.40	37.30	**19.55**	72.80	39.60	20.07	73.80
SMT	43.40	22.36	89.10	38.40	19.13	84.00	**38.70**	18.41	81.80	36.10	17.91	85.10
SMT+Rank	43.10	22.49	88.00	40.70	20.78	88.50	36.50	18.87	82.10	39.00	19.75	86.60
ALL	***46.90***	**24.05**	90.80	***42.20***	**21.91**	89.00	36.90	18.61	82.60	**40.00**	**20.12**	86.70

Table 10: Final evaluation results of language-independent models on the test set

	Spanish			Hungarian			Polish			Swedish		
system	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG	P@10	MAP	CVG
Monolingual	50.30	29.97	99.85	50.30	29.97	99.85	50.30	29.97	99.85	47.10	25.90	99.90
Baseline	44.09	24.72	86.67	40.76	22.31	70.61	36.82	19.92	70.76	36.67	20.60	76.21
SMT	43.79	23.83	87.42	40.00	22.54	89.09	35.61	19.76	85.61	38.33	19.85	88.64
SMT+Rank	43.64	24.28	86.36	38.94	21.91	89.09	**38.18**	20.21	89.24	36.21	20.02	91.97
ALL	***46.36***	**25.30**	90.15	***43.18***	**23.88**	91.67	36.67	**20.38**	89.39	**38.79**	**21.06**	91.97

$\{V_1, V_2, .., V_i, .., V_m\}$; where V_i is the vector of the term T_i in the original query and m is the number of terms in the query.

- For each V_i, we calculate the distance between that vector and all the term's vector in the collection that is used to train *word2vec* using cosine similarity, $C = \{V'_1, V'_2, ., V'_i, .., V'_n\}$; where V'_i is the vector of the term V'_i in the collection and n is the number of terms in the collection.

- After creating a list of terms which have highest similarity scores with the query terms, we choose a term that exists in MedlinePlus list. If all the candidate terms do not exist in that list, we take the one with the highest score. Also if multiple terms appear in the list we choose only one that has the highest score.

Results from QE experiment are shown in Table 3. In the German language, QE improved 14 queries out of 100, 25 queries remained unchanged (in terms of P@10) and 27 queries destroyed. One example of the improved queries is the query: *qtest2014.18*. The *1-best-list* for this query is: *dizziness and hypotension*. The best term that has the highest similarity score with *dizziness* and exists in MedlinePlus list is *lightheadedness*. The expanded query performed 20% more than the baseline and the monolingual query (80% for both) in terms of P@10. For the Czech language, 12 queries improved, 24 queries remained unchanged and 30 queries destroyed. And for French, 11 queries improved, 27 queries did not change and 28 queries destroyed. However, regarding the degraded queries, we can not say anything about that since the assessment information for this run is not completed. We get about 60% of coverage for the top 10 retrieved documents for all languages. The rest of unjudged documents are treated as irrelevant, and this might not be the case.

4.11.2 Assessment procedure

As we showed in the previous section, the unjudged documents in our systems make it impossible to decide about the results if they are better or not. In order to solve this issue, we setup up the assessment tool that is provided by Koopman and Zuccon [2014] in our side. We will ask assessors who have experience in the medical domain to help us to fully assess our systems. After that, we will have a clear idea about which system works better, and which queries degraded or improved. Such information will definitely help us to improve our system and figure out what will work and will not.

5 Our publications

During our research, we published the following papers:

- Shadi Saleh and Pavel Pecina. CUNI at the ShARe/CLEF eHealth Evaluation Lab 2014. In Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Sheffield, UK,2014.
- Shadi Saleh, Feraena Bibyna, Pavel Pecina: CUNI at the CLEF eHealth 2015 Task 2. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, CEUR-WS, Toulouse,France, 2015.
- Shadi Saleh and Pavel Pecina. Adapting SMT Query Translation Reranker to New Languages in Cross-Lingual Information Retrieval. In Medical Information Retrieval (MedIR) Workshop, Association for Computational Linguistics, Stroudsburg, USA, 2016
- Shadi Saleh and Pavel Pecina. Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016, Berlin, Germany, 2016. Springer
- Shadi Saleh, Pavel Pecina: Task3 Patient-Centred Information Retrieval: Team CUNI. Accepted for publication in: CLEF 2016 Working Notes, CEUR-WS, Evora, Portugal, 2016.

6 Future work

First we will finish the documents translation for all languages and complete the assessment process for our systems. Then, we will explore more the use of word-embeddings in the CLIR task, for example: build the embeddings for both of documents and queries and perform the retrieval on the embeddings level. Also, we want to develop our Cross-lingual embeddings, this will help us to get rid of the SMT systems.

The work will include participation in the related tasks and submitting the experiments' results to the relevant conferences.

References

Aronson, A. R., Effective mapping of biomedical text to the umls metathesaurus: the metamap program., *Proc AMIA Symp*, pp. 17–21, 2001.

Aronson, A. R. and Lang, F.-M., An overview of MetaMap: historical perspective and recent advances, *Journal of the American Medical Informatics Association*, *17*, 229–236, 2010.

Bodenreider, O., The unified medical language system (umls): Integrating biomedical terminology, 2004.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L., Findings of the 2013 Workshop on Statistical Machine Translation, in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pp. 1–44, Association for Computational Linguistics, Sofia, Bulgaria, 2013.

Bosca, A., Casu, M., Dragoni, M., and Di Francescomarino, C., Using semantic and domain-based information in CLIR systems, in *The Semantic Web: Trends and Challenges*, pp. 240–254, Springer, 2014.

Choi, S. and Choi, J., Exploring effective information retrieval technique for the medical web documents: Snumedinfo at clefehealth2014 task 3., in *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, vol. 1180, pp. 167–175, Sheffield, UK, 2014.

Crammer, K. and Singer, Y., Ultraconservative online algorithms for multiclass problems, *The Journal of Machine Learning Research*, *3*, 951–991, 2003.

Darwish, K. and Oard, D. W., Probabilistic structured query methods, in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 338–344, ACM, New York, USA, 2003.

Dušek, O., Hajič, J., Hlaváčová, J., Novák, M., Pecina, P., Rosa, R., and 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.

Ferro, N. and Peters, C., Clef 2009 ad hoc track overview: Tel and persian tasks, in *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pp. 13–35, Springer, 2010.

Fujii, A. and Ishikawa, T., Applying machine translation to two-stage cross-language information retrieval, in *Envisioning Machine Translation in the Information Future*, vol. 1934, pp. 13–24, Springer, Berlin, Germany, 2000.

Fujii, A., Utiyama, M., Yamamoto, M., and Utsuro, T., Evaluating effects of machine translation accuracy on cross-lingual patent retrieval, in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pp. 674–675, ACM, New York, USA, 2009.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., and Huang, C., Improving query translation for cross-language information retrieval using statistical models, in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 96–104, ACM, 2001.

Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., and Mueller, H., ShARe/CLEF eHealth evaluation lab 2014, Task 3: User-centred health information retrieval, in *Proceedings of CLEF 2014*, 2014.

Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., and Zuccon, G., Overview of the CLEF eHealth evaluation lab 2015, in *The 6th Conference and Labs of the Evaluation Forum*, Springer, Berlin, Germany, 2015.

Grefenstette, G. and Nioche, J., Estimation of english and non-english language use on the www, in *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, pp. 237–246, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France, France, 2000.

Hieber, F. and Riezler, S., Bag-of-words forced decoding for cross-lingual information retrieval, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, Denver, Colorado, 2015.

Hull, D., Using statistical testing in the evaluation of retrieval experiments, in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329–338, Pittsburgh, USA, 1993.

Hull, D. A. and Grefenstette, G., Querying across languages: A dictionary-based approach to multilingual information retrieval, in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–57, ACM, New York, USA, 1996.

Humphreys, B. L., Lindberg, D. A. B., Schoolman, H. M., and Barnett, G. O., The unified medical language system, *Journal of the American Medical Informatics Association*, *5*, 1–11, 1998.

Järvelin, K. and Kekäläinen, J., Cumulated gain-based evaluation of IR techniques, *ACM Trans. Inf. Syst.*, *20*, 422–446, 2002.

Kim, S., Wilbur, W. J., and Lu, Z., Bridging the gap: a semantic similarity measure between queries and documents, *arXiv preprint arXiv:1608.01972*, 2016.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., and et al., Moses: Open source toolkit for statistical machine translation, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic, 2007.

Kohlschütter, C., Fankhauser, P., and Nejdl, W., Boilerplate detection using shallow text features, in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 441–450, ACM, New York, NY, USA, 2010.

Koopman, B. and Zuccon, G., Relevation!: An open source system for information retrieval relevance assessment, in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 1243–1244, ACM, 2014.

Kuzi, S., Shtok, A., and Kurland, O., Query expansion using word embeddings, in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1929–1932, ACM, New York, NY, USA, 2016.

Levow, G.-A., Oard, D. W., and Resnik, P., Dictionary-based techniques for cross-language information retrieval, *Information Processing and Management*, *41*, 523–547, 2005.

Liu, X. and Nie, J., Bridging layperson's queries with medical concepts – GRIUM @CLEF2015 eHealth Task 2, in *Working Notes of CLEF 2015 Conference and Labs of the Evaluation forum*, vol. 1391, Toulouse, France, 2015.

Macdonald, C., Plachouras, V., He, B., Lioma, C., and Ounis, I., University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming, in *Accessing Multilingual Information Repositories*, vol. 4022 of *Lecture Notes in Computer Science*, pp. 898–907, Springer, Berlin, Germany, 2006.

Magdy, W. and Jones, G., Should MT systems be used as black boxes in CLIR?, in *Advances in Information Retrieval*, edited by P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, and V. Mudoch, vol. 6611, pp. 683–686, Springer, Berlin, Germany, 2011.

McCarley, J. S., Should we translate the documents or the queries in cross-language information retrieval?, in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 208–214, Association for Computational Linguistics, College Park, Maryland, 1999.

McCullagh, P. and Nelder, J., *Generalized Linear Models*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, 1989.

Mikolov, T., Chen, K., Corrado, G., and Dean, J., Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*, 2013.

Nikoulina, V., Kovachev, B., Lagos, N., and 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.

Oard, D., A comparative study of query and document translation for cross-language information retrieval, in *Machine Translation and the Information Soup*, vol. 1529, pp. 472–483, Springer, Berlin, Germany, 1998.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C., Terrier: A high performance and scalable information retrieval platform, in *Proceedings of Workshop on Open Source Information Retrieval*, Seattle, WA, USA, 2006.

Papineni, K., Roukos, S., Ward, T., and 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.

Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlaváčová, J., Jones, G. J., and et al., Adaptation of machine translation for multilingual information retrieval in the medical domain, *Artificial Intelligence in Medicine*, *61*, 165–185, 2014.

Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K., Dictionary-based cross-language information retrieval: Problems, methods, and research findings, *Information retrieval*, *4*, 209–230, 2001.

Pomikalek, J., *Removing Boilerplate and Duplicate Content from Web Corpora*, Ph.D. thesis, Ph.D. thesis, Masaryk University, 2001.

Řehůřek, R. and Sojka, P., Software Framework for Topic Modelling with Large Corpora, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, ELRA, Valletta, Malta, 2010.

Roy, D., Ganguly, D., Mitra, M., and Jones, G. J., Representing documents and queries as sets of word embedded vectors for information retrieval, *Proceedings of the Neural Information Retrieval (Neu-IR) Workshop. A SIGIR 2016 workshop*, 2016.

Schamoni, S. and Riezler, S., Combining orthogonal information in large-scale cross-language information retrieval, in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pp. 943–946, ACM, New York, USA, 2015.

Schuyler, P. L., Hole, W. T., Tuttle, M. S., and Sherertz, D. D., The UMLS Metathesaurus: representing different views of biomedical concepts., *Bulletin of the Medical Library Association*, *81*, 217, 1993.

Smucker, M. D. and Allan, J., An investigation of Dirichlet prior smoothing's performance advantage, Tech. rep., University of Massachusetts, 2005.

Sokolov, A., Hieber, F., and Riezler, S., Learning to translate queries for CLIR, in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1179–1182, Gold Coast, Australia, 2014.

Spink, A., Wolfram, D., Jansen, M. B., and Saracevic, T., Searching the web: The public and their queries, *Journal of the American society for information science and technology*, *52*, 226–234, 2001.

Strohman, T., Metzler, D., Turtle, H., and Croft, W. B., Indri: A language model-based search engine for complex queries, in *Proceedings of the International Conference on Intelligent Analysis*, vol. 2, pp. 2–6, McLean, VA, 2005.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., and et al., Overview of the ShARe/CLEF eHealth evaluation lab 2013, in *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pp. 212–231, Springer, Berlin, Germany, 2013.

Talvensaari, T., Effects of aligned corpus quality and size in corpus-based CLIR, in *Advances in Information Retrieval*, pp. 114–125, Springer, 2008.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H., Accelerated DP based search for statistical translation, in *In European Conference on Speech Communication and Technology*, pp. 2667–2670, Rhodes, Greece, 1997.

Ture, F. and Boschee, E., Learning to translate: A query-specific combination approach for cross-lingual information retrieval, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 589–599, Qatar, 2014.

Ture, F., Lin, J., and 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.

Xu, J., Weischedel, R., and Nguyen, C., Evaluating a probabilistic model for cross-lingual information retrieval, in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pp. 105–110, ACM, New York, NY, USA, 2001.

Zamani, H. and Croft, W. B., Embedding-based query language models, in *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pp. 147–156, ACM, New York, NY, USA, 2016.

Zhai, C. and Lafferty, J., A study of smoothing methods for language models applied to information retrieval, *ACM Transactions on Information Systems (TOIS)*, *22*, 179–214, 2004.