FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Shadi Saleh

Cross-lingual Information Retrieval in the Medical Domain

Institute of Formal and Applied Linguistics

Supervisor of the doctoral thesis:	doc. RNDr. Pavel Pecina, PhD.
Study programme:	Mathematical Linguistics

Prague 2020

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date .... signature of the author

Acknowledgment

The journey during my PhD. has been full of excitement and challenges. However, I would not have been able to complete this research without the extreme support that I got from those whom I would like to thank.

To my supervisor doc. RNDr. Pavel Pecina, PhD. for his ultimate support not only in my research but also in overcoming other obstacles that I have faced during my work, and for his patience and the time he dedicated for me.

To Prof.Gareth J.F. Jones and Prof.Allan Hanbury for their valuable insights on details of the work, and to my colleagues who helped me and gave me great feedback such as Martin Popel, Tom Kocmi and Anna Vernerova.

To Mgr. Martin Mareš, Ph.D. who told me about UFAL during a five minutes lunch break in Italy during the 2012 International Olympiad in Informatics, that was the beginning of my journey to here.

To Joni, Ali, Ammar and Suleiman, my friends who always claimed to enjoy listening to me talking about my research topic, who were able to help me escape from reality for some time. To my girlfriend Marika, for the love, support and care she surrounded me with, for her tolerance for me being busy all time with my thesis.

To my mom and dad for everything they have done for me, to my brother Firas, who introduced me to programming when I was 10 years old, that moment when all my perspectives about my dreams changed, to my brothers and sisters for their extreme love and support.

Finally, I want to dedicate this work to my brother Wessam who's light was dimmed while I was writing the last pages of this work due a terrible accident, I truly believe that he will wake up soon from his coma.

Tomorrow when my country sings
With love flowing from me,
I erase the blackness with my face
And become a nation for every
nation
So no darkness remains in our land
And no evil remains
Thus, say, I am free
And say, you are free.

— Adunis Esber, Selected Poems

Title: Cross-lingual Information Retrieval in the Medical Domain

Author: Shadi Saleh

Institute: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Pavel Pecina, PhD., Institute of Formal and Applied Linguistics

Abstract: In recent years, there has been an exponential growth of the digital content available on the Internet, which has correlated with the increasing number of non-English Internet users due to the spread of the Internet across the globe. This raises the importance of unlocking resources for those who want to look up information not limited to the languages they understand. For example, those who want to use the Internet to find medical content related to their health conditions (self-diagnosis) but they do not have access to resources in their language. Cross-Lingual Information Retrieval (CLIR) breaks the language barriers by allowing search for documents written in a language different from the query language.

This thesis tackles the task of CLIR in the medical domain and investigates the two main approaches: query translation (QT) where queries are machine translated to the language of documents and document translation (DT) where documents are translated to the language of queries. We proceed with our research by employing Statistical Machine Translation (SMT) systems that are tuned for the QT approach and the DT approach in the medical domain for seven European languages (Czech, German, French, Spanish, Hungarian, Polish and Swedish) and empirically show that DT does not outperform QT (the contrary to what had been assumed since the late 1990's). We develop a machine-learning-based system to rerank the translation hypotheses provided by an SMT system towards better CLIR performance. The system is first designed for Czech, French and German CLIR systems and then is adapted to Spanish, Hungarian, Swedish and Polish. Our findings suggest that the best translation produced by SMT is not necessarily the best translation to construct a query in CLIR. Our reranker system produces translations that are optimized towards CLIR, and significantly outperforms the baseline QT system without reranking. To remedy the vagueness of translated queries in CLIR, we present a novel approach that reformulates base queries by adding useful terms to them. The terms are scored for usefulness using a linear regression model. Our approach improves both the performance of CLIR systems in all languages and of the monolingual IR (English reference queries). To compare the performance of SMT versus NMT (predicting a translation using deep neural networks) in the context of CLIR, we train a task-oriented NMT model to translate medical queries. The presented NMT-based QT model significantly outperforms the SMT-based QT one in all languages. During the progress of our research, we developed an extended dataset for CLIR in the medical domain, which is based on existing datasets from the IR tasks of the CLEF eHealth Labs Series 2013–2015, and we make the dataset publicly available via the Lindat/CLARIN repository.

Keywords: Cross-lingual Information Retrieval, Machine Translation, Consumer Health Search

Contents

1. Introduction

The digital medical content available online has snowballed in recent years. This growth has the potential to improve experience with web medical Information Retrieval (IR) systems, which are more and more used for health consultations.

Fox [2011] reported that about 80% of Internet searchers in the U.S. looked for health information online, and this number was expected to grow. The significant increase of non-English digital content on the Internet had been followed by an increase in looking for this information by internet searchers. 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 had grown by 800%, German by 1500%, and Spanish by 1800% in the same period.

Naturally, some information that a searcher is looking for might be available only in a language that they do not understand, which makes such information not accessible to those searchers.

Cross-Lingual Information Retrieval (CLIR) comes to tackle this issue and to help internet searchers break language barriers and access valuable information that is not available in their language.

Medical IR is a task that helps find medical content. According to Hersh [2008], demand for such a system has increased significantly for mainly two reasons. Firstly, searchers for health-related topics, either consumers (laypeople with no experience in the medical domain) or clinicians and medical experts, are relying more on the Internet to find medical advice. Secondly, medical institutions adopt electronic medical records in their systems, which makes more medical data available in a digital and a searchable form.

1.1 Goals

The goals of this thesis can be summarized as follows:

- Studying the challenges of CLIR in the medical domain.
- Developing task-oriented machine translation systems (statistical and neural) to be employed in the CLIR task.
- Comparing the main CLIR approaches, namely: query translation and document translation.
- Improving medical information representation in search queries.

This work is conducted in the context of the Khresmoi project, and as a contribution to the CLEF eHealth IR shared tasks, as we will show later.

1.2 Contributions

Taking into consideration the goals of this research, and the findings and observations that we encountered during our journey, we can summarize our contributions to the task of CLIR as follows:

- Defining the task of CLIR, its challenges and approaches, and the related work that has been conducted.
- Designing a CLIR baseline system that is based on the Query Translation (QT) approach employing Statistical Machine Translation (SMT).
- Developing a method for reranking of the translation hypotheses that are produced by SMT towards better CLIR performance.
- Presenting a novel approach towards building a task-oriented neural machine translation system for query translation in CLIR.
- Studying the effect of morphological processing on monolingual and cross-lingual IR in the medical domain.
- Conducting document translation (DT) experiments using SMT and NMT models and showing empirically that QT outperforms DT (contrary to what has been assumed for more than 20 years) in the context of CLIR in the medical domain.
- Improving the representation of the information need in search queries through a query expansion method based on a machine learning model.
- Contributing to the research community by releasing a test collection where queries are available in eight languages, and a thorough relevance assessment.
- Publishing those findings as nine long papers and two short ones in relevant conferences.

Finally, I want to emphasize that I use *we* in this work to refer to me and you (the reader), and it does not imply any collaborative work.

1.3 Organization of the Thesis

The organization of this thesis is as follows. Chapter 2 introduces the task of information retrieval in the medical domain, describes retrieval models, and popular evaluation methods of IR systems. We move on in Chapter 3 to give an overview of the CLIR task and the related work, tasks, and evaluation campaigns of CLIR, and the released CLIR test collection.

In Chapter 4, we present the test collections that we employed in this work and our contribution to extending these collections and unifying them into a richer one.

Chapter 5 presents an overview of the task of Machine Translation (MT) and how it is employed in CLIR. We study both SMT and Neural Machine Translation (NMT), then we present the MT data that we use in this research for

MT training, the evaluation of MT, we present our training approach and model architecture of MT models (SMT and NMT) for the purpose of the translation part in the CLIR task, and then we present the evaluation of our MT models.

Chapter 6 is dedicated to our CLIR methods, including methods that are based on query-translation using SMT, NMT and public MT. Then we present our translation hypotheses reranking model towards CLIR. The chapter also includes a comparative study of document translation and query translation using SMT and NMT, and then we present our term selection approach for query expansion based on a machine-learning model. The last chapter presents our conclusions and future work.

2. Information Retrieval

The IR task has been discussed since the 1950s. It has always been to understand and improve the way humans seek information.

There are multiple definitions of the IR task; however, according to Manning et al. [2008], it is agreed in academia that the following one defines the task:

> *"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."*

Cleverdon [1960] proposed that an IR system should consider the following perspectives:

- How much the system can distinguish relevant documents.
- The system should be able to distinguish irrelevant documents.
- How much time it takes the system to retrieve the results after being asked a question.
- How the system is going to present the results.
- How much the system is easy to be used by searchers in order to get the results they want.

Information need is a very important concept in the field of IR. According to Wissbrock [2004], information need is defined as the missing information that a searcher is trying to find, and from a searcher's point of view, this information need can be reached by writing one or more terms. These terms are known as *query*. When a document contains that missing information, we call this document a *relevant* document, and when it does not, then it is called *irrelevant*.

The fact that the text is stored in unstructured documents[1] makes the IR task a challenging one, especially when comparing it with information extracted from structured data stored in relational database systems. Because in relational database systems, data is well defined and inserted into tables that have known structures (called schemas), and relations between tables explain how data is connected to each other. Structured Query Language (SQL) helps to retrieve data from a database considering a given criterion through a DataBase Management System (DBMS). Usually, the relevance of the retrieved data is very high, and the criterion of describing this data is clear (not ambiguous). For example, when asking a DBMS to show the author and the publisher of a specific book assuming these fields are stored in a table, the system will effectively retrieve such information, because these fields are defined in the table's schema in a way that enables accessing a specific field (author and publisher) for a specific row

[1] Documents in IR most often do not comply with a predefined schema definition, such as data type and date format. On the contrary, they are usually written in a free text format or *unstructured text.*

(a book in this example) in that table. However, in IR, the situation is totally different. The goal is to retrieve information from unstructured text. Looking up information in unstructured text is more difficult because it requires more advanced text analyzing and parsing techniques, and the same information that the searcher is looking for might be written using different words, where most IR systems follow approaches that are based on term-matching methods.

2.1 Research Terminology

There are multiple terms that are widely used in IR referring to specific concepts, and have different use or meaning in other fields of Natural Language Processing (NLP) researches. To avoid any confusion, we dedicate this section to explain the meaning of the terms used in our work:

- **Information need** is information that a searcher tries to find using a retrieval system.

- **Query** can be either a sequence of terms that is fed by a searcher into an IR system to find a specific information, or written in a more complicated way such as structured queries wherein different terms in a query can have different weights [Strohman et al., 2005].

- **Term** is a token that is processed by an IR system. A term can be a word in its original form, lemmatized version, or its stemmed version.

- **Topic** describes what information is needed behind a given query, assuming that a query cannot fully describe this information.

- **Relevance** determines how much a retrieved document satisfies an information need for a searcher. In searcher-centric relevance, documents should be relevant to the query topic, not to the query itself, because sometimes searchers fail to write a clear or complete query about their information need. In this research, we will focus on searcher-centric relevance not system-centric relevance, which is the case when the goal is to satisfy the searcher's query, not their information need [Smith and Salvendy, 2007].

- **Relevance Assessment** is a manual process wherein human assessors check retrieved documents by an IR system for a given query and the information need behind that query and assign the relevance degree to those documents.

2.2 Consumer Health Search

Health search is the process when searchers use the Internet to look up information that is related to health conditions, symptoms, treatments, or diseases. The term *consumer* had appeared in the literature since the beginning of 2000, referring to the health searchers who do not have a strong background in medicine; hence, they share mutual behavior: the lack of medical terminology when posing a search query. When searchers have a medical background, such users are referred to as *clinicians* [Zeng et al., 2002].

Patrick et al. [1995] defined Consumer Health Search (CHS) as the process when consumers find information online that helps them to understand health topics (either related to their health conditions or to one of their family members), and make actions or take decisions based on what they find.

In July 2017, dotHealth[2] reported their findings in a national survey of 1,509 internet searchers in Canada. They found that 57% of the studied population would search online first when they encountered health-related questions, while 32% would visit a doctor before searching online. Fox [2011] spotted an increase of the percentages of Internet users who used the Internet at least once to look up health-related information in the United States, where the percentages increased from 25% in 2000, to 80% in 2010.

Zeng et al. [2002] studied the characteristics of consumer health search. They found a significant mismatch between the terminology that consumers used to write their queries and the correct terminology in the documents that were relevant to their needs.

Keselman et al. [2008] showed that the lack of medical domain knowledge led consumers to read and to rely on information that was taken from irrelevant resources, which could have significant impacts on their health.

However, when consumer health search was applied to a credible and variable medical content, it helped consumers to improve their clinical interventions and eventually boosted the clinical health search outcome [Gibbons et al., 2009].

Self-diagnosis using health search has been a controversial topic. Giustini [1999] suggested that a governmental organization such as the federal agency of the United States Department of Health and Human Services should regulate and monitor the digital medical content that is available to the public, because some medical resource providers can be considered practicing medicine without being monitored. However, the main advantage of CSH is that searchers can improve their knowledge in the medical domain, and at the same time, gain experience in recognizing irrelevant and untrustworthy resources by exploring multiple documents with different relevance degree.

The quality of the medical content (medical accuracy) is not the only important factor in CHS. Readability measurement, on the other hand, determines how easily a reader is able to read, to understand, and to make the right conclusion after reading a medical text. Kindig et al. [2004] estimated that there are about 90 million adults in the United States who had difficulty reading and understanding health-related text. This issue could cause a barrier to obtain online patient care for those people. However, in this work, we do not focus on the readability of medical content when searching for health-related topics; we keep our focus on the relevance of the retrieved medical information.

[2] `https://www.dothealth.ca/`

2.3 Retrieval Models

We present in the following sections an overview of multiple IR models. An IR model is a matching function that takes as input a set of documents (D) and a user's query (q), and retrieves a list of documents (D'), where $D' \in D$. These retrieved documents are often scored by their matching degree to q.

2.3.1 Vector Space Model

In the vector space model, both queries and documents are represented as vectors, such as $\vec{V}(d_i) = (w_{i,1}, w_{i,2}, .., w_{i.,m})$, where $w_{i,j}$ is the weight of the term j in the document i. TF-IDF is the most common model that is based on vector space model [Salton et al., 1975], in which the weight of a term t in a document d is computed as:

$$w_{ij}\left(t,d\right) = tf_{t,d} \times IDF_t \tag{2.1}$$

where $tf_{t,d}$ is the term frequency of a term t in a document d, and IDF_t is the Inverse Document Frequency (IDF) of a term t in the entire collection. The value of IDF is calculated as shown in Equation 2.2, where N is the number of the total documents in the collection, and df_t is the document frequency, i.e. the number of documents in the collection that contain the term t. IDF can be considered as a measurement of the informativeness of a term t in the collection, where rare terms tend to have high IDF values, and frequent terms have low values, for example stop-words which are words that appear frequently in many documents in the collection but they are not very informative from the perspective of an IR system.

$$IDF_t = log\frac{N}{df_t} \tag{2.2}$$

After representing each document d in the collection as a vector $V\vec{(d)}$ of the weights of its terms, and a query q as a vector $V\vec{(q)}$, the similarity score is quantified using the cosine similarity as given in Equation 2.3, where the numerator is the dot product between $V\vec{(d)}$ and $V\vec{(q)}$, and the denominator is the product of their Euclidean lengths [Manning et al., 2008].

$$RSV\left(q,d\right) = cosine(V\vec{(q)}, V\vec{(d)}) = \frac{V\vec{(q)} \cdot V\vec{(d)}}{\left|V\vec{(q)}\right|\left|V\vec{(d)}\right|} \tag{2.3}$$

Retrieval Status Value (RSV) in IR is a score that is given by retrieval model to define the relevance degree of a given document to an input query [Imafouo and Tannier, 2005].

2.3.2 Probabilistic Retrieval Model

Probabilistic retrieval models rank the documents by their likelihood ratios, which is based on the probabilistic retrieval framework [Robertson and Sparck Jones, 1988]. $R_{q,d} = 1$ if a document d is relevant to a query q, and $R_{q,d} = 0$, if d is irrelevant to q. Probabilistic models estimate the probability of document relevance to a query q. This estimation is written as $P(R = 1|d, q)$.

Okapi BM25 (Best Matching) is a well-known IR model that is based on the probabilistic framework. Documents in this model are ranked for a given query, as shown in Equation 2.4. k_1 and k_3 normalize term frequency and document length, respectively. For this reason, these parameters need to be tuned based on the studied document collection. While tf_d is the normalised term frequency in document d, as shown in Equation 2.5, where dl and avg_{dl} are document length and the average of document length in the collection respectively, and b is a free parameter.

$$RSV(q,d) = \sum_{t \in_d \bigcap_q} \frac{(k_1+1)tf_d}{K+tf_d} * \frac{(k_3+1)*tf_q}{k_3+tf_q} * IDF(t) \tag{2.4}$$

$$tf_d = \frac{tf}{(1+b)+b*\frac{dl}{avg_{dl}}} \tag{2.5}$$

2.3.3 Dirichlet Prior Weighting Model

The main assumption of language modeling based IR models is that when searchers pose a query to find relevant documents to their information need, they use terms that are more likely to appear in those documents. This means that a document d is a good candidate to be relevant to a query q when d is more likely to generate query q (when a searcher wants to find a document d, they will write a query q that represents d from their point of view). To model this, each document d in the collection D has an estimated language model θ_d, and each term t in a document d is assigned a probability as in:

$$P\left(t|\theta_d\right) = \frac{tf_{t,d}}{dl} \tag{2.6}$$

Different smoothing methods are applied to term probabilities. Smoothing with the Dirichlet prior is shown to be an effective method in IR models [Zhai and Lafferty, 2004]. The scoring function in the Dirichlet model is given by Equation 2.7.

$$RSV\left(q,d\right) = log\left(p(q|d)\right) = \sum_{i:c(t_i \in q;d)>0} log\frac{p_s(t_i|d)}{\alpha_d p(t_i|C)} + nlog\alpha_d + \sum_{i=1}^{n} logp(t_i|C) \tag{2.7}$$

where $p_s(t_i|d)$ is the probability that a document d implies a term t_i, t_i exists in the collection C, and α_d is a document-dependent constant, n is the length of the query q, and $p(t_i|C)$ is the probability of a query term t_i in the entire collection C, which is fixed for all documents; thus, the last component in Equation 2.7 is ignored in the ranking function.

$$p_\mu(t|d) = \frac{c(t,d)+\mu p(t|C)}{dl+\mu} \tag{2.8}$$

Dirichlet smoothing is applied to the term probability t_i by the maximum likelihood estimation with the term probability in the entire collection C. It uses a different amount of smoothing based on the length of the document. For longer documents, the smoothing will be less.

```
<query>
    <id>103001</id>
    <title>headaches relieved by blood donation</title>
</query>

<query>
    <id>103002</id>
    <title>high iron headache</title>
</query>

<query>
    <id>103003</id>
    <title>blood donation headache reduction</title>
</query>

<query>
    <id>103004</id>
    <title>headaches caused by too much blood or "high blood
   pressure"</title>
</query>

<query>
    <id>103005</id>
    <title>headache that only goes away with blood loss</title>
</query>
```

Figure 2.1: Samples of query variations of the same topic. These topics were released during the CLEF eHealth 2016 IR task [Kelly et al., 2016]

2.4 System Evaluation

The purpose of system evaluation, in general, is to tell how well a system performs, and which system is better than others. The determination of a better system performance lies in the context of that system application.

When searchers want to find information in a set of documents, they represent this information as a query q. Multiple searchers tend to formulate their information need in different queries. These queries that refer to the same information need are called query variations. Table 2.1 shows samples of multiple query variations for the same information need. These variations were created within the CLEF eHealth IR track 2016 [Kelly et al., 2016].

In searcher-centered IR,[3] the primary purpose of the evaluation is to measure how well results from an IR system satisfy searcher expectation [O'Brien et al., 2016].

A document d is relevant (*rel*) to a query q only if it satisfies the searcher who posted the query. If d does not contain answers to what the searcher is looking for then d is irrelevant (*irrel*).

[3] This is the case in our work where searchers are patients (patient-centered IR).

Evaluation of unranked retrieval:

Unranked retrieval is the case when an IR system returns a set of documents for a given query without a score of relevance, which means all of these documents are treated equally.

The most common two metrics to evaluate performance of an unranked retrieval system are precision and recall.

Precision: Precision is the ratio of the number of relevant retrieved documents to the total number of retrieved documents as in:

$$Precision = P = \frac{\#(relevant\ retrieved\ documents)}{\#(total\ retrieved\ documents)} \tag{2.9}$$

Recall: Recall is the ratio of the number of retrieved relevant documents to the total number of relevant documents in the collection:

$$Recall = R = \frac{\#(relevant\ retrieved\ documents)}{\#(total\ relevant\ documents)} \tag{2.10}$$

The desired value of precision and recall can be controlled by the context of the application of the IR system. For example, while web searchers are usually not interested in all relevant documents, they focus on a few results; thus, it will be more effective for them to use a high precision IR system. On the other hand, in a different search application, searchers might be interested in looking at more results. For that reason, they can tolerate a low precision system and prefer a high recall one. Figure 2.2 shows an example of the relation between precision and recall, which is called *precision-recall* curve.

F-measure: F-measure balances precision and recall by using the harmonic mean between the two metrics.

$$\text{F-measure} = F = \frac{2PR}{P+R} \tag{2.11}$$

F-measure is calculated, as shown in Equation 2.11, where P is precision, and R is recall. This version of F-measure is also called F_1, where each P and R are given an equal weight.

Evaluation of ranked retrieval:

IR systems usually return a list of documents ranked descendingly from the most relevant document. Practically, searchers will not read all the retrieved documents; they usually examine the highly ranked documents only and do not look further in the retrieved list.

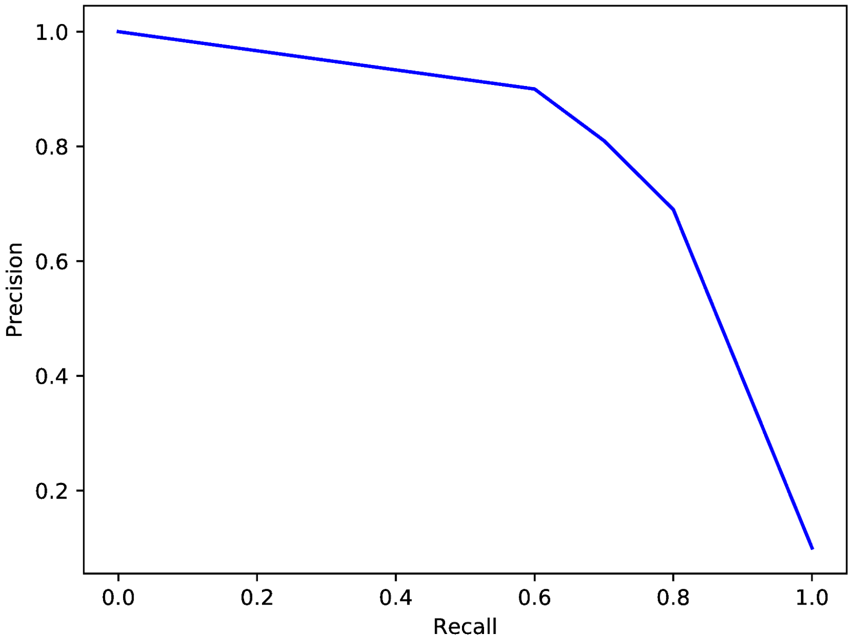


Figure 2.2: Precision-recall curve

Precision at K-documents: $P@K$ is the proportion of the top-K retrieved documents that are relevant to the posed query as shown in Equation 2.12.

$$P@K = \frac{1}{K}\sum_{i=1}^{K} \mathbb{1}(rel(d_i)) \tag{2.12}$$

Where $\mathbb{1}(rel(d_i)$ is an indication function that returns 1 if the document d at position i is relevant to the query, or 0 otherwise. To compute the overall $P@K$ for a set of queries, the average $P@K$ for all queries is taken.

Usually, Precision-at-10 is considered to be a reasonable setup for $P@K$ in the context of web search, because popular web search systems nowadays show by default the top 10 ranked pages (documents) in their first page [Turpin and Scholer, 2006]. The main disadvantage of using P@K is that it does not consider the positions of the relevant documents (within the top K results) with respect to the irrelevant ones.

Mean Average Precision: The precision, as we showed earlier in this section, is used to evaluate an unranked list of retrieved documents. This ignores the position of relevant documents, whether they appear at the top of the list or at its bottom. However, in ranked retrieval, relevant documents are desired to be at the top of the list, so we need an evaluation metric that considers the order of these documents. Average Precision (AveP) computes the average precision of a ranked list at every position of a relevant document. For a set of queries (Q), Mean Average Precision (MAP) is the average of AveP for each query j as shown in Equation 2.13, where m_j is the number of relevant documents for query j, and R_{jk} is a list of the first k ranked documents for query j.

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{2.13}$$

Binary preference-based measure: *bpref* considers how often relevant documents are ranked above the irrelevant ones. This feature makes *bpref* more stable (in terms of system comparison) when the judgment information in the test collection is incomplete.[4] Thus, it is recommended to be used in such a case [Craswell, 2009].

$$bpref = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{d_j} (1 - \frac{|\textit{irrel} \text{ ranked higher than } \textit{rel}|}{min(m_j, n_j)}) \tag{2.14}$$

bpref is calculated for a set of queries (Q) as shown in Equation 2.14, where m_j is total the number of relevant documents for query j as defined in the assessment information, d_j is a relevant document, n_j is the total number of irrelevant documents, and |*irrel* ranked higher than *rel*| refers to the number of times an irrelevant document (*irrel*) is ranked higher than a relevant one (*rel*). The fraction is normalized by $min(m_j, n_j)$ [Sakai, 2007].

[4] Usually, it is costly to manually annotate every document in the retrieval result (thousands of documents are often retrieved for each query) if it is relevant or not. For that reason, a subset of the highly ranked documents retrieved by one or more IR models is automatically selected for annotation. This often causes unjudged documents to appear in the retrieval results, especially when a new IR model is applied.

3. Related Work

In this Chapter, we will present related work to the CLIR task. This also includes campaigns, tracks, and CLIR data collections that have been released by the research community.

3.1 Cross-lingual Information Retrieval

CLIR enables users to search for information by allowing queries in a language that is different from the collection language. This helps break the language barrier between searchers and a vast amount of data that is represented in a different language. The task has gotten the attention of the IR research community since the late 1990s, and the growth of the Internet was a compelling evidence of the need of CLIR systems because digital content across the globe had begun to increase significantly.

A CLIR system usually includes two steps; the first step is the translation step, which includes translating either the queries into the language of the document collection or translating the document collection into the query language. After the translation is done, the task is then reduced into a monolingual IR task. Different approaches and studies investigated two main questions in the CLIR task:

- What is better to translate, queries, or document collection? Or translating both into a common representation?
- How can translation be done? Is the translation task in CLIR similar to the normal machine translation task that aims at generating human-readable translations?

In this section, we will present related work that has been done to answer these questions.

3.1.1 Document Translation Approach

The two main approaches in CLIR are Document Translation (DT) and QT. In CLIR, documents and queries are written in two different languages. To conduct a term-matching based retrieval, it is required to have both documents and queries represented in one language; therefore, either queries should be translated into the collection language (QT), or the document collection should be translated into the query language (DT).

The question of whether to translate document collection into the query language or queries into the document language has been the main focus of the CLIR community for a while.

Oard [1998] investigated the performance of DT, QT, and a hybrid system of both. In their study, they used the Logos translation system[1] to translate between German and English. As for the document collection and the test queries,

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

they used the *TREC-6 CLIR SDA/NZZ* test set, which contained about $251,840$ German documents and 22 English queries.

They compared two systems: the first system translated English queries into German (the collection language), and the second system translated the documents from German into English (the query language). Their study showed that *DT outperformed QT*. They explained that because documents are usually longer than queries, which leads to more contextual and linguistic information that helps reduce translation ambiguity (when one term in a source language has more than one translation candidate in a target language). Their attempts to reduce translation ambiguity showed better machine translation performance, which improved the retrieval performance eventually.

McCarley [1999] presented a hybrid system of DT and QT. To achieve this, they built two translation systems (English to French and French to English) that were similar in performance as much as possible by training them using the same training data. Their experiments showed that a hybrid system that averaged the retrieved document scores from DT and QT systems outperformed both of them.

Fujii and Ishikawa [2000] employed a two-step method, in which QT was used to retrieve a limited number of documents, and then these documents were translated into the query language and reranked by their final scores.

Yarmohammadi et al. [2019] followed the document translation approach. They translated the documents into English, and then performed the retrieval in a monolingual setup. To achieve that, they trained both NMT and SMT systems for the translation step and investigated the CLIR performance for both systems. For MT training, they used the MT data that was provided in the OpenCLIR (Open Cross Language Information Retrieval) evaluation within the MATERIAL (Machine Translation for English Retrieval of Information in Any Language) project.[2] The OpenCLIR dataset contained documents in Somali, Swahili and Tagalog, and English queries. They showed that DT using SMT outperformed DT using NMT. Authors stated in their work that DT outperforms QT, because documents have more context; thus DT can produce more accurate translations, adopting the findings of Croft et al. [1991].

3.1.2 Dictionary-based Query Translation

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

- Untranslatable words due to Out-Of-Vocabulary (OOV) problem (words did not appear in the training data, so could not be translated).

- Processing inflected words.

- Identification of phrases and collocations to be translated correctly.

- Lexical ambiguity in source and target languages.

[2] https://www.nist.gov/itl/iad/mig/openclir-evaluation

Pirkola [1998] showed that query translation using Machine-Readable Dictionary (MRD) could be effective when the used MRD was a domain-specific one (medical MRD in their case); this helped to reduce OOV.

As a method of query expansion (query expansion is the process of adding new informative terms to queries, details are presented in Section 3.2), the authors combined multiple translation candidates for each query term. This approach considered alternative translations as synonyms for query terms.

Ballesteros and Croft [1997] claimed that word-by-word translation failed to correctly translate phrases in queries. To solve this issue, they used a database of phrase and word translations provided in the form of Spanish-English MRD. In some cases, phrasal translation significantly outperformed word-by-word translation; however, in some cases, phrase translation degraded the retrieval performance compared to word-by-word translation. This happened when phrases were incorrectly translated. To filter bad phrasal translations, they employed local context analysis by excluding translated phrases that did not frequently appear in the top-ranked documents.

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

Ballesteros and Croft [1998] tackled the issue of translation ambiguity and short queries by expanding query terms before and after translation. This was done by applying local feedback from the collection. The local feedback assumes that top retrieved documents are relevant to the base query, and terms from these documents can be relevant to the information need. Authors showed that this kind of expansion helped improve the translation of base queries, and eventually improved the retrieval performance. The motivation of this approach was that a base query usually is written in a short and incomplete form, which leads to ambiguity in translation. The translation was done by using MRD between English and French.

3.1.3 Corpus-based Query Translation

This approach employs a dictionary that is extracted from an aligned corpus (on document level) to translate queries into the document language.

To produce a corpus of aligned documents, a set of comparable documents in two languages is created first. One way to create this set is by crawling news websites in a given period of time and obtain news articles in these two languages, or by using Wikipedia articles and use the interlanguage links between each article in the source language and its translation in the target language (if a translated article exists) [Tholpadi et al., 2017].

After preparing a set of documents in two languages that are aligned at a document level, a basic dictionary of vocabularies is needed to perform translations (word-by-word) of the sentences on the target side of the corpora. Then each sentence in the source language is used as a query to find the most similar sentence in the target side and consider it the translation for that query. If the two languages use the same alphabet, named entities (such as city names and dates) can also be used in the translation scoring function. The sentence with the highest similarity score on the target side of the corpus is then considered as a translation of the input sentence [Michelbacher et al., 2010].

One way to improve this method is to iteratively update the initial dictionary. This can be done by capturing lexical similarities and co-occurrences between words in two aligned sentences. This can help to predict the translations of more vocabularies [Rogati and Yang, 2002].

The dictionary-based approach can be supported by information that is extracted from the collection, Bosca et al. [2014] used multilingual semantic and domain-based information from the collection during indexing in order to map query fragments into concepts.

Talvensaari [2008] showed in their work that the main three factors which could 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.

Preparing comparable corpora requires documents to be available in all supported languages. We keep this approach out of our research scope, since we are aiming at developing CLIR systems for multiple languages, and we do not have access to sufficient resources to build comparable corpora for all of them. In addition to that, QT that is based on MT systems showed to be more effective, as we present in the following section.

3.1.4 Query Translation using MT system

Usually in CLIR, an MT 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 retrieval as a query.

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 queries into collection language. They also compared monolingual queries that were provided in English and the automatically translated ones, and they found that there was big difference in quality between them. This confirms the claims that further investigation should be put to improve the translation quality and disambiguating query terms.

Users usually use only 2 terms on 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 are usually not grammatically correct. 2) Queries expressed with 2 terms might not be sufficient 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 the CLEF eHealth 2013 dataset [Goeuriot et al., 2013] and it supported Czech-English, German-English and French-English pairs. The MT systems improved by an average of 55% in terms of the BiLingual Evaluation Understudy (BLEU) metric (BLEU is an automatic MT evaluation metric, presented in Section 5.4.1) 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 the translation quality does not guarantee to improve the performance of CLIR system.

Fujii et al. [2009] investigated the correlation between translation quality and retrieval quality in the cross-lingual question answering task (where the goal was to find answers to questions, not full text or documents as in IR), which is comparable to the CLIR task. They created search topics from the patent applications that were rejected. 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. 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 MT perspective) does not necessarily lead to the best retrieval quality.

A general MT system showed to perform well in CLIR, even when these MT systems were not adapted to translate queries in the domain of the collection. The overview of CLEF 2009 [Ferro and Peters, 2010] showed that using Google Translate to translate queries improved the CLIR results from 55% of the monolingual baseline in 2008 to more than 90% in 2009 for French and German languages. This can be explained because some improvements might be brought to the Google MT system during this year. However, using a generic MT system for CLIR has several drawbacks:

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

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

1. MT systems assume that the input sentence is syntactically correct, and the word order of it is linguistically meaningful, thus, this information is used for translation.

2. Some MT systems can produce multiple alternative translations (ranked by their translation score). However, when using generic MT systems, these alternative translations are ignored and one translation (the best one) is considered.

3. MT systems are usually trained to produce translations that are good to be read by humans. However, this is not important for CLIR performance.

MT should be adapted and integrated as an internal component of an IR system. This will help to keep the main objective goal to produce translations that perform better in retrieval, regardless if they have correct language structure or not (respect word order and follow a grammatically correct word morphological variation), like weighted alternative translations or translations that are represented in different morphological forms (forms, stems, lemmas).

In order to improve the translation of queries towards better CLIR performance, different approaches tried to expand or lexically process the query after translating it into target language.

Choi and Choi [2014] participated in the multilingual CLEF eHealth 2014 Task 3 [Goeuriot et al., 2014]. Firstly, they translated the queries (from Czech, French and German) into English using public MT system (Google Translate) following the QT approach. Then, they annotated each translated query with medical concepts using MetaMap [Aronson, 2001]. MetaMap is a tool that recognises medical entities in a given text, and annotates them with concepts that are taken from Unified Medical Language System (UMLS) entries. MetaMap returns a list of concepts ranked by their matching scores with the original text, thus, they selected top 5 concepts and added these concepts to the translated query.

Query terms which did not appear in the query's discharge summary were removed from the original query, assuming that these terms were not medical (not informative). Discharge summary is an official document that is usually released by a hospital for a given patient containing their diagnosis and treatment procedure during their stay at a hospital.

Discharge summaries were used in the query creation process of the CLEF IR task as a source of an information need. Lastly, they used structured query language (as it is implemented in Indri search engine [Strohman et al., 2005]) to weight the original query and the expanded one with 0.9 and 0.1 respectively. The authors reported that the medical concept annotation approach outperformed their baseline system (a system that uses the translated queries without any processing) by 18% for Czech, 4% for German and 4% for French.

A similar approach was followed 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, 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. The 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, only using Wikipedia to expand the queries did not help. Only a system that combined the Wikipedia approach with MetaMap improved the baseline system (original queries).

Translation Hypotheses Reranking

Translation hypotheses are usually ranked with respect to the translation quality, the main idea behind translation hypotheses reranking is to rerank these hypotheses towards better CLIR performance.

Nikoulina et al. [2012] defined the following two main challenges when adaptation an MT system for CLIR:

1. The adapted MT system should be able to produce translations that are good enough to hold correct information that is represented in the input query.

2. It is not enough to have good translations that hold the presented information from the source sentence, but also these translations should perform well in retrieval.

The previous two aspects (translation quality and retrieval quality) do not correlate with each other: sometimes syntactically wrong translations can lead to better retrieval performance than the correct ones.

Nikoulina et al. [2012] presented two approaches to tackle these two aspects: **query-genre tuning** and **reranking approach**.

Query-genre tuning focuses on the first challenge (translation quality). The authors tackled this challenge by tuning an SMT system to translate short queries taken from the Conference and Labs of the Evaluation Forum (CLEF) data set (news domain). This is done during the tuning step using Minimum Error Rate Training (MERT) algorithm [Och, 2003], and optimizing the SMT component weights towards higher BLEU scores (translation quality of queries). The motivation behind this approach is that SMT features have different impact on retrieval quality. Thus, the weights of these features should be tuned on query-like sentences rather than normal sentences. More details about the SMT parameters and tuning process are presented in Section 5.1.

Reranking approach tackles the second challenge by exploiting multiple translation hypotheses after being produced by SMT. This is done by selecting the best translation in terms of the retrieval aspect. To selection process is done by using Margin Infused Relaxed Algorithm (MIRA) [Crammer and Singer, 2003] that is trained directly towards retrieval quality, namely, the MAP metric. To conduct this, they first take the list of candidate translations for each training query, and generate a vector of features for that query, also they conduct retrieval for each translation hypothesis and obtain MAP score using the provided relevance assessment for the training queries. The loss function for their machine-learning model is the difference between MAP of each translation hypotheses, and the hypothesis that gives the highest MAP score (oracle translation). As features, they used internal features from the SMT decoder and syntax-based features extracted from the source queries and the translation hypothesis. They reported an improvement between 1% and 2.5% absolute on the CLEF AdHocTEL 2009 task (French to German) [Macdonald et al., 2006].

Ture and Boschee [2014] employed a similar approach. They used a set of binary classifiers to produce *query-specific* weights of various different features to select optimal translations from the translation hypotheses. They used three types of features: *surface features* such as number of token in the translation hypothesis, how many stop words appeared in the translation, and the category of the query using a pre-trained models (question query, cause-effect query etc.), *parse-based- features* such as number of named entities in the query, and part-of-speech tagging features (for example if the translation contains VVB in its parsing tree), *translation-based features* that are taken from the verbose output of the decoder and its alignment information, and *index-based features* that are taken from the collection. They reported significant improvements on several English-Arabic and English-Chinese tasks.

Sokolov et al. [2014] adapted an SMT system in CLIR in a different way. They added a new component inside the decoder of the SMT to directly consider the retrieval performance (using MAP metric) when generating translation hypotheses. This is done by combining IR-based weights and MT-based weights within the decoder itself, which makes the decoder prefer translation hypotheses that give better IR performance rather better MT performance. They reported stable improvements on the BoostCLIR task of Japanese-English patent CLIR [Sokolov et al., 2013].

Khwileh et al. [2017] proposed an approach to select the best translation from an n-best-list that is produced by an SMT system. To achieve this, they weighted each translation hypothesis as shown in Equation 3.1; where k is the number of terms in the query translation hypothesis Q, cf_t is the number of times t appears in the collection, and df_t is the number of documents that contain the term t. They called this weighting method Average Term Fluency (*AvgFL*). *AvgFL* predicts whether the translation hypothesis contains the same terms as in the documents that are relevant to the original query.

$$AvgFL(Q) = \frac{1}{k}\sum_{t\in Q}^{k}(log(cf_t + 1))/(log(df_t + 1) + 1)) \tag{3.1}$$

Their experiments on an Arabic collection and English queries (in the news domain) showed that *AvgFL* outperformed their baseline, which used only 1-best-list for each query as given by their English into Arabic SMT system.

In this section, we presented two approaches of reranking of MT hypotheses towards better CLIR performance. Both approaches were shown to outperform the use of the best translation from MT systems which are not tuned towards the CLIR task. In hypotheses reranking, we do not need to have access to the machine translation training data nor be involved in training the internal components of the MT system. We can employ any available MT system, which produces translation hypotheses, as a black box and develop the reranker after the translation process. It is also possible to integrate more features and make use of external resources in this approach. On the other hand, tuning MT systems to produce already optimised translations for CLIR requires access to training data with an IR metric instead of an MT metric, which might be an intensive task in terms of computational complexity.

3.1.5 Neural Approaches for Query Translation

Employing neural networks in CLIR has shown to be effective when used in different components of the CLIR task, either the translation part, or the retrieval ranking function. In this section, we will present the related work of employing word embedding and NMT models in the CLIR task.

Query Translation Using Word Embeddings

The main goal of using word embedding in NLP is to capture the context of words in documents and to find semantic and syntactic similarity between text. This is done by introducing a distributed representation of words as dense vectors. However, the word-embedding approach is not the first attempt to do so in NLP.

The idea goes back to Latent Semantic Analysis (LSA), which is considered to be the first approach that represents words as vectors in a semantic space [Deerwester et al., 1990]. The main hypothesis that LSA depends on is that similar words appear in the same parts of text (paragraph).

To create this vector representation, LSA uses Singular Value Decomposition (SVD) [Golub and Reinsch, 1970]. SVD is based on converting a matrix (normally 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 tells if the term exists in that document or not. The decomposition of this matrix using SVD will give us a matrix containing concepts, a matrix representing the strength of these concepts and a matrix representing terms as concepts (moving into semantic space). The first challenge that will come to mind is that loading the entire collection into the memory might 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 a wide set of algorithms including LSA.

Mikolov et al. [2013a] presented two word embedding models: the Continuous Bag-of-Words (CBOW) model and the continuous skip-gram model. CBOW predicts words given a context. The range of the context is called windows size (c). The 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 the CBOW model on the semantic questions set but CBOW outperformed it on the syntactic set. These two algorithms and vector representation of the top N most frequent words from a huge training corpus are presented as a neural-network based open source tool called *word2vec* [Mikolov et al., 2013a].

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 such as cosine similarity. The vector-based similarity is then combined with text-based similarity to rank the documents for a given query. For their experiments, they used TREC6, TREC7, TREC8 and TREC Robust datasets (these datasets are described in details in Section 3.4.1), and Lucene[5] for indexing and retrieving from the collection. To create an index for the collection, they used the *doc2vec* tool.[6] Each document was represented as set of vectors and had one or more clusters of words K (topics). Then the Language Model (LM) retrieval model was combined linearly with word2vec-based query likelihood. The final score of combination was given to each document for a given query. Their experiments showed that the hybrid model outperformed the text-based LM model on the Robust and the TREC-8 collection when using $K{=}100$ clusters. However, the experiments also showed that when using one cluster $K = 1$ (single-point representation) for representing the documents, results were similar to the text-based model. This occurred because using one cluster to represent a topic for each document was very little.

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 million documents. A candidate term was 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 is used for 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 candidate terms that are chosen by the model. Their method showed to be effective for query expansion.

Kim et al. [2016] used word embeddings to calculate 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 a *word2vec*

[5] `https://lucene.apache.org`

[6] `https://github.com/gdebasis/txtvecir`

model on 25 million articles from PubMed using their titles and abstracts and made the model available online.[7]

Litschko et al. [2018] employed an unsupervised shared cross-lingual word-embedding model for the translation part of CLIR, which was trained using monolingual data only. They used the embeddings to translate query terms (term by term) into the collection language. However, they recommended this method only when there is insufficient parallel data to train an MT system.

Query Translation Using NMT

NMT has shown a significant improvement in the machine translation task and researchers in the CLIR task have started recently investigating NMT employment in the translation part of the process.

Sarwar et al. [2019] proposed a model that is inspired by Relevance-based word embeddings to train an NMT model for query translation. The model is based on the Transformer architecture to translate Finnish and Italian queries into English (collection language). They employed the document collection in training the model by retrieving the top ranked document for each sentence in the target side of the parallel corpus (Europarl V7). Then they shuffled the retrieved document (after appending the sentence to it) and used it to train a word embedding model that was eventually used to train the NMT model. Creating a word embeddings from a document in the collection and a sentence from the parallel corpus helped the translation model to access the vocabularies in the document collection; hence, when translating a query into the target language, the model would be more likely to select a translation term that appears in both sources (document collection and parallel corpus). This approach, which they called Relevance-based auxiliary, was shown to outperform a strong baseline that employs an NMT model trained on the same parallel corpus by 16% of MAP of the baseline result.

Rücklé et al. [2019] presented a method for the Cross-lingual question-answering task, wherein the setup was to retrieve answers to German questions from an English collection in the technical domain (AskUbuntu and StackOverflow). They trained an NMT model using the WMT'14 English-German parallel data. The trained model did not perform well, since the WMT training corpora was created from an out-of-domain resource. To overcome this problem, they used the model to translate in-domain monolingual data from English into German. Then, they used the translations to create synthetic data, which eventually enriched the parallel data that is needed for the NMT model. This approach in NMT is called back-translation.

3.2 Query Expansion in IR

Web search user queries tend to be short. The average web search query length, as reported by Gabrilovich et al. [2009], is about 2.5 terms. The information

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

represented in these terms might be too brief and/or vague. This is considered to be a challenge for IR systems that follow the term-matching approach since they fail to find relevant documents that do not contain the terms specified in the query.

Query Expansion (QE) is the process of reformulating searcher's query by adding useful terms that can improve the information represented in the base query. The goal of QE is to reduce the term mismatch problem between a query and its potential relevant documents, which leads eventually to improvement in the retrieval performance [Efthimiadis, 1996].

QE can be done automatically, or by interaction with users (e.g., selecting one or more terms to be added to the query), which is known as interactive query expansion [Harman, 1988]. In this study, we will focus on automatic query expansion. Blind Relevance Feedback (BRF) is one of the most popular techniques for QE, also known as pseudo-relevance feedback [Rocchio, 1971]. First, an initial retrieval is conducted using the base query, and top n ranked documents are selected as a source for term candidates. Then each term in these documents is scored using some approaches like a combination of its Term Frequency (TF) in these documents and its IDF in the collection. Finally, the highest scored m terms are added to the base query, and a final retrieval is done. However, there is a risk when following this approach because one or more of these n documents might be irrelevant; thus, adding terms from these documents might drift the information away from the intended one. QE can have a significant improvement on one of the main evaluation metrics (such as MAP, P@10 or recall) and degrades the others; thus, the use of QE should consider the context of the IR application when using query expansion [Harman, 1992].

Pal et al. [2014] employed WordNet [Miller, 1995] to weigh candidate terms and measure their usefulness for expansion. They leveraged the similarity score of the top retrieved documents using the BRF assumption and excluded terms from WordNet, which do not appear in these documents. They calculated various similarity scores between the query term and the candidate term based on term distribution in the document collection. Then they linearly combined these scores to select the weights of the expansion terms. This approach brought an improvement over the use of base queries on multiple TREC collections.

Ermakova and Mothe [2016] used local context analysis by choosing terms that surround query terms from documents retrieved from the initial retrieval. They assumed that document terms that appear close to query terms are more likely to be good candidates for expansion. They tested their method on the TREC Ad-Hoc track datasets from three years (1997–1999) and the WT10G dataset [Chiang et al., 2005].

Cao et al. [2008] showed that when QE is only based on term distribution, it can not distinguish good terms, which will improve the IR performance and bad terms, which will harm it. They presented a classification model that is integrated into a BRF method. It uses features from the collection to predict the usefulness of the expansion terms and select only the good ones.

3.2.1 Expansion Using Word Embeddings

In the word embeddings model (word2vec [Mikolov et al., 2013b] and GloVe [Pennington et al., 2014]), the main objective is to capture the semantic and syntactic similarities between words. However, the goal in the word embeddings model for IR application is different, wherein the goal is to predict words in documents that are relevant to the query (information need eventually).

Recent researches have demonstrated the use of word embeddings based query expansion. The main idea is to expand the query with terms that are semantically related and appear in a position close to the query terms [Zamani and Croft, 2016, Nogueira and Cho, 2017a, Zamani and Croft, 2017]. Multiple researchers confirmed that embeddings models that are trained on medical data like PubMed articles are not significantly better than those that are trained on general domain data, such as news [Zuccon et al., 2015].

Zamani and Croft [2017] introduced Relevance-based Word Embedding (RWE), which is similar to BRF. But the main difference between BRF and RWE is that BRF is an online approach that requires to conduct retrieval for each query in the test set, while RWE is an offline model which does not require retrieval for each query. They evaluated their RWE for query expansion using the TREC collection. Their experiments showed that using the terms that are suggested by RWE improved the base query (as a whole). While using a typical word embeddings model suggested terms that are related to one or more query terms, not the whole query.

Kuzi et al. [2016] employed word embeddings in QE by training Word2Vec on the document collection, and then they select candidate terms from the collection that are related to the query terms by computing cosine similarity between query centroid $\overrightarrow{q}_{Cent}$ and each term in the collection. Query centroid is a vector calculated as is shown in Equation 3.2, by summing the vectors of all query terms. This helps represent multiple vectors (query term vectors) in one vector that is semantically similar to the query.

$$\overrightarrow{q}_{Cent} = \sum_{q_i \in q} \overrightarrow{q_i} \tag{3.2}$$

Then they selected v terms from the collection that have the highest score and use these terms for expansion, where v is a free parameter. They also experimented with different scoring methods like Fusion-based methods that normalize the cosine similarity between each query term vector and each term vector in the collection. Their experiments showed that employing word embeddings in QE can significantly improve the initial retrieval that is done without expansion.

Nogueira and Cho [2017b] employed word embeddings as a source of candidate terms by selecting *top-N* terms based on cosine similarity between their vectors and query term vectors. These terms form a candidate pool for expansion. They used pre-trained word embeddings released by Mikolov et al. [2013a]. Their supervised method used a binary classifier (based on CNN) to predict if a term from the candidate pool could improve the performance of the retrieval when added to the base query or not. Their supervised method outperformed the model that used only base queries when test on the TREC CAR dataset.

3.2.2 Kullback-Leibler Divergence

Kullback-Leiber Divergence (KLD) for query expansion is one of the most well-known query expansion approaches in IR. In KLD, the top n ranked documents (pseudo-relevant documents) are retrieved using a base query, then each term in these documents is scored by Equation 3.3, where $P_r(t)$ is the probability of term t in the pseudo-relevant documents, and $P_c(t)$ is the probability of term t in the document collection c. Finally the top m scored terms are added to the base query and a final retrieval is done using the new expanded query.

$$Score(t) = P_r(t) \cdot log\left(\frac{P_r(t)}{P_c(t)}\right) \tag{3.3}$$

3.3 Query Expansion in CLIR

In the QT approach, popular MT techniques struggle to translate short queries because of the lack of linguistic information that is required to solve the ambiguity, which eventually causes information loss in the translated queries [Pirkola et al., 2001]. Query expansion in CLIR helps solve translation ambiguity by adding relevant terms to the translated queries. In order to improve the translated queries, different approaches tried to expand or lexically process the query after translating it into the target language. Query expansion in the medical domain is a more difficult task than the general domain CLIR. Approaches that work on the general domain might not work well when applied in the medical domain.

Nikoulina et al. [2012] reported that simply merging the top 5 scored translation hypotheses (as a special QE approach) to create queries in the CLIR task outperformed the baseline system in the general domain data. However, the same approach did not work when we tested on the medical domain [Saleh and Pecina, 2016a].

KLD for query expansion (explained in Section 3.2.2) failed to outperform the baseline system (using initial queries) during the CLEF 2011 medical retrieval task [Kalpathy-Cramer et al., 2011].

Choi and Choi [2014] used Google Translate to translate the queries into English (from Czech, French, and German) during their participation in the CLEF eHealth 2014 CLIR task Goeuriot et al. [2014]. Then, they annotated each query with medical concepts using MetaMap [Aronson, 2001], and the top scored concepts were added to the original query. Finally, they weighted the original query and the expanded query with 0.9 and 0.1, respectively. The query expansion approach outperformed their baseline system relatively by 18% for Czech, 4% for German, and 4% for French.

Liu and Nie [2015] participated in the monolingual task of CLEF eHealth 2015 [Goeuriot et al., 2015], and presented a system which expanded queries with UMLS [Humphreys et al., 1998] concepts and terms extracted from Wikipedia articles. The main motivation by using Wikipedia was that a layperson usually poses a medical query using ordinary terms (not 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. The authors claimed that Wikipedia abstracts are similar to the way that users pose queries (more generic), while the titles of Wikipedia articles contain medical terms. However, only using Wikipedia to expand the queries did not help. Only a system that combined Wikipedia with MetaMap Aronson [2001] improved the baseline system. Employing Medical Subject Heading (MeSH)[8] for QE was investigated thoroughly.

Wright et al. [2017] presented a simple method that expands queries with five synonyms from MeSH. Nunzio and Moldovan [2018] expanded a query with one MeSH term that is related to the base query. In case there was more than one MeSH candidate term, they created multiple expanded queries. Then for each expanded query, they conducted retrieval and merged the retrieved documents by different approaches like averaging document scores or summing them.

Cao et al. [2007] considered query translation to be the first step in formulating the final query by expanding query translation with related terms. The authors integrated multiple relations between terms (monolingual using co-occurrences and cross-lingual using their dictionary translations) into a Markov Chain (Fok) model. Where the nodes represent terms, and edges between nodes represent the probability of relation between them. These relations can be term co-occurrences, translations from the dictionary, or *contain* relation (like *computer science* contains *computer*). Then formulating the final query is done by the Random Walk approach on the Fok model. The term cross-lingual similarities are adjusted during the Random Walk process, which helps produce stronger related terms to the base query. Because probabilities are adjusted according to the distribution of the original English query, their experiments on three CLIR collections significantly outperformed CLIR systems without expansion in which queries were translated using a dictionary-based method.

Chandra and Dwivedi [2017] used Google Translate to translate queries from Hindi into English in the FIRE 2008 dataset. Expansion terms were selected according to their Term Selection Value (TSV), as shown in Equation 3.4. Where R is the number of documents that are used to create a candidate pool from the

[8] https://www.nlm.nih.gov/mesh

top-ranked ones, and r_t is the number of documents that contain term t. Best results were achieved when R was set to top-3 ranked documents. They also reported that candidate terms that have the highest frequency in the retrieved documents are less important for expansion as those with lower frequency.

$$TSV_t = \left(\frac{f_t}{N}\right)^{r_t} \binom{R}{r_t} \tag{3.4}$$

3.4 Test Collections and CLIR Tracks

Cyril Cleverdon (a British librarian) is considered the first one who started working on developing systematic evaluation methods and test collection for IR.

Cleverdon was responsible for the famous IR project called the Cranfield project in 1960s [Richmond, 1963]. The goal of the project was to evaluate performance of various indexing systems for academic papers.

During the development of the project, Cleverdon realised that multiple assessors could not agree on some documents if they were relevant or not to the information need, because the information need was not clear enough. This caused the evaluation process to stop. This was the main obstacle of evaluating IR systems. Cleverdon suggested that before the evaluation process starts, the goal of the entire process should be clear. This includes the descriptors of the information need (query), agreeing on which documents are relevant to these queries, and define 5 levels of relevance: a document completely answers a given question, a document contains a high degree of relevance, a document contains information that can be useful as a background, a document contains very little amount of relevant information that can be considered as historical interest, or a document does not contain any interest to the original asked question. This is known as the Granfield paradigm [Cleverdon, 1960].

Later in 1990, National Institute of Standards and Technology (NIST)[9] built a new text collection to be used by DARPA (the Defense Advanced Research Projects Agency of the United States department of defense) for a project called TIPSTER IR [Tassey et al., 2010]. Later, the style of that test collection became known as the Text REtriveal Conference (TREC) style. It mainly contains the following:

- A document collection in which every document has a special identifier (document ID).

- A set of queries, where every query represents a topic or an information need and has a query ID.

- Relevance assessment (typically known as qrels set) that includes the relevance information for each assessed document-query pairs. A relevance degree can be either binary (*relevant*, *irrelevant*), multi-graded with 3 grades (*relevant*, *partially relevant*, and *irrelevant*), or 4 grades where the fourth one is called *highly relevant*.

We will discuss more in depth relevance assessment when building an IR test collection. Because relevance assessment is considered to be the most challenging part, since it includes an intensive human effort that is needed to annotate retrieved documents.

[9] `https://www.nist.gov`

The definition of *relevant* has two perspectives: *system view* and *user view*. Sun and Kantor [2006] considered relevance from end the user's perspective to be the gold standard in IR evaluation, however, evaluation should not rely entirely on users, since they sometimes cannot make a good judgement if the document content meets their need or not [Belkin, 1980].

Assessment can be done following one of two approaches: *dual assessment* in which two assessors judge the same document-query pair independently, and *review* approach wherein one assessor does an initial judgement and the results are later reviewed and corrected by a second assessor [White et al., 2005]. Different assessors tend to assess the same document-query pair differently. To evaluate the reliability of the assessment process, agreement rate is calculated using different approaches such as kappa statistics. If the kappa value is above 0.6 then the agreement is “acceptable“, 0.8 means the agreement is “perfect“ [Cohen, 1960].

The question of how many document-query pairs should be assessed in a test collection in order to be reliable has been always difficult to answer. Losada et al. [2019] studied different approaches to reduce the manual effort for relevance judgment, without reducing the quality of the test collection. The importance of this work is that basically if we do not define a systematic method to stop assessment, it means that we have to assess all document-query pairs in the collection, which can require a massive amount of resources (time and money). The authors presented the following stopping methods for relevance assessment (after creating a pool of retrieved documents for each query using some retrieval model):

- We keep assessing document-query pairs until we reach the *nth* documents for each query. This method considers a fixed number of assessed documents for all queries.

- The second method considers a variable number of documents to be assessed as follows:

  - Assess $x\%$ of the pool for each query.
  - Assess the pool for each query until we finish judging n relevant documents or n irrelevant documents.
  - Assess the pool until encountering n irrelevant documents consecutively.

They showed that sorting the documents in the pool using the relevance predictive method helps reduce the assessment effort. The relevance predictive method estimates the similarity between relevant documents and unjudged ones, where relevant documents are obtained using a small set of training queries.

We present in this section an overview of multiple CLIR tracks and campaigns. Table 3.1 shows a summary of the datasets that were released during these tracks, including statistics about each one's document collection, queries and the supported languages.

3.4.1 TREC

TREC is an annual event organized by NIST.[10] In 1997, TREC-6 was the first TREC event accommodating a CLIR track [Voorhees and Harman, 2000a]. The document collection included three sets of English, French and German documents taken from news agencies. 25 test topics in the same languages were created based on the interest of the participating assessors who performed binary relevance assessment for these queries. The TREC-7 CLIR track used the same document collection as in TREC-6 plus a set of documents and topics (28) in Italian [Voorhees and Harman, 1998]. The TREC-8 CLIR track used the same document collection as in TREC-7 with new set of 28 queries in the same four languages [Voorhees and Harman, 2000b]. TREC-9 ran a CLIR track with document collection aggregated from Chinese news agencies and 25 queries in English and Chinese [Gey and Chen, 2000]. In the TREC-10 CLIR track, an Arabic newswire document collection was used with a set of 25 topics created by assessors in Arabic and English and afterwards translated into French [Gey and Oard, 2001]. In TREC-11 [Oard and Gey, 2002], the same Arabic document collection as in TREC-10 was used with 25 newly created English topics then translated into Arabic. Assessors fluent in Arabic and English created corresponding Arabic topics and English versions of them. That was the last CLIR track organised by TREC.

3.4.2 NTCIR

NII (National Institute of Informatics in Japan) Testbeds and Community for Information access Research (NTCIR) is a project of NII.[11]

The first NTCIR workshop (NTCIR-1) was held on 1999 and aimed to improve linguistic research of Asian languages [Kando, 2001]. NTCIR-1 released a test collection which included scientific documents in Japanese and English, plus 83 Japanese topics with graded relevance assessment: (very relevant, relevant, partially relevant, irrelevant). NTCIR-2 worked with a collection of academic conference papers in Japanese and English and 49 topics in both languages. NTCIR-3 used a document collection of news in Chinese, Japanese and English with 50 topics in Chinese and 30 topics in Japanese and their translations into Chinese, Korean, Japanese and English. The same dataset was used in NTCIR-4 CLIR. The NTCIR-5 CLIR test collection included documents from news agencies in Chinese, Japanese, Korean and English and 50 search topics in all these languages with graded relevance assessment. NTCIR-6 exploited a document collection of newspaper articles. It reused the collection from NTCIR-5, 4 and 3 CLIR tasks and included 50 topics in Chinese, Japanese, Korean and English and additional documents from newspaper articles in Chinese, Japanese and Korean with graded relevance assessment too. NTCIR-7 ACLIA included CLIR as a subtask which included news articles in Chinese, Japanese and Korean, with 100 topics in Japanese and 100 topics in Chinese and 300 English topics and 3-level relevance assessment. NTCIR-8 ACLIA also launched CLIR subtask with documents in Chinese and Japanese with 300 topics in English.

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

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

3.4.3 FIRE

Forum for Information Retrieval Evaluation (FIRE) [Majumder et al., 2013] has been running since 2008 and aims to support research in multilingual information access for Asian languages. In FIRE 2008, a document collection of news articles in English, Hindi and Marathi was used with 50 queries in the same languages. In FIRE 2010, the 2008 document collection was enriched with new documents in Bengali. A set of 50 topics is manually translated into English, Gujarati, Marathi, Tamil and Telugu. FIRE 2011 used the same collection as in 2010, the queries were refined and interactive search was used to improve the relevance assessment.

3.4.4 CLEF

CLEF is one of the most famous initiatives that tackles the multilingualism of information system through multiple tasks.[12]

CLEF Ad-hoc aimed at developing retrieval system in monolingual and multilingual settings, and focus on a news document collection. This task was organized annually between 2000 and 2009. The document collections in 2000–2007 were collected from news agencies in several European languages and topics were generated in multiple languages to allow CLIR evaluation. In 2008 and 2009, the document collection was created in cooperation with the European Library [Di Nunzio et al., 2008].

The CLEF CL-SR (Cross-Language Speech Retrieval) task was organized annually in 2003–2007 and focused on searching in spoken English news archives using queries in five languages (Czech, English, French, German and Spanish)[Pecina et al., 2008].

3.4.5 CLEF eHealth IR Evaluation Labs

The CLEF ShARe/eHealth IR evaluation series has been organized since 2013 aiming at improving access to medical and health-related documents by laypeople and medical experts in monolingual and cross-lingual settings.[13]

We present a summary of each lab, while more details about the test collections in these labs will be presented in Section 4, since those test collections are the base of the data we use in our experiments.[14]

ShARe/CLEF eHealth 2013

The main purpose of this lab was to improve medical information access for patients and laypeople rather than medical experts. This was the main difference between this task and previous tasks [Goeuriot et al., 2013].

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

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

[14] http://catalog.elra.info/en-us/repository/browse/ELRA-E0042/

Dataset	Domain	#Docs	Doc lang.	#Queries	Query lang.
TREC-6	News	350K	DE, EN, FR	25	DE, EN, FR
TREC-7 & TREC-8	News	698,773	DE, EN, FR, IT	28	DE, EN, FR, IT
TREC-9	News	126,937	ZH	25	ZH, EN
TREC-10 & TREC-11	News	383,872	AR	25	AR, EN
CLEF eHealth 2014	Medical	1,09M	EN	50	CS, DE, EN, FR
CLEF eHealth 2015	Medical	1,08M	EN	66	AR, CS, DE, EN, FA, FR, PT
CLEF eHealth 2016&2017	Medical	52M	EN	300	CS, DE, ES, FR, HU, PL, SV
CLEF eHealth 2018	Medical	5.5M	EN	50	CS, DE, EN, FR

Table 3.1: Statistics of the presented CLIR datasets.

The lab included three tasks, Task 3 is an IR task. The organisers released document collection, queries, and relevance information, and participants were asked to submit their results (a list of ranked documents for each query in the test set) using their own approaches and resources.

For querying, the English queries were generated by clinical documentation reporters and nurses based on real discharge summaries to mimic the realistic patients' queries. Five queries were used for development purposes and 50 queries for testing. The document collection contained about one million English pages crawled from medical websites. No CLIR task was organized that year.

ShARe/CLEF eHealth 2014

In ShARe/CLEF eHealth 2014 Task 3 [Goeuriot et al., 2014], the queries were generated in the same fashion as in the previous year. In addition to the monolingual task, a CLIR task was introduced. Five development and 50 test queries were generated in English and then manually translated into Czech, German and French to simulate cross-lingual setting. The document collection was the same as in 2013.

CLEF eHealth 2015

In CLEF eHealth 2015 Task 2 [Palotti et al., 2015], the query creation aimed to implement self-diagnosing case. Non-expert student volunteers were shown images of symptoms of specific conditions and asked to create three different queries (in English) for each symptom. 66 queries were then randomly selected and used for testing (plus 5 queries for development). The queries were manually translated into Arabic, Czech, French, German, Farsi and Portuguese. The 2015 collection was a subset of the 2014 collection (a few websites were removed).

CLEF eHealth 2016

In CLEF eHealth 2016 Task 3 [Kelly et al., 2016], a new document collection was introduced (ClueWeb12 B13[15]). The collection contained web documents from both medical and non-medical domains in an attempt to give more realistic representation when users look-up information from the web (generic collection). As queries, first a set of posts where extracted from the AskDocs forum[16] which contains medical questions that were asked by users to get answers from online experts regarding their health conditions. Only posts which contained clear and comprehensible questions where chosen to create a pool of queries. Then for each query in that pool, six query variations were created by three medical experts and three people without medical knowledge resulting into the final set of 300 queries representing 50 topics. This approach aimed to design an information retrieval system that is robust to different representations of the same information need.

The queries were translated (by medical experts) into Czech, French, German, Hungarian, Polish, Spanish and Swedish to allow CLIR experiments.

CLEF eHealth 2017

CLEF eHealth 2017 IR Task used the same collection and queries as in 2016. However, an additional assessment was performed [Palotti et al., 2017].

CLEF eHealth 2018

CLEF eHealth 2018 Consumer Health Search Task released a document collection created using CommonCrawl platform [Jimmy et al., 2018] containing $5,560,074$ documents that were crawled from $1,653$ websites. 50 queries were provided in English in the monolingual task (IR Task 1 Ad-hoc search). In IR Task 4 (Multilingual Ad-hoc Search) the same English queries were provided in French, German and Czech.

3.5 Conclusion

We presented in this chapter an overview of the CLIR task and its related work. This included the main approaches (QT and DT) to conduct CLIR, and how multiple MT approaches (dictionary-based, SMT, NMT and word embedding) can be employed in the translation part of CLIR. Lastly, we summarized various tracks and campaigns that were organized in the CLIR context. We focused on the evaluation lab series of CLEF eHealth since we adopted the test collections that were released during this series, and we extended them as we show in the following chapter.

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

[16] https://www.reddit.com/r/AskDocs/

4. Test Collection

The test collection, which we use in our work, is based on three test collections that were released during the CLEF eHealth patient-centered IR tasks 2013–2015 [Goeuriot et al., 2015, 2014, Suominen et al., 2013]. We extend the test collection mainly by translating the queries into more languages, and enriching the relevance assessment. The extended test collection is published online on the LINDAT/CLARIN repository.[1] We described the extended test collection in a short paper that was published in the 41th European Conference in Information Retrieval (ECIR) 2019 [Saleh and Pecina, 2019a].

In the following sections, we present more details about the collection and our contribution to its parts.

4.1 Document Collection

As the document collection, we use the one that was released during the 2015 eHealth Task 2: User-Centred Health Information Retrieval [Goeuriot et al., 2015]. The collection was created within the Khresmoi project [Aswani et al., 2012]. It includes $1,096,879$ English documents that were crawled from medical web sites. These web sites were verified by the Health On the Net (HON) foundation.[2] In addition to those websites, the collection also includes famous websites that include medical-related topics such us Trip Answers,[3] Diagnosia[4] and Drugbank.[5]

It is important to mention that the document collection in the 2015's task is almost identical to the collections that were used in the two previous years of the CLEF eHealth lab. However, some documents were removed from the 2015's collection due to copyright issues.

Document Processing

The documents in the CLEF eHealth IR Task2 dataset were provided by the organizers in the HTML format. Each document contains HTML markup, and possibly CSS and javascript code. These scripts are not informative, and they do not contain the actual text; thus, we decide to exclude them from the index by cleaning the document collection from such scripts.

In order to decide about the cleaning approach, we investigate three main methods. In the first method, we apply a simple script that uses the Perl module HTML-Strip.[6] It removes all the HTML code, CSS and other scripts in HTML pages, and keeps only raw text. We make an exception for meta keywords and description tags. Important information about content of HTML page is sometimes

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

[2] https://www.hon.ch/en/

[3] http://tripanswers.org

[4] http://diagnosia.com

[5] http://drugbank.ca

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

method	size (MB)	%	length (mil. tokens)	%	avg length (tokens)
none	41,628	100.00	–	–	–
HTML-Strip	6,821	16.38	1,006	100.00	911
Boilerpipe	3,248	7.80	423	42.11	383
JusText	2,853	6.85	452	44.93	409

Table 4.1: Collection size (in MB and millions of tokens) and average document length (in tokens) after applying the different cleaning methods

encoded as attributes of an HTML tag such as *meta* tag or *keywords* tag, so stripping off these tags will cause loss of these attribute values. For that reason, we keep the content of these attributes. After cleaning the collection using this script, its size reduced from 41,628 MB to 6,821 MB, which means around 16% of the original size.

The second cleaning method is Boilerpipe [Kohlschütter et al., 2010]. It reduces the total number of tokens in the collection by 58% (the average document length is 383 tokens). Boilerpipe attempts to remove the noisy text (like menus, bars, header and footer) around the main text (body) of a web page. It employs text features for classifying each text element in a web page if it is noise or not.

Lastly, we experiment JusText for document cleaning [Pomikalek, 2001]. JusText removes the boilerplate and duplicate content and keep the main text in web pages. It reduces the collection by 55% (the average document length is 409 tokens).

Table 4.1 shows statistics about the document collection after being cleaned with these three methods.

To compare the effects of the cleaning methods on the IR performance, we create three IR indexes from the documents that are cleaned using each method separately. Then we run a monolingual IR system using the English training queries, and evaluate each system results (queries are described in Section 4.2). We find that the IR system that uses HTML-Strip tool for document cleaning significantly outperforms other results. Even this cleaning method reduces the collection size by about 84% of its original size, documents still contain a significant part of the informative text, which seems to be enough to distinguish the relevant documents from the irrelevant ones, at least compared to the more advanced cleaning methods. This can be explained because the two other advanced methods are too aggressive and remove informative content from the documents. In all our following experiments, the document collection is cleaned by HTML-Strip.

We reported our document cleaning approaches during our participation in the ShARe/CLEF eHealth 2014 shared IR task [Saleh and Pecina, 2014].

4.2 Queries

The queries in this work are adopted from the test sets that were released during the CLEF eHealth IR tasks 2013–2015.

Queries from 2013 and 2014 In the CLEF eHealth IR task 2013 [Goeuriot et al., 2013] and CLEF eHealth IR task 2014 [Goeuriot et al., 2014], the queries were generated by medical experts from discharge summaries of patients. The motivation of choosing medical experts (nurses and clinical practitioners) for query generation was that those experts were in touch with patients on daily basis; thus, they could understand their information needs. The queries were generated as follows: medical experts were given discharge summaries and they were asked to select randomly a disorder from them, then to write a short query describing it. They assumed that patients would use the same query when they want to find more information about the same disorder. Involving medical experts to generate queries from discharge summaries affected the nature of the queries in a way that they contain medical terms, and they tend to be short. The queries in 2013 were available only in English, and the queries in 2014 were officially translated by medical bilingual experts from English into Czech, French and German.

Queries from 2015 In the CLEF eHealth Evaluation Lab 2015, the IR task was called *Retrieving Information About Medical Symptoms* [Palotti et al., 2015]. The goal of the task was to design IR systems that could help laypeople (users without medical experience) find information related to their health conditions and understand what caused their symptoms (self-diagnosis). Thus, the creation of the queries in this task attempted to simulate the real case as much as possible. Participants in the query creation step were university students without medical experience, as an attempt to simulate the case of an average search engine user. They were shown images and videos that contained symptoms of medical issues. Then, they were asked to generate queries for each case, as they thought those queries would represent their information need, and eventually would lead them to relevant documents. The queries in 2015 were created in English and officially translated by medical experts into Czech, French, German and Spanish.

New Data split We showed in the previous two paragraphs the main difference between the queries in the 2013, 2014 and 2015 IR labs of CLEF eHealth. Our motivation for introducing new split is to design a CLIR system that is stable for such a diversity of user queries, rather than designing a system that is biased to one type of them (short queries with medical terms, or long queries without medical terminology).

To remedy this, we get the test queries from each IR task in 2013 (50 queries), 2014 (50 queries), and 2015 (66 queries). We mix them to get more representative and balanced query set, and then split these queries into two sets: 100 queries for training (33 queries from 2013 test set, 32 from 2014 and 35 from 2015) and 66 queries for testing (17 queries from 2013 test set, 18 queries from 2014 and 31 from 2015). The two sets are stratified in terms of distribution of the year of their origin, number of relevant/irrelevant documents that exist in the relevance assessments, and the query length (number of tokens). The query ID tags in our new split preserve the original IDs which allows mapping the queries to their original year.

Table 4.2 shows samples from all years. It is clear from the samples how queries in 2015 are relatively longer than others, and written in a simpler language without the use of advanced medical expressions.

query id	query 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 4.2: Samples of the English test queries from the CLEF eHealth IR tasks 2013–2015

4.3 Manual Translations of Queries

The motivation of the manual translation of the queries is to support new languages for CLIR experiments. Our goal is to make all queries available in the supported languages in this research: Czech, French, German, Hungarian, Polish, Spanish and Swedish. Howevere, as we showed in the previous section, queries from the CLEF eHealth IR labs were not available in all languages. To achieve our goal, we asked medical experts who were fluent in English and one of target languages to translate the English queries into the target language. This manual translation followed the same instructions that were provided during the CLEF eHealth tracks to the official translators [Goeuriot et al., 2014, Urešová et al., 2014].

First, the translators tried to translate the queries into the target language and keep the syntax as much as possible, however, in case that was not possible (since the input queries were not grammatically correct in some cases), they conducted term-by-term translation. Then, linguist experts were asked to look at the queries to check their fluency and adequacy. The last step was to ask the medical experts again to check and find if there was any harm to the medical terminology during the linguistic check.

At the end of the manual translation process, 166 queries were available in a total of 8 languages (the original English plus human translations into Czech, French, German, Hungarian, Polish, Spanish, and Swedish). English queries allow monolingual IR since the collection is available in English. And non-English queries open the doors for CLIR experiments in those 7 languages.

Our contribution of manual translation includes translation of the 2013 queries into 7 languages, since in the 2013 CLEF eHealth IR task only English queries were available, translation of the 2014 queries into Spanish, Hungarian, Polish and Swedish, and translation of 2015 queries into these languages too.

4.4 Machine Translation of Queries

In addition to the human translation of the queries from English into the target languages, we include in the data package queries that are machine-translated back into English. This allows researchers to conduct CLIR experiments without having an access to an MT system. The MT system, that is used to translate the queries into English, is a phrase-based SMT system that is adapted to translate queries from the medical domain. The system is fully described later in Section 5.1 (QT-SMT-form). For each input query, the system generates a list of 1000 ranked translation hypotheses (*n-best-list*) including internal system information and scores for each one of them (the verbose output of the MT system). We provide in our extended dataset both *1-best-list* (best translation) and *1000-best-list* English translations for queries in all the languages.

4.5 Relevance Assessment

We present in this section our contribution to the assessment of the extended test collection. Our assessment procedure contains the following steps:

1. Document pooling, in which a set of highly ranked documents for each query is created to be assessed.
2. Relevance judgement where judges (humans with experience in the domain) determine if each document is relevant or not to a given query.
3. Dual assessment to check the reliability of the assessed documents.

4.5.1 Document Pooling

Relevance assessment is an expensive task; thus, it is often impossible to assess all document-query pairs in a test collection. Document pooling is the process of determining which documents need to be assessed taking into consideration minimizing the time that is needed by the assessors as much as possible, on the other hand, to consider obtaining a full coverage of the system results.

Full assessment depends on the metric that is used for evaluation. For example, if the goal is to calculate MAP, then we need to assess the entire retrieved documents for each query. We consider in our research P@10 as the main IR metric (more information about IR evaluation is presented in Section 2.4). To guarantee fully evaluated results, we need to ensure that the top 10 ranked documents are assessed for each query, then we can confirm that P@10 scores are reliable. Because when a document is not assessed, it is considered as irrelevant (the behaviour of the standard TREC evaluation tool), which is a risky assumption because if an unjudged document is relevant but treated as irrelevant, then the evaluation of the system will be less than the reality. Which makes system comparison and drawing an accurate conclusion about which system is better not accurate.

To build a document pool for additional assessment, we run the following experiments:

- Monolingual system (Section 6.1).
- Baseline query translation system using QT-SMT-form system (Section 6.2.1).
- Query translation system using public MT systems (Section 6.2.4).
- Translation hypotheses reranker (Section 6.2.2).
- Query expansion based on term selection (Section 6.4).
- Document translation system (Section 3.1.1).
- Query translation using QT-NMT-form system (Section 6.2.3).

For each CLIR system in each language, we collect all the unjudged documents from the above experiments (only from the top 10 ranked ones), then we remove duplicated document-query pairs among all systems. At this point, we have a document pool of a set of $14,368$ document-query pairs that need to be manually assessed by human experts.

4.5.2 Relevance Judgement

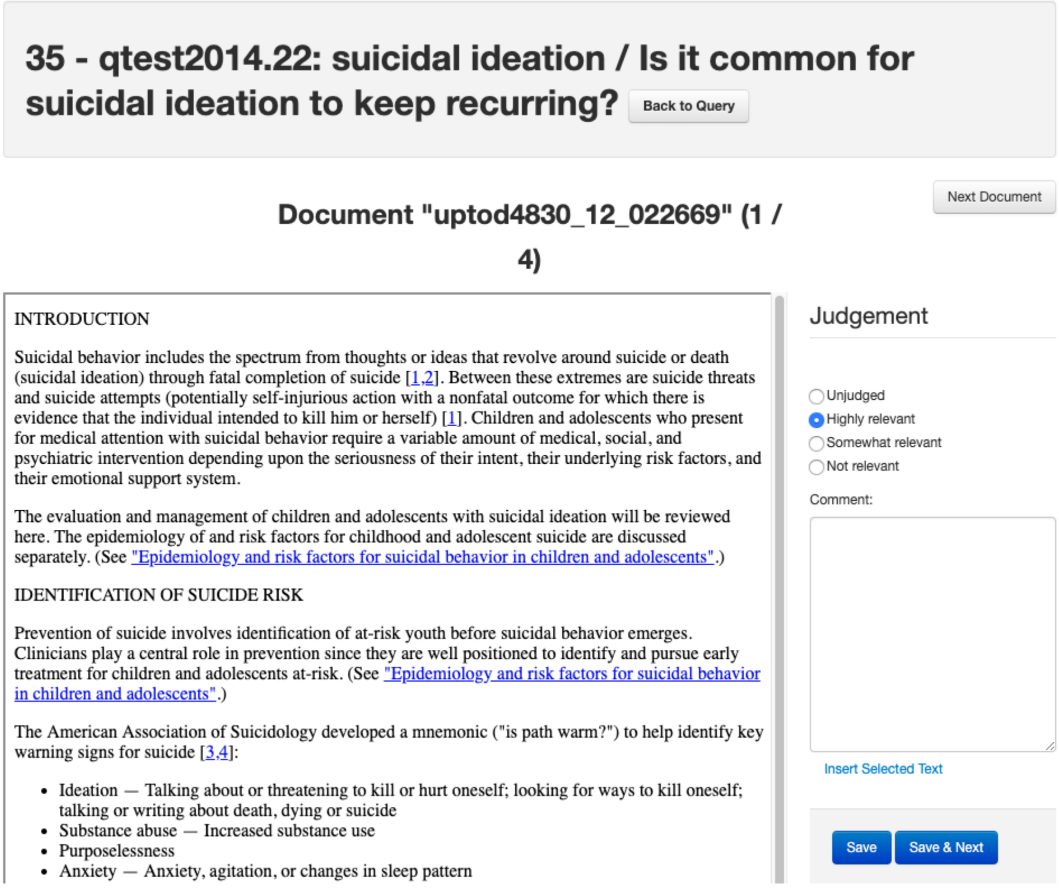
After we prepared the document pool, we setup the Relevation system online, which is an open-source tool for conducting relevance assessment for IR evaluation [Koopman and Zuccon, 2014]. Then we asked medical experts who were fluent in English to conduct the assessment, each assessor had to assess a set of queries, which were taken randomly.

Figure 4.1 shows the web interface of Relevation for a sample document-query pair. At the top of the page, the assessor can see the title of the query, and its narrative title which describes to the assessor what exactly a relevant document should include. Using only query title is not enough for assessing the relevance of the document, since the query title represents how a user thinks their information need should be formulated, which can be vague and ambiguous. The assessor can make their judgment on a document after they read it.

The assessor judgement can be one of the following choices:

- **Not Relevant** when the document is not related at all to the information need.
- **Somewhat relevant** the document partially answers the information need. This means that some information is missing and searcher needs to read more documents to completely get their question answered.
- **Highly relevant** the document completely satisfies the information need, and no need to read any other documents.

During the assessment, the assessors gave us feedback that on average, it takes around 20 seconds to judge a document if it is irrelevant, and up to 2 minutes if a document is relevant. In other words, it is faster to judge an irrelevant document than a relevant one.



35 - qtest2014.22: suicidal ideation / Is it common for suicidal ideation to keep recurring?

Document "uptod4830_12_022669" (1 / 4)

INTRODUCTION

Suicidal behavior includes the spectrum from thoughts or ideas that revolve around suicide or death (suicidal ideation) through fatal completion of suicide [1,2]. Between these extremes are suicide threats and suicide attempts (potentially self-injurious action with a nonfatal outcome for which there is evidence that the individual intended to kill him or herself) [1]. Children and adolescents who present for medical attention with suicidal behavior require a variable amount of medical, social, and psychiatric intervention depending upon the seriousness of their intent, their underlying risk factors, and their emotional support system.

The evaluation and management of children and adolescents with suicidal ideation will be reviewed here. The epidemiology of and risk factors for childhood and adolescent suicide are discussed separately. (See "Epidemiology and risk factors for suicidal behavior in children and adolescents".)

IDENTIFICATION OF SUICIDE RISK

Prevention of suicide involves identification of at-risk youth before suicidal behavior emerges. Clinicians play a central role in prevention since they are well positioned to identify and pursue early treatment for children and adolescents at-risk. (See "Epidemiology and risk factors for suicidal behavior in children and adolescents".)

The American Association of Suicidology developed a mnemonic ("is path warm?") to help identify key warning signs for suicide [3,4]:

- Ideation — Talking about or threatening to kill or hurt oneself; looking for ways to kill oneself; talking or writing about death, dying or suicide
- Substance abuse — Increased substance use
- Purposelessness
- Anxiety — Anxiety, agitation, or changes in sleep pattern

Judgement

○ Unjudged
● Highly relevant
○ Somewhat relevant
○ Not relevant

Comment:

Figure 4.1: The web interface of Relevation as it is used by the assessors to conduct relevance assessment

4.5.3 Dual Assessment

Different assessors can argue on the relevance of a document-query pair; this is a normal phenomenon in manual assessment not only in information retrieval but also in different fields such as machine translation. The degree of agreement among multiple assessors is referred to as *agreement rate*.

Cohen [1960] introduced the *kappa statistics*, which is a measurement of agreement on judgments that were conducted by two judges. The *kappa statistics* was used in social sciences for the categorical outputs from the judgment (or ratings) decisions, and it was adopted later to calculate the agreement rate in relevance assessment in IR [Manning et al., 2008].

We use *kappa statistics* as shown in Equation 4.1, where $P(A)$ is the probability that two assessors agreed on the relevance degree for a given document-query pair, $P(E)$ is the probability that they agreed by chance. $P(E)$ can be considered as the probability of making a two-class decision, then it can be set always to 0.5, or it can be considered as the probability of a marginal decision which is defined as shown in Equation 4.2. When two assessors agree on all document-query pairs, *kappa* is 1, and 0 if their agreement rate is equal to agreeing by chance.

$$\text{Kappa} = \frac{P(A) - P(E)}{1 - P(E)} \tag{4.1}$$

$$P(E) = P(irrelevant)^2 + P(relevant)^2 \tag{4.2}$$

In order to accept the relevance assessment and use it in system evaluation, the agreement rate between multiple assessors (*kappa*) is recommended to be higher than 60%, as shown by Manning et al. [2008].

We conducted relevance assessment three times as we proceeded with our experiments. After each phase, we randomly picked up 20% of the assessed document-query pairs, which were assessed by assessor A, and asked assessor B to assess them again, without telling B what was A's assessment to avoid any bias in B's decision. We similarly picked up 20% from B's pairs, and asked A to assess them. At the end, the agreement rate between the two assessors was 80.2%, which is considered to be high enough to accept the assessment as reliable for evaluation.

Table 4.3 shows statistics of the official assessment versus our extension in terms of the number of assessed documents. The extended dataset contains a total of $38,109$ document-query pairs, $14,368$ pairs of them are assessed by us.

	CLEF 2013	CLEF 2014	CLEF 2015	Our extension	total
relevant	1,174	3,209	2,515	2,517	9,415
irrelevant	3,676	3,591	9,576	11,851	28,694
total	4,850	6,800	12,091	14,368	38,109

Table 4.3: Relevance assessment statistics

4.5.4 Conclusion

We employed the official datasets that were released during the CLEF eHealth Evaluation Lab IR tasks 2013–2015 to present a unified and enlarged dataset for CLIR experiments. Our contribution to the new extended dataset includes the manual translations of the queries into seven languages: Czech, German, French, Hungarian, Polish, Spanish and Swedish. In addition to the manual translations, we provided machine translation of these queries including 1000-best list for researchers who do not have access to medical MT systems from the seven studied languages into English.

The official assessments suffer from low coverage, which makes it difficult to fully and reliably evaluate IR systems. To remedy this limitation of the existing assessment, we ran assessment three times (January 2016, July 2017 and December 2017) as we proceeded in our research. At the end, we enlarged the relevance assessment by more than double times of the original ones. The document pool in these assessment operations was created based on diverse CLIR methods, which makes the assessment more likely to cover more methods and approaches by those who want to use it. We published this work [Saleh and Pecina, 2019a], and made the extended dataset available via the LINDAT/CLARIN repository.[7]

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

5. Machine Translation for CLIR

In this chapter, we present background and theories of the recent approaches in implementing MT system, namely statistical machine translation and neural machine translation approaches. These two approaches are used later in this work to handle the translation part within the CLIR task. We also present methods for data selection and domain adaptation. These two concepts are critical when working in a domain-specific task (the medical domain in our research).

5.1 Statistical Machine Translation

SMT systems had been the state-of-the-art approach for MT for a long time, before NMT systems emerged and significantly outperformed SMT systems [Bojar et al., 2016].

In phrase-based SMT, sentences are first segmented into phrases which are translated as atomic units, contrary to the word-based models where each word is translated separately. Translating phrases allows the use of context within those phrases. This helps to solve translation ambiguity, since context includes information that is employed in the decoding (finding the best translation) process, which gives SMT the advantage of better translation. If enough training data is available, sometimes SMT considers one long sentence (practically up to 10 words) as one phrase, which produces more adequate and comprehended translations.

The translation model is learnt from a parallel training corpus. It is based on the noisy channel approach [Shannon, 1948] as shown in Equation 5.1. Where f is a foreign sentence in a source language, e is its translation in a target language, $p(f|e)$ is a translation model that is trained on a parallel corpus and it is based on the noisy channel model (Bayes rule is used here), and $p(e)$ is an LM trained on a monolingual corpus [Koehn et al., 2003].

$$\hat{e} = \arg\max_{e} p(e|f) = \arg\max_{e} p(f|e)p(e) \tag{5.1}$$

SMT employs monolingual data in the target language to train $p(e)$ (LM) that is involved in the decoding process. LM helps to overcome the ambiguity when a word has multiple translations by capturing contextual information that is represented in the surrounding words. This helps produce translations that are more grammatically correct and more adequate in the target language.

Equation 5.1 can be expressed as a log-linear model, which can combine more features (h_m). These features come from different separate models, and each feature has a weight (λ_m) that is assigned during the SMT tuning process [Koehn et al., 2007].

$$p(e|f) = \frac{1}{Z} exp(\sum_{m} \lambda_m h_m(e,f)) \tag{5.2}$$

The most important features are:

- Translation log probability calculated by the translation model for each phrase (or word) w in a sentence t, known as $p_w(t|s)$.

- Log probability $p(w)$ of the candidate translation w given by the language model of the target language.

- Word penalty which penalizes the number of words in the target translation, and phrase penalty that penalizes the number of phrases used in the translation. These two features help control the length of the translations by penalising the long ones.

- Log probability of the lexical re-ordering model, which is learnt from the parallel data, and it determines the likelihood of a phrase to follow the previous one, or to be disconnected from it [Koehn et al., 2005].

Figure 5.1 shows samples of 5-best-list translations of the Czech query *příznaky a aortální insuficience*, as given by a Czech-to-English QT-SMT-form system (see Section 5.1 for more details). The translation hypotheses are ranked according to their final scores as given by the decoding function (Equation 5.2).

This verbose output is taken from the Moses decoder [Koehn et al., 2007]. The output contains valuable information that we employ in our work as we will show later, such as scores given for each translation by multiple internal components of the SMT system. These components are the translation model, the language model, the distortion (re-ordering cost) model and word penalty. The last score in each sentence is the log probability score of the entire translation hypothesis which is a linear combination of these model scores.

The figure also shows the alignment information of the input sentence and the output translation (numbers between the two pipes after each segment in the sentence). This defines the phrases in the output and their alignment to the phrases in the input. One limitation of the diversity of these translations in CLIR is that sometimes multiple translations differ only in word order, punctuation marks and stop-words (such as *of*, and *and* in Figure 5.1), which are ignored when conducting a bag-of-words based retrieval with a stop-words removal.

We can observe here the difference between the translation task in CLIR and the normal translation task, when such information (stop-word and word order) is important and sometimes change the entire meaning of the sentence in the normal translation task, but it is totally ignored in CLIR systems that follow bag-of-words approach (treating documents and queries as a sequence of words ignoring their order).

```
0 ||| symptoms |0-0| of |1-1| aortic |2-2| insufficiency |3-3|
  ||| LM0= -25.1603 Distortion0= 0 WordPenalty0= -4
  PhraseDictionaryCompact0= -7.57569 -5.72601 -6.74157 -8.71513
  3.99959 ||| -2.10535

0 ||| symptoms |0-0| and |1-1| aortic |2-2| insufficiency |3-3|
   ||| LM0= -32.8526 Distortion0= 0 WordPenalty0= -4
  PhraseDictionaryCompact0= -3.72686 -1.24893 -2.05201 -3.79505
  3.99959 ||| -2.15809

0 ||| aortic |2-2| insufficiency |3-3| and |1-1| symptoms |0-0|
   ||| LM0= -28.982 Distortion0= -7 WordPenalty0= -4
  PhraseDictionaryCompact0= -3.72686 -1.24893 -2.05201 -3.79505
  3.99959 ||| -2.29053

0 ||| signs |0-0| and |1-1| aortic |2-2| insufficiency |3-3|
  ||| LM0= -34.3881 Distortion0= 0 WordPenalty0= -4
  PhraseDictionaryCompact0= -4.50501 -2.3473 -3.22388 -4.60657
  3.99959 ||| -2.37274

0 ||| signs |0-0| of |1-1| aortic |2-2| insufficiency |3-3|  |||
   LM0= -28.0266 Distortion0= 0 WordPenalty0= -4
  PhraseDictionaryCompact0= -8.35384 -6.82438 -7.91344 -9.52665
  3.99959 ||| -2.41201
```

Figure 5.1: Samples of 5 translation hypotheses from Czech-English QT-SMT-form system

5.2 Machine Translation Data

In this section, we describe the data that is used to train our MT systems to translate medical queries from seven languages (Czech, French, German, Spanish, Hungarian, Polish and Swedish) into English and to translate documents from English into these seven languages.

5.2.1 Parallel Data

The parallel data is used to train both the SMT (Section 5.1) and NMT (Section 5.3) systems. It is taken from various resources for all seven languages. These resources are split into *in-domain* data and *general-domain* data.

In-domain data

The in-domain data in this work is taken from medical resources, such as:

- The EMEA corpus (The European Medicines Agency), which contains parallel data constructed from biomedical documents in the pdf format obtained from the European Medicines Agency and prepared by Tiedemann [2009]. The EMEA corpus is available in 22 European languages.

- The UMLS metathesaurus, which contains $12,966,290$ distinct medical concepts that are translated into multiple languages such as English, Czech, French, German, Hungarian, Spanish, Swedish and Polish [Humphreys et al., 1998].

- MuchMore corpus, which includes around 1 million tokens in each language extracted from the abstracts of 6000 papers published in 41 scientific journals [Buitelaar et al., 2003].

- The MAREC patent collection that includes around 19 million patent documents provided in 19 languages including English, French and German [Wäschle and Riezler, 2012].

- The Corpus of Parallel Patent Applications (COPPA), it contains sentences extracted from the titles and the abstracts of the patent applications that were submitted to the Patent Cooperation Treaty[1] during 1990 and 2010. This parallel corpus is available for the English-French pair only [Pouliquen and Mazenc, 2011].

- Titles from Wikipedia articles in the medical categories, translations of these titles are obtained using the inter-lingual links of Wikipedia articles.

General-domain data

The general domain data is more easy to obtain than the in-domain data. In this work, we employ an extensive amount of parallel data that is taken from

[1] https://www.wipo.int/pct/en/

multiple bilingual resources, such as CommonCrawl [Smith et al., 2013], JRC-Acquis [Steinberger et al., 2006], News commentary of the Syndicate project [Callison-Burch et al., 2012], OJEU corpus [Forcada et al., 2011], the European Parliament interpretation corpus [Koehn, 2005], and DBpedia-based dictionary entries.[2]

5.2.2 Monolingual Data

The monolingual data is used to build a language model of a target language during development of an SMT system. The language model helps select a candidate translation that is as coherent and fluent as possible in the target language (for our CLIR experiment, this is important for document translation, but less important for query translation). We also use the monolingual data in development of NMT models through the back-translation approach as we will show later in Section 5.3.4.

The source of the monolingual data came from the English part of the MultiUN corpus [Eisele and Chen, 2010], which includes around 14K sentences, 22,197 sentences from the Gigaword news headlines [Parker et al., 2011] and 44,285 sentences extracted from news articles that were distributed during WMT (Workshop on statistical Machine Translation) 2009–2012. The monolingual English data also includes medical text taken from the document collection that is used in CLEF eHealth 2013 IR task [Goeuriot et al., 2013]. The data is described in details by Pecina et al. [2014].

5.2.3 Development and Test Sets

Our goal is to develop and tune two types of MT systems in the medical domain, one for query translation (short and incomplete sentences) and one for document translation (long and complete sentences). For that purpose, we adopt two sets for tuning and for the final evaluation:

The Query Test Set

We use dataset of medical-domain queries coming from two sources: the first source is the HON website,[3] where 749 English queries were randomly chosen from queries which were asked publicly. The second source is queries asked by medical experts in Trip database,[4] which includes 759 English queries. This set was prepared and published by Urešová et al. [2014].

We use from these two sources 508 queries for tuning and 1,000 queries for the final MT evaluation. Samples of this set are shown in Table 5.1.

The Summary Test Set

This set includes summaries of 1500 English sentences selected randomly from English medical documents that were crawled from medical websites. Then

[2] https://wiki.dbpedia.org

[3] http://www.hon.ch

[4] http://www.tripdatabase.com/

Language	**Query**
English (Ref.)	*gastric bypass*
Czech	*žaludeční bypass*
German	*Magenbypass*
French	*bypass gastrique*
Hungarian	*gyomor bypass*
Polish	*Ominięcie żołądkowo - jelitowe*
Spanish	*derivación gástrica*
Swedish	*gastric bypass*

Table 5.1: Samples of queries in English and their manual translations in the seven languages taken from the Khresmoi Query Translation Test Data

professional native speakers translated these 1500 document summary sentences from English into the target languages. This set was developed within the Khresmoi project,[5], and supported three language pairs: English-Czech, English-French and English-German. These translations were extended within the KConnect project to include translations into Hungarian, Spanish, Swedish, and Polish. Table 5.2 shows samples of these sentences, which are complete and long in general; thus, we use summaries to tune and evaluate MT systems that are used for document translation experiments (Section 6.3), since the objective goal is to translate long sentences (documents) rather than short queries. The development set contains 500 sentence summaries used for tuning.

For the final MT evaluation, we use 1000 sentence summaries from the Khresmoi test set. We also use the Himl test set for evaluation,[6] which contains two subsets: the *NHS* set containing 1044 summaries that are taken from the NHS 24 website,[7] and the *Cochrane* set with 467 summaries from the Cochrane website.[8] Both sets are available in English and the seven languages. Samples of these two sets are shown in Table 5.3 and 5.4 respectively.

Table 5.5 shows statistics of both development and test sets including the number of sentences in each set and their source.

5.2.4 Data Selection

Naturally, data in machine learning often contains noise, and machine translation is no exception. Such noise can be low quality translations and sentences mixed with words in multiple languages. In addition to that in our case, the available data came from different domains not only from the medical domain, thus, translations of some sentences in the medical domain might be different in in other domains. For example, the word *development* in the medical domain means the growth or the spread of a disease or a tumor. However, we can not tell what this word means in a general domain without a context.

[5] http://khresmoi.eu/

[6] http://www.himl.eu/test-sets

[7] https://www.nhs24.scot

[8] http://www.cochrane.org

Language	Summary
English (Ref.)	*wound healing and treatments for people with diabetic foot ulcers*
Czech	*hojení ran a léčba u lidí s diabetickou nohou*
German	*wundheilung und behandlungen von menschen mit diabetischen fußgeschwüren*
French	*la cicatrisation des plaies et le traitement pour les personnes souffrant d'ulcères du pied diabétique*
Hungarian	*sebgyógyítás és kezelések diabéteszes lábszár fekélytől szenvedő betegek számára*
Polish	*gojenie ran i leczenie u ludzi ze stopą cukrzycową*
Spanish	*la curación de las heridas y los tratamientos para las personas con úlceras de pie diabético*
Swedish	*sårläkning och behandling för personer med diabetiska fotsår*

Table 5.2: Samples of sentence summaries in English and their manual translations in the seven languages taken from the Khresmoi summary tuning set

Language	Summary
English (Ref.)	*tests and treatments*
Czech	*vyšetření a zákroky*
German	*Tests und Behandlungen*
French	*les tests et les traitements*
Hungarian	*vizsgálatok és kezelések*
Polish	*badania i leczenie*
Spanish	*pruebas y tratamientos*
Swedish	*tester och behandlingar*

Table 5.3: Samples of sentence summaries in English and their manual translations in the seven languages taken from the NHS set

Language	**Summary**
English (Ref.)	*antithrombin also reduces inflammation in the human body*
Czech	*antitrombin také snižuje zánět v lidském těle*
German	*antithrombin reduziert ebenso entzündungen im menschlichen körper*
French	*l'antithrombine réduit également l'inflammation dans le corps humain*
Hungarian	*az antitrombin emellett csökkenti a gyulladást a szervezetben*
Polish	*antytrombina zmniejsza również stany zapalne w organizmie człowieka*
Spanish	*la antitrombina también reduce la inflamación en el cuerpo humano*
Swedish	*antitrombin reducerar också inflammation i människokroppen*

Table 5.4: Samples of sentence summaries in English and their manual translations in the seven languages taken from the Cochrane set

set	Khresmoi	Cochrane	NHS
dev	500	-	-
test	1000	467	1044

Table 5.5: Statistics of development and test sets for MT evaluation (number of sentences)

Data selection is considered to be a method for MT domain adaptation. The goal of data selection is to maintain a high quality of parallel data for MT training. Data in this context is high quality when a sentence in the target language represents a good translation for its target sentence in both semantic and syntactic perspectives. Data selection showed to improve the performance of the translation significantly [Koehn and Schroeder, 2007]. The common practice for data selection is to filter out out-domain data (data that does not belong to the domain in which the model will be used) and keep in-domain data (data that belongs to a specific domain) for training or tuning of the MT model.

We follow the work of Moore and Lewis [2010] to apply data selection method on the entire available data, wherein we train two language models, one trained using in-domain data, and the second one is trained on general-domain data. In order to choose the data that is used to train the in-domain language model, we select only sentences that have at least two entries in dictionaries that are created from medical multi-lingual resources such as UMLS. Then each sentence (from all available data in *in-domain* and *general-domain*) is scored by the difference of its cross-perplexity from both language models. Sentences which have low final scores are chosen for training the monolingual language model for SMT as we will show later in this section.

For parallel data, each side of a parallel sentence (source and target) is scored separately, and then the final score of that sentence is the average score of those two scores. Filtered sentences in parallel data are limited to the best 10 million sentences in each language pair in order to reduce the size of the model. This data is used to train both SMT and NMT models.

5.2.5 Data Preprocessing

We clean and process MT data as follows:

Cleaning: we clean all the data by removing non-UTF8 characters from the text, we also remove sentences that are longer than 80 words because these sentences make the training very slow.

NMT models showed to be more sensitive to noisy data than SMT models [Popel and Bojar, 2018b]. We present a simple yet effective way to clean the noise in the presented parallel data when the sentences either contain untranslated words or are swapped, for example, English-German pairs in the parallel data were in German-English order, though the translations were correct.

To filter out these sentences, we loop over each sentence in each language pair, if a sentence in the source side contains at least one stop-words in the target language, we remove that sentence from both sides. We use stop-words lists provided by Ranks NL.[9] Some stop-words lists in two languages share similar words (with different meaning in each language); such as the preposition word *to* in English and the demonstrative pronoun *to* in Czech. We discard the words that exist in the intersection set of the two stop-word lists to avoid removing correctly aligned and clean sentences. We apply this cleaning method to the parallel data in all languages but only for the purpose of developing NMT models. This cleaning method removes around 5% to 10% from the data, depending on the language.

Lemmatization: In order to train an SMT system that produces lemmas (as we present in Section 5.5.1), we lemmatize the training data (only the target side of the corpus) using UDPipe, which is a pipeline that supports various morphological analysing tasks (such as tokenization, tagging and lemmatization) of a raw text. It also provides trained models for most of the universal dependency treebanks [Straka and Straková, 2017], including all the languages in our work.

It is known when doing lemmatization that the number of words in the input sentence might be different from the lemmatized one. This happens because some words in the input are multi-word. For example, the word *im* in German is originally two words (*in dem*). This might cause inconsistency in the alignment table because we apply alignment on forms and then replace the forms with the lemmatized version of the data (more details in Section 5.5.1). To avoid such a case, we first apply the lemmatization on the parallel data, then take from the output of the lemmatizer both forms and lemmas, and then we replace the output forms with the original ones, and keep lemmas to be replaced later after the alignment. For example, if the input sentence (forms) is “w_1, w_2, w_3, w_4“; the lemmatizer produces, in column format, both forms and lemmas, and it is likely to have a new token in both forms, such as: “$w_1^f, w_2^f, w_3^f, w_4^f, w_5^f$“, and lemmas as in: “$w_1^{lem}, w_2^{lem}, w_3^{lem}, w_4^{lem}, w_5^{lem}$“. The forms from the lemmatizer's output are used instead of the original forms. This is done only on the parallel data (the target side), since the alignment is only done between the source and the target side of the parallel sentences.

Tokenization is the process of splitting sentences into individual words including punctuation marks which are usually written without space after the previous words. All the data is tokenized using the Moses tokenizer.[10]

[9] https://www.ranks.nl/stopwords

[10] http://statmt.org/moses/

5.3 Neural Machine Translation

In this chapter, we first present related work on NMT, and we describe the architecture of the state-of-the-art methods. Then, we present our training method to train the Transformer NMT model to translate queries and documents for CLIR. We follow the iterative back-translation approach in NMT. For model selection, we present a novel method that predicts which model among the intermediate training models is ideal for the QT approach.

5.3.1 Introduction

NMT has recently achieved superior results in the task of MT, and led to a significant improvement over SMT systems. The success of word embedding and deep learning methods were the main reasons that boosted NMT models, also the raise of enhancement of parallel computing, which was achieved by the recent development of Graphics Processing Units (GPU), helped make developing NMT models much easier than ever. Beside the outstanding MT performance, NMT is considered to be simpler than SMT in terms of the internal system components. In SMT, we need to build multiple components that are integrated together in the entire MT system; such as language model and translation model (as we showed in Section 5.1). However, it is not the case in NMT, where only one model is needed to implement an MT system.

5.3.2 Sequence-2-Sequence Model

The sequence-2-Sequence model consists of two Recurrent Neural Network (RNN) networks, the goal is to predict the conditional probability of the output sequence $y = (y_1, ..., y_{T'})$ in the target language, given the input sequence $x = (x_1, ..., x_T)$ in the source language, and taking into consideration that the output length T' might be different from the input length T. The encoder (Long Short-Term Memory (LSTM) network) reads the input sequence in a reverse order (the last word in a sentence is the first input token to the encoder) and then computes a fixed annotation vector v for that sequence.

$$p(y_t, .., y_{T'}|x_1, .., x_T) = \prod_{t=1}^{T'} p(y_t|v, y_1, .., y_{t-1}) \tag{5.3}$$

The decoder performs language modeling using LSTM formulation as shown in Equation 5.3, and given the annotation vector v from the encoder LSTM.

Attention Mechanism

The main challenge in sequence-to-sequence approach is using an annotation vector w with a fixed length. This vector might fail to hold all the contextual information that is needed to translate long sentences. To remedy this issue, the attention extension of the encoder-decoder model was introduced by Bahdanau et al. [2015]. The attention technique enables the decoder to take advantage of the complete tokens in the input and extracts the needed information for decoding. In the attention sequence-to-sequence model, the annotation vector v is replaced by a context vector c_i. This context vector is the sum of the annotations h_j for the hidden vector s_{i-1}, and the weight of the annotation h_j is computed using an alignment model a_{ij} which is the probability that the token y_i is aligned to the input word x_j. The alignment model is built using a single-layer feedforward neural network.

Each token in the output sequence y_i is predicted based on a recurrent hidden state s_i, the previously predicted word y_{i-1}, and a context vector c_i . Mapping the input sequence into an annotation vector makes the model able to produce number of tokens that is not necessarily the same as the number of input tokens. The reverse order of the input tokens helps the LTSM to perform well when dealing with long sentences. The reason for that is to make the optimisation of the stochastic gradient descent simpler, which is different from the gradient descent algorithm (an optimization algorithm usually used to find a local minimum of a given function) by choosing randomly a data point (called a batch in this context) from the training data.

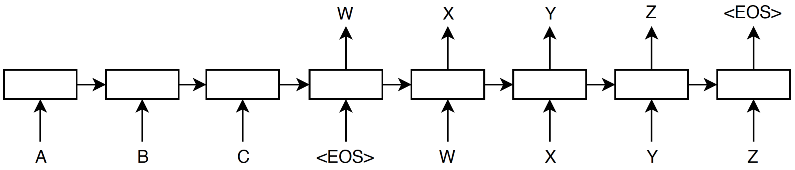


Figure 5.2: The encoder maps the input sequence of tokens (A,B,C) to an annotation vector W, and the decoder starts producing the output sequence of tokens (X, Y,Z) after reading end-of-sentence (<EOS>) token, and stops prediction after producing <EOS> token. Input is in a reverse order [Sutskever et al., 2014].

The Transformer Model Transformer is based on the encoding-decoding model, and it achieves state-of-the-art performance in the machine translation task Vaswani et al. [2017], which replaces the RNN layers with self-attention layers.

The input is firstly embedded into multi-dimensional space vector and then a positional encoding is applied to those embeddings. The positional encoding generates a new presentation for a given word embeddings considering that word's position in the original sentence.

Both encoder and decoder have two mutli-head attention layers as shown in Figure 5.3. The multi-head self-attention computes multiple attention blocks from the source input and linearly combines them onto a space with initial dimensions. The special feature of multi-head attention is that it combines information from different seen states into one vector. Attention (as in Equation 5.6) is a function (referred as *MultiHead*) that maps three vectors: queries (Q), key (K) and value (V) pairs, and outputs the weighted sum of these values using softmax function.

$$MultiHead(Q, K, V) = Concat(head_1, head_2, .., head_h)W^o, \tag{5.4}$$

$$\text{where } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \tag{5.5}$$

$$\text{and } Attention(Q, K, V) = softmax(\frac{QK^K}{\sqrt{d_k}})V \tag{5.6}$$

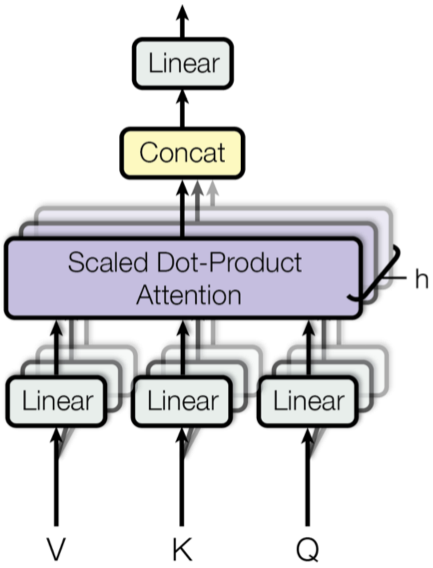


Figure 5.3: The multi-head attention in the Transformer model [Vaswani et al., 2017]

5.3.3 Byte-Pair Encoding

Machine translation is an open-vocabulary problem, which means that an MT system has to translate any word during test, even if this word did not appear in the training data.

The main challenge in NMT model is that it deals with a fixed vocabulary size of both source and target language. This is usually around tens of thousands of words depending on the training hyper-parameter. The most frequent N words (usually between 20 000-80 000 words) in the source and the target language are used to create the vocabulary for each side. Because using a large vocabulary is very challenging in terms of time and space complexity. When an out-of-vocabulary word appears in the data, it is either left untranslated in the output, or replaced with a special token (usually *<UNK>* token).

The reason behind using fixed vocabulary size by encoding an input sentence into a vector of a fixed length is because softmax function is computationally expensive. Thus, NMT model deals with the most frequent vocabularies in each side of a language pair [Luong et al., 2015]. The approach of using fixed vocabulary showed to be effective when there are a few unknown words in the target sentences, but the performance degrades significantly in case too many unknown words are observed [Bahdanau et al., 2015].

Jean et al. [2015] tackled this issue by using an extensive vocabulary of the target language and apply sampling of the vocabulary based on the importance of the vocabulary to solve the issue of the time complexity. They reported an improvement in English-French and English-German models compared to those which used small vocabulary or replaced the unknown words with the most likely translation that are selected by a pre-trained alignment model (back-off models) [Luong et al., 2015].

Sennrich et al. [2016b] employed the Byte-Pair Encoding (BPE) compression algorithm [Gage, 1994] by encoding all the words in the training corpus using a small set of vocabularies. During training, their bottom-up character merging algorithm computes frequencies of all symbols (or characters) in the corpus and then applies n merges (a hyperparameter of the experiment) of these symbols iteratively by choosing the most frequent pair of symbols, and applies merge of these symbols by using a special character that indicates the merge positions, and finally adds the merged symbols into the vocabulary set. Then to apply BPE, first, they get all bigrams, and then apply the merge on a symbol of pair that appeared first in the merges. This helps NMT models deal with unseen words as subwords units, and showed to be more effective than back-off models. The authors also showed that NMT models usually perform poorly with rare words, but BPE segmentation helps to translate such words effectively.

5.3.4 Back Translation

The NMT approach is known to be more greedy for training data than the SMT approach. Koehn and Knowles [2017] reported that for a low amount of data, SMT significantly outperforms NMT. However, when they gradually added more data to both models, NMT performance kept improving steadily until it outperformed SMT models, as shown in Figure 5.4 for the English-Spanish pair.

The NMT greediness of training data has been a main motivation for researchers to look for a way to enrich the training data and make use of what has been available. The use of monolingual data has shown to be very effective in the SMT approaches by giving the model the ability to prefer translations that are more fluent and coherent [Koehn, 2009], and NMT is no exception, multiple works succeeded in employing it in NMT approaches.

He et al. [2016] improved the performance of NMT model (English-Chinese) by around 2.33 of BLEU score when they log-linearly integrated SMT features (translation model and language model) within the NMT model.

Gulcehre et al. [2017] proposed two ways to make use of monolingual data in NMT models. The first one is a neural language model that is combined in the hidden state of the NMT model (deep fusion), and the second one is a language model that scores candidate translations of the NMT decoder (shallow fusion). They showed that deep fusion model significantly improved the translation performance on low-resource language (Turkish-English) and high-resource languages (German-English and Czech-English).

Sennrich et al. [2016a] investigated the use of monolingual data in the target language to improve the translation quality of NMT. They translated monolingual data in the target language into the source language, and added the translations as parallel data into the training data. They called this data that is generated using back translation as *synthetic* data. They also showed that this approach can be used for domain adaptation when there is no authentic domain-specific parallel data available for training.

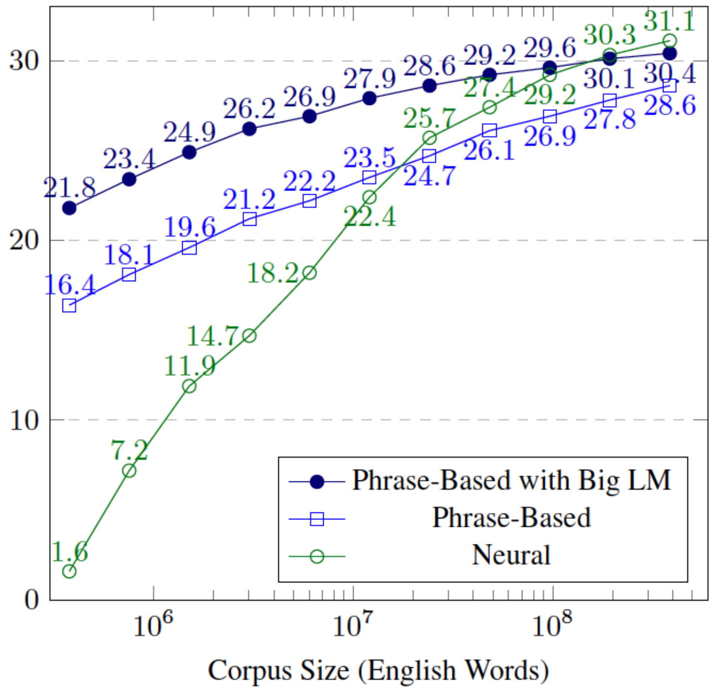


Figure 5.4: The affect of training corpus size on BLEU (English-Spanish pair) [Koehn and Knowles, 2017]

Przystupa and Abdul-Mageed [2019] studied the effect of the amount of monolingual data that can be used in back translation. They confirmed that back translation can be useful to improve NMT models for low-resourced languages, furthermore, they showed that the optimal amount of synthetic data depends on the amount of the original authentic bilingual data. However, when having a good amount of authentic bilingual data in hand, around $20-30\%$ should help to boost the performance of the original model.

Hoang et al. [2018] showed that the quality of back-translated data could significantly improve the translation system. They improved the back-translation process by a presented method called iterative back translation. This is done by implementing another NMT system in the opposite direction of the main system (target to source language). At this point they have two NMT systems, each one of them is used to generate synthetic data that is used to continue training of the system in the other direction. Each time a back-translation is done, both systems improve; thus the quality of the synthetic data that is generated from these two systems will improve too. Repeating this process multiple times leads to a remarkable improvement of the final model.

5.4 MT Evaluation

The purpose of MT evaluation is to measure the effectiveness of the system, and to optimise its performance. It can be either human evaluation or automatic evaluation. Human evaluation needs native speakers in the source and the target languages, and experts in the domain if the task is domain-specific such as the medical domain. Those native speakers are asked to look at the output of an MT system and give a graded score determining its correctness. Mainly, they are required to judge the output in terms of fluency and adequacy. Fluency means how much the output sentence in the target language is fluent ignoring its content or the information it represents. While adequacy ignores the output fluency, and tells if the information that is represented in the input source sentence is present in the target translation sentence and coherent [Snover et al., 2009].

The process of MT development involves an intensive number of sub-experiments, such as data selection and tuning various hyper-parameters for the chosen model. A combination of such different parameters might lead to hundreds or maybe thousands of translation outputs from each model. These translations need to be evaluated in terms of their quality; thus, we can choose the MT model that produces the best translation quality. Manual evaluation for these models can be an expensive process in terms of time and money.

This raises the importance of developing an automatic evaluation method that can measure the performance of a system in near real-time. In order to conduct an automatic evaluation in MT, we need a set of sentences (usually written by domain experts and reviewed by linguists) in the source language that is large and diverse enough in terms of sentence lengths and their vocabularies. Then, human translation experts translate these sentences into the target language to create reference translations (ground truth). Finally, a set of translation hypotheses (candidates) are generated by an MT system which needs to be evaluated, These candidates are used as input together with the reference translation for an automatic evaluation metric. The automatic metric is expected to generate a number that represents the quality of translation of the evaluated MT system. We will consider in this work two metrics, BLEU and Position-independent word Error Rate (PER).

5.4.1 BLEU

BLEU is one of the most popular automatic evaluation metrics in MT; its values lie between zero and one; a higher BLEU score indicates a better MT performance. It is based on measuring similarity of n-gram counts between a translation hypothesis and its reference translation (usually up to $N = 4$) [Papineni et al., 2002].

For a translation hypothesis c and a reference translation r, the clipped count of n-gram g is calculated as shown in Equation 5.7, where $count(g, s)$ refers to the number of n-gram g appears in a sentence s, and precision p_n is then calculated as in Equation 5.8. BLEU is calculated as shown in Equation 5.9, given the Brevity Penalty (BP) as in Equation 5.10, and n-gram weight w_n, usually n-gram counts have uniform weights $w_n = 1$. BP is used to penalise the short sentences, where lc and lr are the length of translation hypothesis and reference respectively.

$$count_{clip}(g, c, r) = min\left(count(g, c), max_{i \in I}(count(g, r_i))\right) \tag{5.7}$$

$$p_n = \frac{\sum_{g \in \text{n-grams}(c)} count_{clip}(g, c, r)}{\sum_{g \in \text{n-grams}(c)} count(g, c)} \tag{5.8}$$

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{5.9}$$

$$BP = \begin{cases} 1 & \text{if } lc > lr \\ e^{(1-lr/lc)} & \text{if } lc \leq lr \end{cases} \tag{5.10}$$

MT systems are usually employed in CLIR either to translate the queries into the collection language, or to translate the collection into the query language. During the process of developing such an MT system, tuning its parameters using one of these automatic metrics is needed. Tuning towards BLEU is not preferred in developing MT systems for query translation approaches, because most IR systems use term-matching models based on bag-of-words approach, which ignores word order, and queries tend to be short and written in free word order form.

However, it is not the case when following the document translation approach, because documents usually consist of complete sentences, thus, BLEU is preferred in this case. An interested reader is referred to the work of Pecina et al. [2014].

5.4.2 PER

PER on the other hand, does not penalise word order between a hypothesis and its reference translation as BLEU does, instead, it measures the difference of the word counts that appear in both. PER captures all words that appear in the translation hypothesis but do not exist in the reference. These words are known as PER errors; thus, a higher PER value indicates more errors and lower MT performance.

PER is calculated as shown in Equation 5.11, where $d_{PER}(ref_k, hyp_k)$ (for each token k in the reference translation ref) is given in Equation 5.12, the reference translation is ref_k and translation hypothesis is hyp_k, $n(e, hyp_k)$ and $n(e, ref_k)$ are the counts of a word e in hyp_k and ref_k respectively, as shown in Equation 5.12

$$PER = \frac{1}{N^*_{ref}} \sum_{k=1}^{N^*_{ref}} \min \mathrm{d}_{PER}(ref_k, hyp_k) \tag{5.11}$$

$$d_{PER}(ref_k, hyp_k) = \frac{1}{2}\left(|N_{ref_k} - N_{hyp_k}| + \sum_{e} |n(e, ref_k) - n(e, hyp_k)|\right) \tag{5.12}$$

In addition to the fact that queries are short and usually do not formulate syntactically correct sentences, most IR models (including Dirichlet-smoothing IR model that is used in our work) treat queries and documents as bag-of-words;

hence word order is not important, which makes PER ideal for tuning MT models. For these reasons, we focus on PER for MT tuning.

5.4.3 Training with MERT

MERT is the most popular tuning algorithm for SMT model parameters. The main objective of MERT is to optimise the parameters (feature weights) λ_i, towards a better translation performance by minimising an error rate [Och, 2003].

The optimisation step in MERT is done by generating a list of alternative translations (n-best-list) in the target language for each sentence in the target language. Each translation of this list has a set of feature values (such as language model score, translation model score). MERT grid-searches feature weights until it finds a combination of these weights that gives the best evaluation metric. BLEU or PER can be as the objective function of MERT.

5.5 MT Training

In the following sections we present our training methods of mainly two MT systems: SMT and NMT. We employ these systems for the query translation approaches and the document translation approaches in CLIR. Thus, we consider CLIR performance (P@10 metric) as the final objective function of our MT training methods.

5.5.1 SMT Training

In this section, we present the process of developing SMT models for our CLIR experiments. It is important to mention that the methods, which are applied in this section, are adopted from the work of Dušek et al. [2014]. Our main contribution in this work is: 1) Replicating their work and evaluating the results on the CLIR task and 2) Developing an SMT system that produces lemmatized sentences as we will show later.

We build two systems, the first MT systems are tuned to translate queries rather than long sentences. These systems are used in our CLIR methods that follow the query translation approach.

The second systems are tuned to translate normal sentences, which makes the systems ideal for the CLIR document translation based approaches. The main difference between the two systems lays in the parameter tuning part and the development data set, as we will show later in this section.

All the SMT experiments are done using eman [Bojar and Tamchyna, 2013], which is an open source tool for experiment planning and automation. Eman can be used to design any kind of experiments that include different dependent and independent steps. However, it is optimised for SMT experimenting, and integrates tools like Moses [Koehn et al., 2007] and GIZA++ [Och and Ney, 2003], in addition to a set of tools for MT evaluation and text preprocessing.

SMT for Query Translation (QT-SMT-form)

The goal of this approach is to design an SMT system that translates medical queries into the language of the document collection (English in our work). We refer to this MT system in the following text as *QT-SMT-form*. The system is based on Moses [Koehn et al., 2007], an open source tool that supports training and evaluation of SMT systems. For word alignment, we use *fast_align* [Dyer et al., 2013] on the lowercased and tokenized data.

For training language models, we use SRILM (Stanford Research Institute Language Modeling toolkit) [Stolcke, 2002] with order of 5 (5-gram). The language models are trained using the monolingual data that we presented in Section 5.2 for each of the target languages. And finally MERT is employed to tune the model parameters using the development data set towards PER.

SMT for Document Translation

For the DT experiments, we train two SMT systems:

DT-SMT-form
This system is a replication of the SMT systems that translate standard sentences by Dušek et al. [2014]. This system is identical to the previous SMT system that we presented for query translation, with only one difference that it is tuned with MERT using complete sentences (summaries) towards BLEU (rather than using queries towards PER). The rest of the training steps are the same.

DT-SMT-lemma
The document collection in this work is in English, which is not morphologically rich as some target languages of the queries (e.g. Czech, German and Swedish). According to Schultz et al. [2002], morphological variations of terms in information retrieval cause degradation of recall performance, because IR systems that are based on term-matching fail to match terms in search query with their morphological variations in document collection.

To avoid the influence of the morphological variations of the document collection when translating it from English into a morphologically rich language, we attempt to reduce the morphological variations in the output translations by introducing *DT-SMT-lemma* as our own modification of *DT-SMT-form*, which translates English sentences into lemmatized sentences in the target language using UDPipe [Straka and Straková, 2017]. The only difference between these two systems is that in *DT-SMT-lemma*, we lemmatize the monolingual data and the target language part of the parallel data.

We use *fast_align* [Dyer et al., 2013] to compute word alignment on the lowercased word forms between English and each target language. Then we replace the word forms in the target language with word lemmas. Moses (with its default settings) is used to train a phrase-table using the tokenized and lowercased English word forms and the tokenized and lemmatized data in the target language. The evaluation of the presented SMT systems is shown in Section 5.6.

5.5.2 NMT Training

In this section, we describe our training approach for the NMT models that we use for CLIR experiments (both the query translation and the document translation approach). The main purpose is to compare the performance of CLIR systems when using SMT versus NMT. The training objective of our NMT systems is to maximise the performance of the retrieval rather than the quality of translations as it is done in a typical MT task.

Our NMT experiments are conducted using the Marian library, an open source tool that is written in C++ and supports multi GPU training and translation. Marian also includes implementation of the state-of-the-art NMT models such as deep RNN and Transformer model [Junczys-Dowmunt et al., 2018].

As for the training, dev and test data, we use the same data as in SMT training to make the comparison as fair as possible. However in NMT, all data sets (monolingual and parallel) are encoded by BPE using subword units segmentation script by Sennrich et al. [2016b]. The model parameters (for example: learning rate, beam-size, dropout etc.) are the same for all language pairs, and identical the parameters that were reported by Vaswani et al. [2017].[11]

We follow the iterative back-translation approach as suggested by Hoang et al. [2018] and presented in Section 5.3.4. In addition to the fact that back transaltion training boosts the performance of the NMT approach, it is shown that it helps for domain adaption of NMT when the monolingual data is taken from a specific domain. Which is ideal for our case, since we translate medical text (medical domain).

Source of the monolingual data
The monolingual data is in English for the pairs that have English as a target language, and in one of the seven languages (e.g. Czech) for the another translation direction (e.g. English -> Czech). It is recommended to be taken from medical-domain data to adapt the model more for medical text translation (Section 5.2.4); thus we select data from the following two resources:

- For English monolingual data, we use the retrieval collection, which is entirely medical-domain data. The collection is described in details in Section 4.1. We use the seven initial models (English to target) to translate it into each language (Czech, French, German, Hungarian, Polish, Spanish and Swedish).

- For non-English monolingual data, we use the target side of the parallel corpus. The parallel corpus is taken from the medical domain as we presented in Section 5.2.

In both cases and in each back-translation iteration, we randomly select 2 million sentences from the monolingual data, ignoring sentences that are longer than 80 words.

[11] `https://github.com/marian-nmt/marian-examples/tree/master/transformer`

Figure 5.5 shows the architecture of the proposed iterative back-translation NMT model. It includes the following three steps:

- **Initial models:** For each language pair, we first train the initial models in both directions, English to target, and source to English. We use the authentic parallel data that we presented in Section 5.2 for training the initial models. During the training of the transformer models, multiple epochs (iterations through the entire training data) are needed. It is known that too many training epochs can cause over-fitting of the model, and too few iterations might cause under-fitting [Popel and Bojar, 2018a]. To avoid this, early stopping of the training is employed to terminate the process when the intermediate model satisfies a stopping criterion (training objective).

  We stop training when there are 3 consecutive checkpoints without any improvement in the translation performance of the tuning data . The checkpoint in Marian is when Marian saves the trained model at that point (as an intermediate model), translates the tuning data using that model and evaluates its performance. This information is stored in a log text file, which can be used to monitor the training process.

- **Translating monolingual data:** When the training in the previous step finishes, we take the last saved model and use it to translate the monolingual data in the target language into the source language. Then, we add the translated data into the authentic one and shuffle all the data. In every back translation step, we add the translated data into the authentic one rather than accumulating the translated data in each translation process. This guarantees a fixed ration of authentic data (80%) in the synthetic one.

- **Updating initial models:** At this step, we continue training using the last saved models and using the new synthetic data.

We applied back translation three times during the developing of our NMT model. The improvement in the third one was not substantial, however, we applied a fourth back translation on the Czech-English pair and the performance was not different from the third one; so we decided to stop at this point.

NMT Model Selection

We setup Marian to save the intermediate models (checkpoints) after each 5000 iterations, where each iteration is a batch-sized instances from the training data. This is done instead of saving each epoch to avoid loosing effective intermediate models in between. The model is evaluated at each checkpoint using two MT metrics: BLEU and PER on the MT development set.

We employ MT metrics to select the best model for CLIR rather than employing IR metrics for this purpose. Figure 5.6 shows the evaluation of the intermediate models using MT metrics and how they correlate with P@10 (IR metric). P@10 is calculated by query translation of the Czech training queries into English using the corresponding NMT model, and then conducting retrieval using the baseline system as we describe in Section 6.2.1.

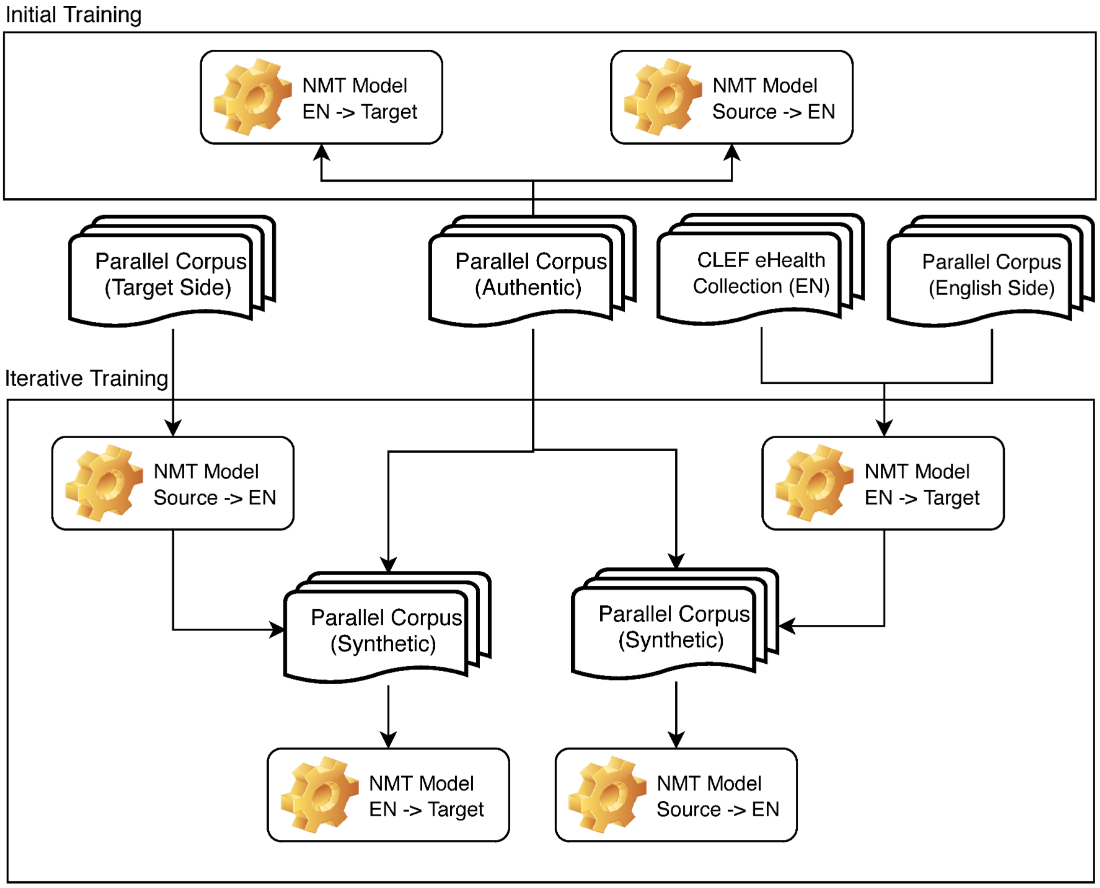

Initial Training
NMT Model EN -> Target
NMT Model Source -> EN
Parallel Corpus (Target Side)
Parallel Corpus (Authentic)
CLEF eHealth Collection (EN)
Parallel Corpus (English Side)
Iterative Training
NMT Model Source -> EN
NMT Model EN -> Target
Parallel Corpus (Synthetic)
Parallel Corpus (Synthetic)
NMT Model EN -> Target
NMT Model Source -> EN

Figure 5.5: Iterative back-translation architecture

The best BLEU score (iteration 400000) does not correspond to best value for P@10, nor the best score for PER (500000). This is understandable, because these metrics evaluate the quality of the translation.

In order to select the best checkpoint that guarantees the advantages of both metrics (BLEU which penalizes word order, and PER which does not), we ensemble the two models together (best BLEU and best PER) during decoding by setting up the weights for both models equally to 1.0. Marian decoder supports model ensembling since they share the same vocabularies. For the document translation experiments, we select the NMT models with the best BLEU scores. We plot only Czech-English system for the sake of simplicity.

5.6 MT Results Evaluation

We present our evaluation of the MT systems that we presented previously in this chapter. The purpose of our MT development is to use translation of either search queries or document collection in CLIR. For that reason, the performance of CLIR system (after employing MT) is considered the objective of our MT development.

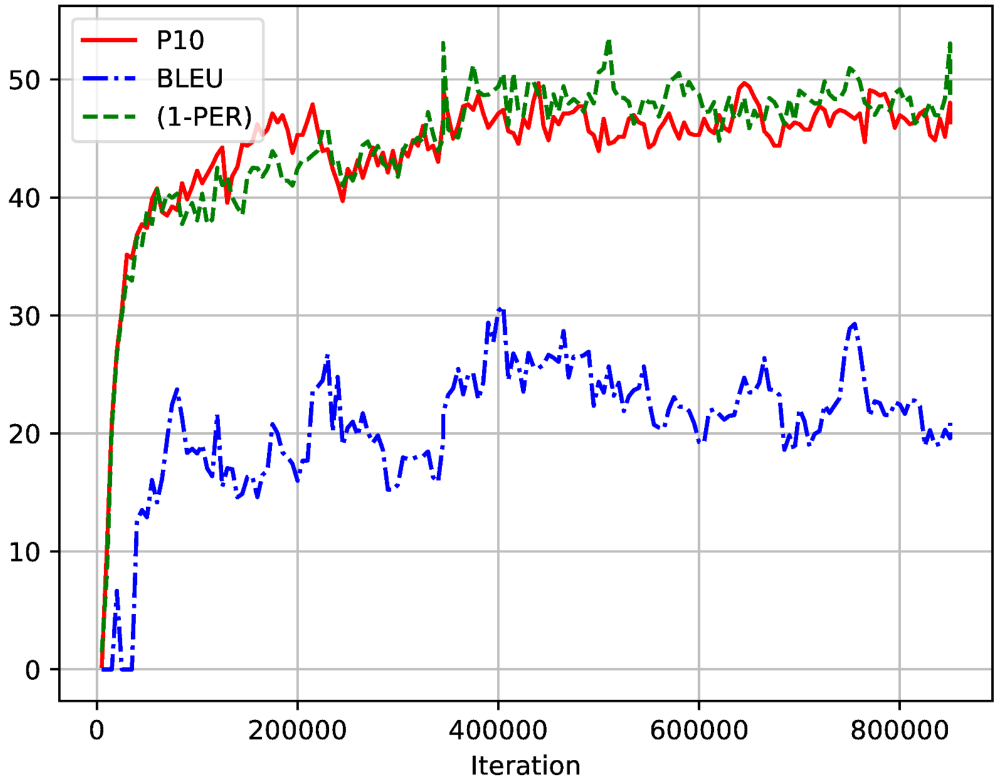


Figure 5.6: Performance comparison of intermediate NMT models at each checkpoint (each checkpoint is 5000 iterations) in terms of BLEU, 1-PER and P@10 (in percentages), for Czech→English MT model and the corresponding Czech (QT) CLIR system

5.6.1 MT for Document Translation

For MT systems that are used in document translation experiments, we report the following models:

- **DT-SMT-form**: This system is a replication of the SMT systems that translate standard sentences by Dušek et al. [2014].

- **DT-SMT-post-lem** In this system, we perform post-translation lemmatization of the output from **DT-SMT-form** using UDPipe.[12] The purpose of this system is to study the effect of the lemmatizated sentences on the MT evaluation metrics. Reference sets are also lemmatized.

- **DT-SMT-pre-lem** The only difference between this systems and **DT-SMT-form** is that here we lemmatize the monolingual data and the target side of the parallel data. Then, we translate the test sets, in this case the produced translations are already lemmatized, then we evaluate them with the lemmatized reference translations.

- **DT-NMT-form** This system translates the test sets using the NMT model that we described in 5.5.2.

Table 5.6 shows the results of our evaluation of MT for sentence translation systems using the three test sets (Khresmoi summary, Cochrane and NHS).

[12] http://ufal.mff.cuni.cz/udpipe

The scores cannot be directly compared across languages and for the *form* and *lemma* systems (since the test sets differ) but they indicate how the translated sentences differ from the reference translations which in term-matching IR is important. Also, the results of the two systems producing lemmas instead of the forms are indicative only. They cannot be directly compared to those producing forms. In all language pairs (except English-Hungarian), DT-SMT-post-lem (lemmatizing the output of the SMT systems) achieves the best results in terms of BLEU scores for the Khresmoi summary test set.

Pair	System/Set	Khresmoi		Cochrane		NHS	
		BLEU	PER	BLEU	PER	BLEU	PER
EN-CS	DT-SMT-form	19.06	51.16	17.63	49.66	12.35	41.81
	DT-SMT-post-lem	30.97	65.64	28.43	64.51	20.41	54.48
	DT-SMT-pre-lem	28.72	64.26	25.04	60.91	18.99	53.56
	DT-NMT-form	25.92	56.57	24.56	54.62	14.64	44.04
EN-FR	DT-SMT-form	37.85	68.30	31.18	62.39	27.94	57.31
	DT-SMT-post-lem	43.50	74.78	41.62	71.50	36.60	65.60
	DT-SMT-pre-lem	41.26	72.62	31.52	63.54	26.67	58.20
	DT-NMT-form	38.84	66.52	29.36	59.90	18.46	48.04
EN-DE	DT-SMT-form	18.79	53.42	22.19	55.89	16.78	49.78
	DT-SMT-post-lem	23.68	60.46	27.78	62.76	19.76	55.05
	DT-SMT-pre-lem	13.08	48.01	15.00	50.61	11.66	47.40
	DT-NMT-form	19.83	51.45	20.69	50.96	9.77	36.43
EN-ES	DT-SMT-form	25.77	63.29	33.79	67.10	24.83	59.67
	DT-SMT-post-lem	35.46	72.31	44.41	75.85	36.19	69.60
	DT-SMT-pre-lem	28.41	65.73	36.06	70.05	26.96	62.10
	DT-NMT-form	23.20	55.25	26.24	56.16	16.54	43.08
EN-HU	DT-SMT-form	10.51	41.68	8.62	36.15	6.85	27.85
	DT-SMT-post-lem	13.24	48.60	11.53	44.03	10.04	35.17
	DT-SMT-pre-lem	14.35	51.96	12.10	46.32	11.00	39.96
	DT-NMT-form	8.23	39.58	7.83	39.01	6.00	35.29
EN-PL	DT-SMT-form	11.56	41.34	15.84	44.66	11.53	36.70
	DT-SMT-post-lem	16.19	50.59	21.84	53.60	16.36	44.12
	DT-SMT-pre-lem	12.52	46.90	18.06	52.04	14.38	44.78
	DT-NMT-form	10.23	35.91	13.65	41.58	9.86	35.50
EN-SV	DT-SMT-form	33.69	64.69	34.16	64.38	30.58	62.48
	DT-SMT-post-lem	40.93	69.93	41.92	70.51	37.93	68.24
	DT-SMT-pre-lem	39.16	70.07	39.39	70.61	36.41	68.48
	DT-NMT-form	35.18	64.49	38.84	67.63	34.52	62.51

Table 5.6: Results of DT MT system evaluation in percentages, DT-SMT-form, lemmatization of translated forms (DT-SMT-post-lem), translations of DT-SMT-pre-lem, and DT using NMT system. These MT systems are used for the document translation CLIR experiments

5.6.2 MT for Query Translation

For the query-translation experiments, we present two systems:

- **QT-SMT-form** This system is described in Section 5.5.1 which follows the work of Dušek et al. [2014]. The model is optimised to translate medical search queries by tuning its parameters using the Khresmoi query test set, and employing MERT to tune its parameters towards PER rather than the common approach in MT which is tuning towards BLEU metric.

- **QT-NMT-form** This system follows our proposed approach in Section 5.5.2, wherein the NMT model is an ensembled model of two models: the model that achieves the best BLEU and the model that achieves the best PER using the development data set.

The results of our evaluation of MT for query translation are presented in Table 5.7 and Table 5.8 using Khresmoi query test set. We can observe that SMT systems significantly outperform the NMT systems in terms of BELU score in all language pairs. However, NMT models produce the best PER scores in all pairs (except for the Swedish-English pair), although our NMT models are ensembled based on two models (best PER and best BLEU), it seems that the model with best PER has stronger weights in the output layer.

Metric/System	CS-EN		FR-EN		DE-EN	
	NMT	SMT	NMT	SMT	NMT	SMT
BLEU	22.58	**36.49**	30.63	**38.70**	28.73	**37.09**
PER	**48.96**	70.25	**65.41**	75.91	**58.10**	65.26

Table 5.7: Evaluation of translation quality of QT-SMT-form and QT-NMT-form systems against the Khresmoi query test set using BLEU and PER metrics in percentages for the pairs : CS-EN, FR-EN and DE-EN. These MT systems are used for query-translation CLIR experiments

Metric/System	HU-EN		ES-EN		SV-EN		PL-EN	
	NMT	SMT	NMT	SMT	NMT	SMT	NMT	SMT
BLEU	36.77	**39.79**	17.83	**31.22**	**40.94**	39.28	18.72	**26.06**
PER	**63.28**	67.39	**45.52**	73.73	63.06	**62.70**	**47.99**	58.64

Table 5.8: Evaluation of translation quality of QT-SMT-form and QT-NMT-form systems against the Khresmoi query test set using BLEU and PER metrics in percentages for the pairs: HU-EN, ES-EN, SV-EN and PL-EN. These MT systems are used for query-translation CLIR experiments

6. Experiments

In this chapter, we report our approaches to implement CLIR systems in the medical domain. It is organized as follows: First, we present our monolingual system. This system is considered as a reference system to other CLIR methods. Then, we move to present our CLIR methods, which mainly follow two paradigms: QT and DT. We study each method separately, and then we compare the two methods and report the state-of-the-art approach among all the presented systems. Finally in this chapter, we present our term-selection method for query expansion in CLIR and monolingual IR.

6.1 Monolingual Settings

In all CLIR methods, we reduce the task to a monolingual setting. The architecture of the monolingual system is shown in Figure 6.1. At the beginning, the English documents are indexed using Terrier [Ounis et al., 2005]. We use the embedded English tokenizer in Terrier and its default English stopword list to remove stopwords from both queries and documents. Finally, for each query, we retrieve the top 1000 ranked documents as scored by the retrieval model. The following sections present our methods for IR model selection process.

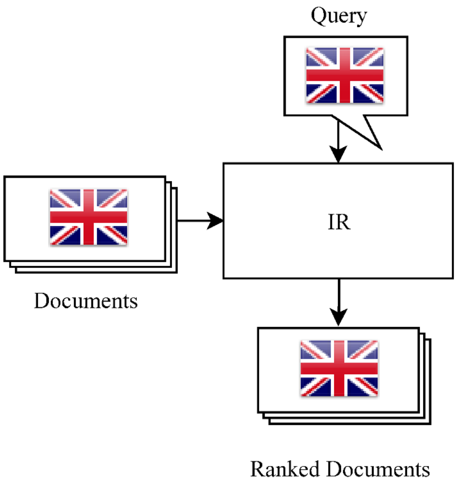


Figure 6.1: Monolingual system architecture

6.1.1 IR Model Selection

To select the best monolingual IR model for our test collection, we use the original English training queries (100 queries) to tune each model parameter and evaluate its performance using the P@10 metric, since P@10 is the main IR evaluation metric in our research. We evaluate three models: Okapi BM25, TF-IDF, and Dirichlet LM.

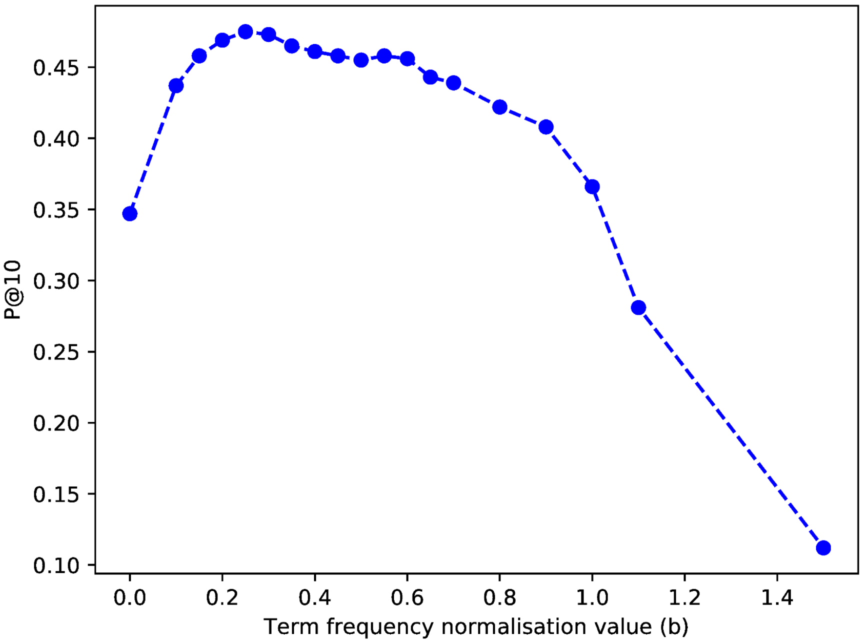


Figure 6.2: Tuning the b parameter (term frequency normalisation) using the training English queries

Okapi BM25 model was presented in Section 2.3.2. The model has a free parameter (b). We tune this parameter, and we set it to 0.25. Tuning k_1 and k_3 do not bring any significant difference over the best system that is achieved by tuning b previously, we keep the default values ($k_1 = 1.5$ and $k_3 = 1000$) as they are set in Terrier. Figure 6.2 shows how the value of b parameter affects the retrieval performance.

The second model is TF-IDF that is based on the vector space model, as explained in Section 2.3.1. The third model is Dirichlet LM, which was presented earlier in Section 2.3.3. The model has one smoothing parameter called μ, we keep this parameter set to 3000, which is its default value in Terrier, since tuning it did not bring any improvement as shown in Figure 6.3. The results of the three models on the training queries are presented in Table 6.1. Dirichlet LM achieves 50.70% of P@10, which is significantly better than the other two models. The significance testing is conducted using the Wilcoxon signed rank test, with α set to 0.05. We choose the Dirichlet LM model as the retrieval model in this research, and we consider these monolingual settings the reference system for our CLIR experiments.

Retrieval model	P@10	MAP	BPREF
DirichletLM	***50.70***	***26.14***	***37.76***
BM25	47.50	22.44	36.11
TF-IDF	44.30	22.88	36.58

Table 6.1: Results (in percentage) of our evaluation of multiple monolingual systems against the English training queries

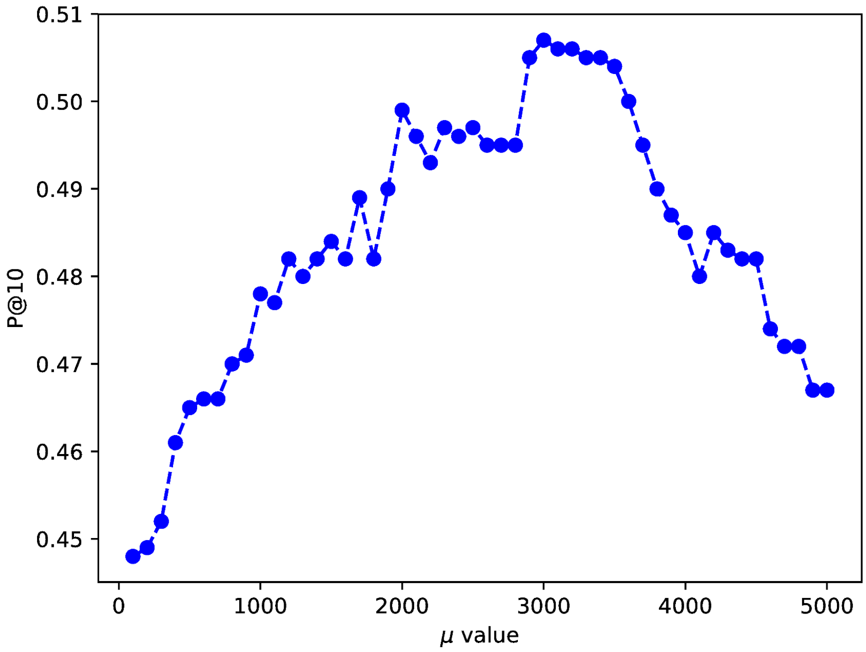


Figure 6.3: Tuning μ parameter in Dirichlet model using the training English queries

6.2 Query Translation

In this section, we present our methods of designing CLIR systems that follow the QT approach in which queries are translated into the language of the document collection.

6.2.1 Baseline (QT-SMT-form)

Our baseline system follows QT as shown in Figure 6.4. First, queries in a source language are translated into English (the document language) using the QT-SMT-form system that is described in Section 5.1. The queries are constructed from the *1-best* translation, as it is scored by the SMT decoder.

Then, these translated queries are used for retrieval from the indexed English collection. In all CLIR experiments in this work, after constructing the queries in the language of the collection, Dirichlet model is used for retrieval as described in Section 6.1, with the same retrieval parameters.

Results of the CLIR baseline systems on the training and the test sets are shown in Table 6.2 . We also report the results of the monolingual system that uses English queries for comparison and we consider it as a reference system.

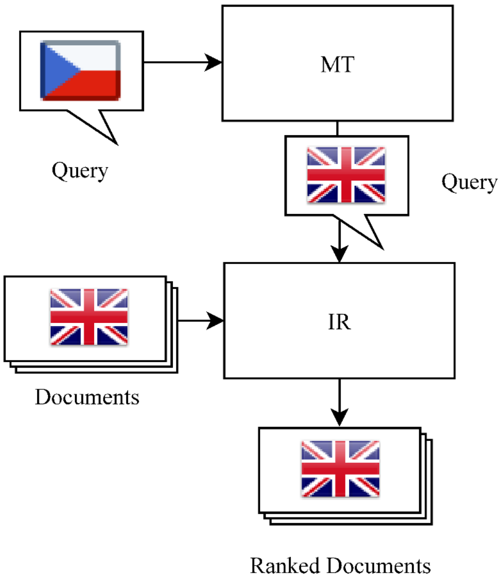


Figure 6.4: Baseline (QT-SMT-form) architecture

Lang/Set	training set			test set		
	P@10	BPREF	MAP	P@10	BPREF	MAP
Mono	***50.70***	***37.76***	***26.14***	***53.03***	***39.94***	***28.31***
Czech	49.10	35.97	23.76	47.27	36.79	23.30
French	48.00	35.78	23.34	48.03	35.65	24.45
German	44.90	32.60	20.04	44.24	35.38	22.51
Hungarian	45.80	34.95	21.88	45.91	37.08	23.60
Spanish	48.00	35.97	23.05	46.97	37.24	24.11
Swedish	44.60	33.22	20.54	40.00	33.24	20.94
Polish	45.40	35.24	21.72	42.12	33.77	21.00

Table 6.2: Results of CLIR baseline systems (QT-SMT-form) and the monolingual system on the training and test sets in percentages

Table 6.3 shows performance comparison between the baseline system and the monolingual system in terms of P@10 and the translated queries.

The baseline system outperforms the reference system in some cases when the translation contains a different synonym of the same concept as in the French query *2015.34*. The term *caries* appears in the query translation instead of the term *cavity*. The same case in the Czech query *2015.40*, the terms *infant* in the baseline query versus the term *stain*, and the term *stain* versus *blotch* in the reference query. This makes the query outperform the reference one by 40% absolute. By conducting further analysis, we find that *infant* has TF in the document collection equal to 173505, while TF of *baby* is 2 only.

The monolingual English corpus that is used to train LM for SMT systems contains the English CLEF eHealth document collection that we use in our CLIR experiment. This makes the SMT system give a higher weight for translations that contain terms or expressions that have high TF in the document collection, and eventually reduces the mismatch problem between translated queries and relevant documents.

However, the main problem of the baseline system is OOV, wherein the translation system is unable to translate query terms in the source language to the target language because such terms did not appear in the training data of the SMT system, as shown in the Czech query *2013.3* where *asystolická* is an unknown word and left untranslated, and the OOV *makulablutung* in the German query *2013.41*. There are 11 terms in the translated Czech test queries left untranslated, 12 in the French queries, and 16 in the German queries. We also find in the translated Spanish queries a total of 10 OOVs. 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, which explains the low performance of these CLIR systems.

Per query analysis of the test set is presented in Figure 6.5 and Figure 6.6. The bars represent the difference between P@10 of each query in the baseline CLIR system and the monolingual system for each language. A positive value means that the corresponding query improved in the baseline compared to the monolingual system. A negative value indicates a degradation.

Query: 2015.40 (Czech)
ref: baby red blotch on face (00.00)
base: infant red stain on face (40.00)

Query: 2015.42 (Czech)
ref: eye iris large (50.00)
base: big eye iris (40.00)

Query: 2013.3 (Czech)
ref: asystolic arrest (70.00)
base: asystolická arrest (50.00)

Query: 2013.41 (German)
ref: right macular hemorrhage (50.00)
base: makulablutung right (00.00)

Query: 2013.45 (German)
ref: symptoms and disease of aortic insufficiency (90.00)
base: symptoms and aortic insufficiency (100.00)

Query: 2015.34 (French)
ref: cavity problem (60.00)
base: problem with caries (90.00)

Table 6.3: Examples of translations of queries including reference (English queries) (*ref*) and baseline (QT-SMT-form) referred to as *base*. The scores in parentheses refer to query P@10 scores in percentages

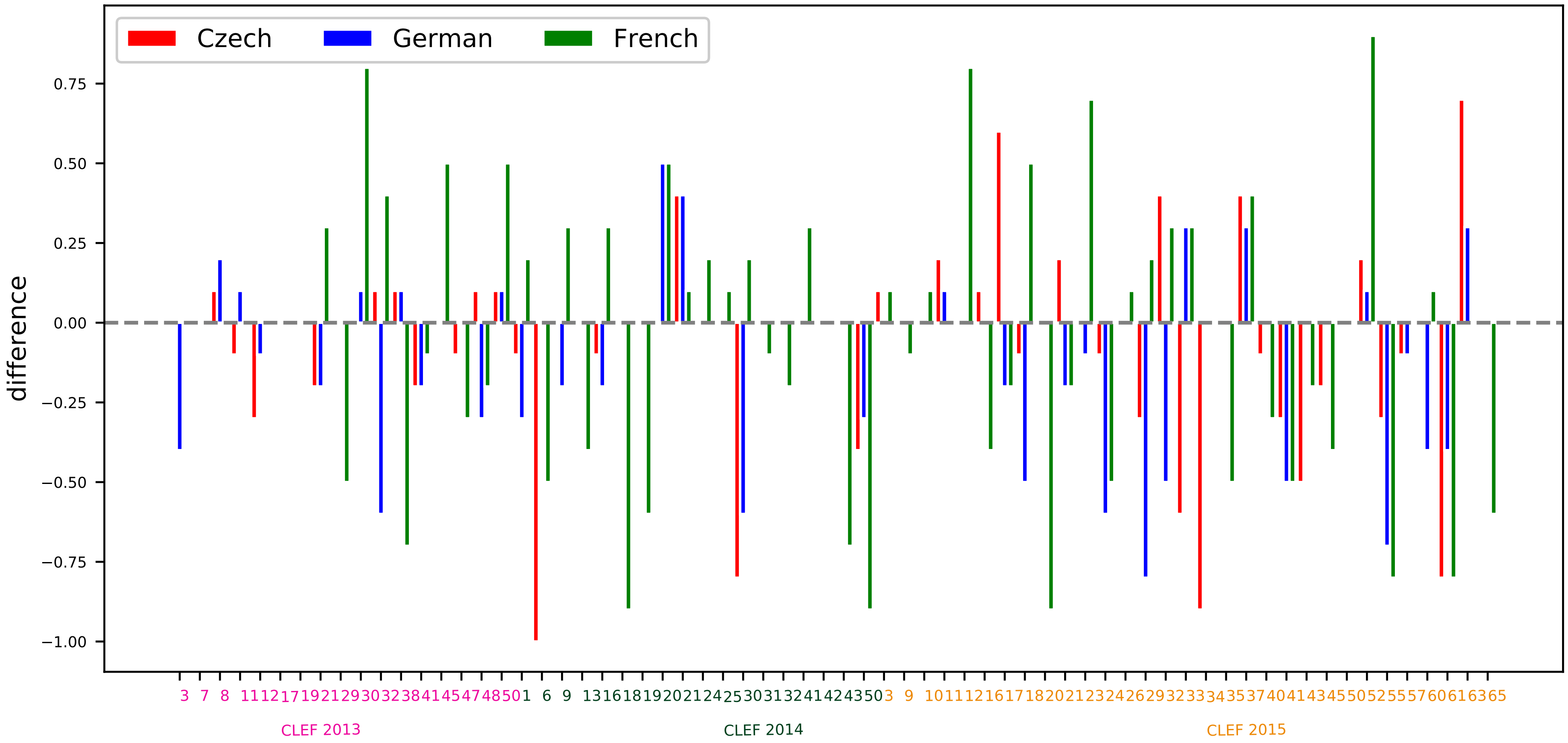


Figure 6.5: Per-query results on the test set of Czech, French and German CLIR systems. The bars represent the difference of P@10 of each CLIR system (QT-SMT-form) and the monolingual system for each query. Labels in X axis represent the query IDs from each year of CLEF eHealth IR test collections

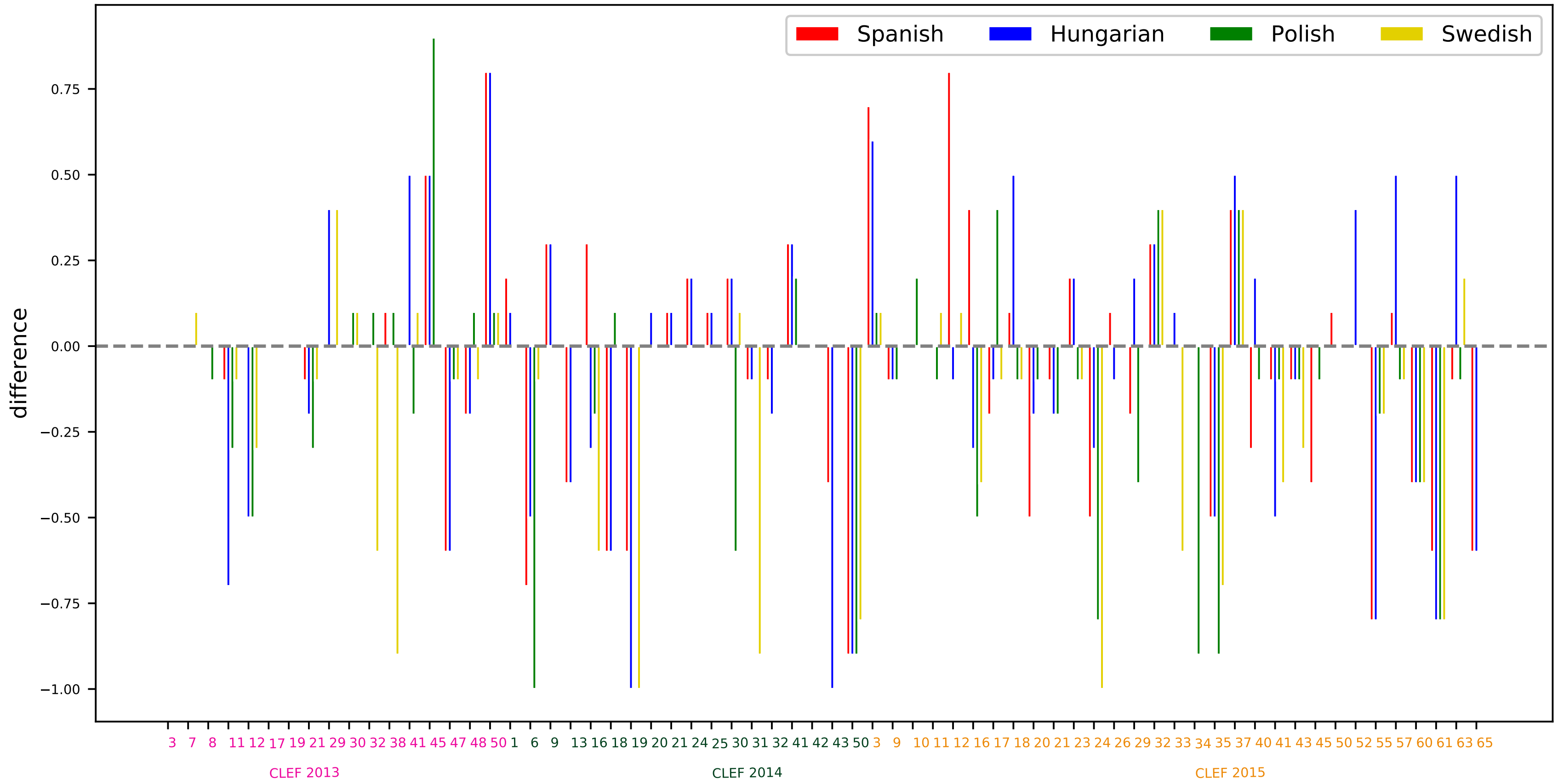


Figure 6.6: Per-query results on the test set of Spanish, Hungarian, Polish and Swedish QT-SMT-form systems. The bars represent the difference of P@10 of each CLIR system and the monolingual system for each query. Labels in X axis represent the query IDs from each year of CLEF eHealth IR test collections

6.2.2 Translation Hypotheses Reranking

Previously, we used the MT system as a black box in our query-translation experiments, in which we always took the best translation for the query construction. However, MT systems usually generate a list of alternative translations. This list is rich of multiple synonyms and candidate translations at phrase and word level of an input sentence. We use the term *rerank* to refer to the process of reranking the SMT translation hypotheses with respect to CLIR performance.

Method

There are two approaches to rank translation hypotheses towards better CLIR performance, either to directly tune the decoder to do that by replacing the translation metric during tuning with P@10 (IR metric) or reranking the produced translations externally. We choose the latter because it will be easier to use the reranker with other SMT systems that produce *n-best-list* translations. The first approach is also more computationally expensive to include P@10 as an objective function in MERT. Because it requires conducting retrieval from the collection for every translation hypothesis during the optimization process.

Our method is limited to translations that are produced by SMT . Because in NMT, we do not have access to internal MT scores as the case in SMT. Internal scores in SMT system play important role in our method.

Training

When translating a query q_i to a target language, we get a ranked list of translation hypothesis $q_{i,j}$. This list is ranked descendingly according to the scores produced by the SMT decoder.

We represent each translation hypothesis by a vector of features (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 reason behind subtracting the oracle P@10 (O_j) from the hypothesis P@10 ($P_{i,j}$) is that some queries do not have relevant documents by any of their translation hypotheses, even these hypotheses might be good translations. If we do not do that, we will be assuming that all of these hypotheses are bad examples, which will not be helpful for the system to distinguish between good and bad training samples. 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]; thus, the problem of reranking is converted into a regression problem.

We consider the number of translation hypotheses to be ranked as an experiment hyperparameter, thus it needs to be tuned. A small list does not give a linguistically rich list for ranking, and having a big list makes it difficult for the model in test time to distinguish between the good and the bad candidates. We follow grid search optimisation to decide how many hypotheses to use and decide to use 15 hypotheses based on the model performance.

Training samples are generated by *15-best-list* of translation hypotheses from 100 queries in 3 languages (Czech, German, and French). This leads to 4500 training instances.

To normalize the data, we apply standardization on each feature of the training set. Then get the coefficient values (feature's mean and variance) and use them to normalize the test data as well. This normalization approach will transform the data to have zero mean and unit variance.

We apply the Leave-One-Out Cross-Validation (LOOCV) approach on the 100 queries, then testing on 1 query. The model predicts the P@10 value for each hypothesis. Then the query is generated from the hypothesis that has the minimum response value (the highest P@10).

We employed the GLM implementation in R,[1] which optimizes the model parameters by the iteratively reweighted least squares algorithm.

Feature Set Description

We present a set of features that are extracted from multiple resources including the SMT system, the document collection and external resources like Wikipedia and PubMed articles. Our feature set includes the following:

SMT Scores The main set of features are the eight scores from the SMT models plus the final translation score. They include scores of the following models:

- *Phrase translation model* that ensures that the individual phrases correspond to each other.
- *Target language model*, which estimates the fluency of the output sentence.
- *Reordering model* capturing different phrase orders in the two languages.
- *Word penalty* penalizing translations that are too long or too short.

For more details about these scores and final translation scores, see Section 5.1.

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

[1] `https://www.r-project.org/`

Retrieval Status Value (RSV) RSV is the value of the retrieval scoring function, given a query and a document, this value is an estimation of the relevance degree of that document by Dirichlet model. This feature is the score of the top-ranked document for each hypothesis after conducting the retrieval. This feature is motivated by the work of Nottelmann and Fuhr [2003], where they investigated the correlation between the RSV and the probability of relevance.

WIKI This feature is motivated by Liu and Nie [2015] and Herbert et al. [2011], who found that Wikipedia abstracts in the medical domain usually contain a description of diseases and health-related topics using simple language (less medical terms), while the titles are usually expressed using medical terms. This helps us in this work to match between queries that are written in a simple language by laypeople and the related medical concepts.

Our approach is as follows: First, we index all Wikipedia articles using their titles and abstracts. Second, for each query, we translate it into English taking its *1-best-list* and conduct retrieval using Dirichlet model. Then we create a pool from the top 10 ranked articles. Third, from this pool, we calculate the sum and the average of term frequency for each term in a given hypothesis.

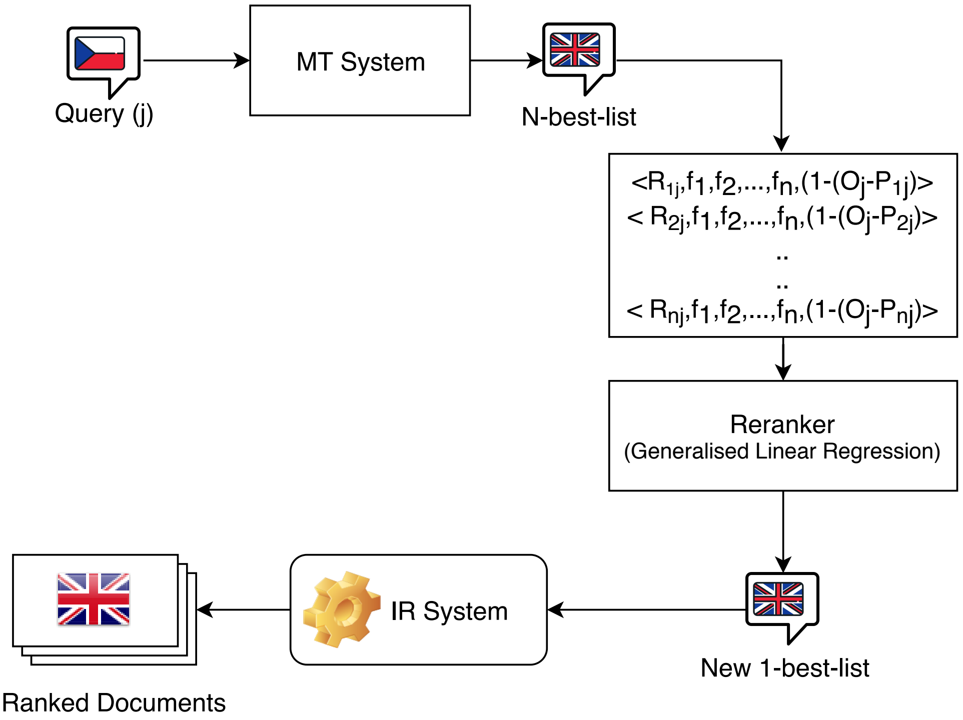
BRF Motivated by the blind-relevance feedback approach for query expansion [Yu et al., 2003], a single best translation provided by QT-SMT-form for each query is used to retrieve the highest ten ranked documents, and each hypothesis is scored by the sum and average of term frequencies extracted from the retrieved documents.

IDF To distinguish translations containing informative terms, we score each hypothesis by the sum and average of inverse document frequency of its terms. IDF is calculated from the document collection.

TP Synonyms and acronyms appear in the alternative translations of the same term in the source sentence because the SMT decoder generates lattice paths to produce the final translation, and different paths (hypotheses) might contain the same terms. When multiple hypotheses share the same term, it gives that term a higher probability to be more relevant. To investigate this effect, we create a Translation Pool (TP) by concatenating all hypotheses for a given query. Then each translation hypothesis is scored by the sum and average of term frequencies extracted from the merged n-best-list.

UMLS Two features based on the UMLS Metathesaurus [Schuyler et al., 1993]: First feature is 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). This is done by annotating all translation hypotheses with medical concepts using MetaMap. The second feature is the number of unigrams and bigrams that match entries in the UMLS Metathesaurus. This helps to give extra weight to the hypotheses that contain medical terms. Researchers have investigated the use of UMLS concepts in the CLEF eHealth IR tasks [Goeuriot et al., 2015], [Goeuriot et al., 2014]. Also, the work reported by Choi and Choi [2014] inspired us to use this feature.

Testing: For testing, translation hypotheses of the test queries are scored by this model, and the lowest-scored hypothesis (the greatest P@10) is selected to be used for retrieval. After we generate queries, we conduct retrieval using the Dirichlet model (Section 2.3.3).



Query (j)
MT System
N-best-list
$<R_{1j},f_1,f_2,...,f_n,(1-(O_j-P_{1j})>$
$< R_{2j},f_1,f_2,...,f_n,(1-(O_j-P_{2j})>$
..
..
$< R_{nj},f_1,f_2,...,f_n,(1-(O_j-P_{nj})>$
Reranker
(Generalised Linear Regression)
New 1-best-list
IR System
Ranked Documents

Figure 6.7: The architecture of the reranker model: first the SMT system produces English translation hypotheses for a given query in the source language, then each hypothesis will be converted into training sample by assigning a response and a set of feature values, then the GLM will be trained to distinguish good translations from bad ones and produces a new 1-best-list that is used to create a query for retrieval

Experiments and results

For each query (both in the training and test sets), we consider 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).

The test data is normalized using the coefficients that are obtained when we normalize the training data.

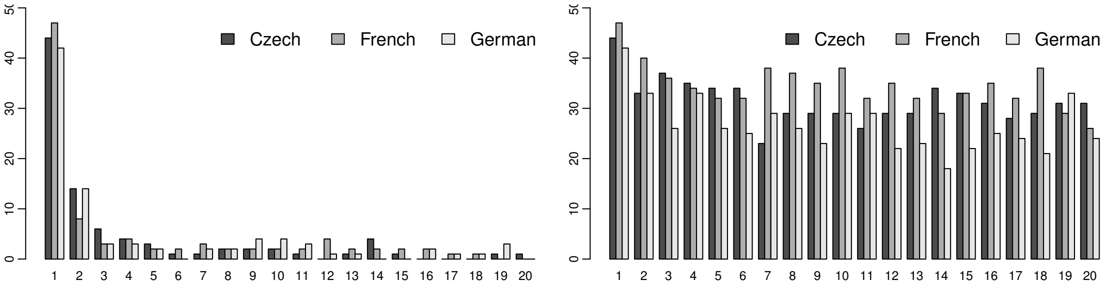


Figure 6.8: Histograms of ranks of translation hypotheses with the highest P@10 for each training query: the first such ranks only (left), all such ranks (right)

Oracle Experiment

To confirm the hypothesis that 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 6.8. The two plots visualize the distribution of the rank of 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 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 that can be selected to achieve optimal performance.

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). Results of the 5-best-list concatenation on the test data are shown in Table 6.6 (row *5-best*).

Reranking

We test our reranking method with several combinations of features. The complete results are displayed in the middle section of Table 6.6. 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 (QT-SMT-form). The significance test is performed using the Wilcoxon signed-rank test with α=0.05.

The system based only on the SMT features does not bring any substantial improvement over the baseline (denoted as **SMT** in the table) for any of the

	Czech		French		German	
system	P@10	MAP	P@10	MAP	P@10	MAP
Mono	53.03	28.31	53.03	28.31	53.03	28.31
QT-SMT-form	47.27	23.30	48.03	24.45	44.24	22.51
5-best	38.94	22.30	41.06	23.05	30.45	17.28
SMT	44.70	24.77	48.79	25.81	42.73	22.65
+RANK	48.64	***25.73***	48.48	26.07	44.55	24.09
++IDF	48.03	25.22	48.64	26.10	44.39	***24.11***
++BRF	47.27	24.99	49.70	26.64	43.64	23.76
++TP	45.76	23.74	48.48	26.26	44.39	24.07
++WIKI	48.64	***25.73***	49.24	26.36	43.64	23.76
++UMLS	48.64	***25.73***	49.09	26.09	44.55	24.09
++RSV	48.64	25.66	48.94	25.95	43.03	23.55
ALL	***50.15***	***25.73***	***51.06***	***27.86***	***45.30***	23.71

Table 6.4: Complete results of the final evaluation on the test set queries (in percentage)

languages. P@10 is improved by less than 1 point for French only, which is not a statistically significant improvement. However, it is degraded significantly for both Czech and German.

The traditional way of SMT tuning towards translation quality does not seem sufficient if no additional features are available. However, adding the explicit features derived from the SMT rankings helps a lot (row *+RANK*), especially for Czech and German, where the increase of the P@10 and MAP scores is statistically significant.

The effect of the other features is studied independently by adding those features to the model with the *SMT+RANK* features. However, in terms of P@10, none of them brings any notable improvement. Although the *BRF*, *WIKI*, and *UMLS* features improve the results for French, they are not statistically significant even in comparison with the QT-SMT-form.

The baseline (QT-SMT-form), however, is outperformed significantly by a system combining all the features (row *ALL*). P@10 is increased by 2 points on average for all languages. In comparison with the monolingual results, the *ALL* system performs at 94.56% for Czech, 96.28% for French, and 85.42% for German.

We observe in the alternative translations of the German language less diversity in the available synonyms compared to other languages, some queries have translation hypotheses that are just different in word order compared with 1-best-list translation.

In Figure 6.9, we present a detailed comparison of the QT-SMT-form 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 QT-SMT-form system. Positive values denote improvement, which is observed for a total of 10 queries in Czech, 15 queries in German, and 8 queries in French. Negative values denote degradation, which is observed in 2 cases for Czech, 4 cases for German, and 3 cases for French. A good example of a query with improved translation is 2015.11 (reference translation: *white patchiness in mouth*). The Czech QT-SMT-form translation *white coating mouth* is improved to *white coating in oral cavity* (P@10 is increased from 10.00 to 80.00) and the French QT-SMT-form *white spots in the mouth* is improved to *white patches in the mouth* (P@10 is increased from 10.00 to 70.00). More examples are given in Table 6.5.

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)

Table 6.5: Examples of translations of training queries including reference (*ref*), oracle (*ora*), QT-SMT-form (*base*) and best (*best*) translations (system using all features). The scores in parentheses refer to the query P@10 scores in percentages

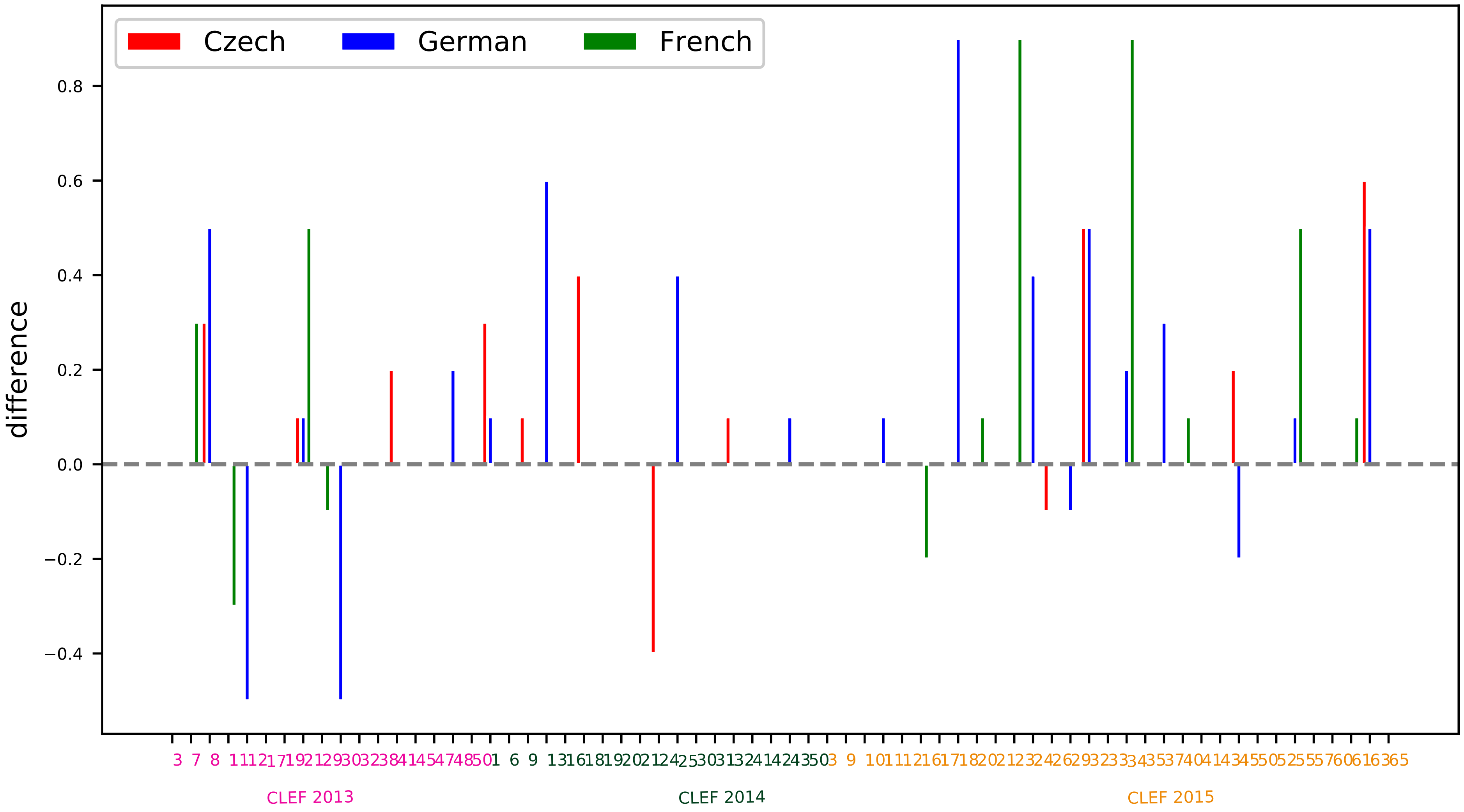


Figure 6.9: Per-query results on the test set. The bars represent absolute difference of P@10 of the best system (*ALL*) and the baseline system (QT-SMT-form) for each query and each language

	Spanish		Hungarian		Polish		Swedish	
system	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
Mono	50.03	28.31	50.03	28.31	50.03	28.31	50.03	28.31
QT-SMT-form	46.97	24.11	45.91	23.60	42.12	21.00	40.00	20.94
SMT	43.79	23.83	40.00	22.54	35.61	19.76	38.33	19.85
SMT+Rank	43.64	24.28	38.94	21.91	38.18	20.21	36.21	20.02
ALL	***50.15***	**25.11**	***48.48***	***25.28***	**43.18**	**21.76**	**41.36**	**21.29**

Table 6.6: Final evaluation results of the monolingual system (mono), the baseline system (QT-SMT-form), and the translation hypotheses reranker (with multiple combinations of features) on the test set for Spanish, Hungarian, Polish and Swedish in percentages

Adaptation to New Languages

The presented reranking model is designed for three CLIR systems: Czech, French and German, because at some point in our research timeline, we did not have queries available in other languages. However after the queries became available, we want to study if the model can be adapted to the new 4 languages and if we can make use of the training data that we obtain from the 3 languages for the adaptation process. We want to investigate in this section if the features in our presented supervised machine learning model are language independent or language specific. If the features are language independent, it means that the training dataset can be easily used to train new model for new languages. This can help when introducing a new language with limited training data by employing training data from other languages.

We first translate the queries in the extended languages (Spanish, Hungarian, Polish and Swedish) using the QT-SMT-form into English, and for each query, we get 15-best-list translation hypotheses, then we generate the feature values as described in Section 6.2.2. We merge the training data from the Czech, German and French languages with the training data from these languages, which leads to a larger training data.

Table 6.6 shows the results for the reranking models for the new languages against the test set.

For the Spanish and Hungarian languages, the system combining all features (ALL) significantly outperforms the baseline system (QT-SMT-form). A small and not statistically significant improvement is observed for Swedish and Polish by the system based on all the features.

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*). 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.

The same case can be observed in query *2015.35* (reference query: *lot of irritation with contact lenses*). The performance of this query is 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.

6.2.3 Query Translation using NMT System (QT-NMT-form)

In this section, we present our experiments conducting the query-translation CLIR approach using the NMT models that we presented in Section 5.5.2.

The goal of these experiments is to compare the CLIR performance when translating the queries using NMT systems versus the performance when translating them using SMT systems.

We translate the queries from all languages into English using QT-NMT-form; then, we conduct retrieval using our monolingual settings.

Results of the CLIR systems using NMT are shown in Table 6.7 (QT-NMT-form). The bold and italic numbers indicate the best system in each language and numbers in bold only indicate a system that is not statistically different than the best system.

QT-NMT-form system significantly outperforms the CLIR system that employs the SMT models (QT-SMT-form) in all languages. Moreover, the Czech QT-NMT-form outperforms the monolingual IR system. This can be explained because the NMT models in our work are adapted to translate medical content by employing the collection itself in the back-translation process. This gives the model access to the collection vocabularies that are frequent in the retrieval collection, and in the relevant documents eventually.

Table 6.8 shows a comparison between the performance of queries that were translated using SMT (QT-SMT-form) and NMT (QT-NMT-form), together with the public MT based CLIR systems and the monolingual IR system.

We observe that when translating the queries using the QT-NMT-form, there were no OOVs in the translations comparing to the OOVs that appeared in the SMT translations. In NMT, dealing with OOVs is different than SMT, BPE

Lang	MT System/Year	P@10	MAP	BPREF
Mono	-	53.03	28.31	39.94
Czech	Baseline (QT-SMT-form)	47.27	23.30	36.79
	QT-NMT-form	***57.27***	**26.02**	***40.74***
	Google 2016	52.88	***26.69***	**40.56**
	Google 2017	55.76	**26.06**	**40.00**
	Bing 2016	48.94	24.78	38.05
	Bing 2017	48.03	21.92	35.47
French	Baseline (QT-SMT-form)	48.03	24.45	35.65
	QT-NMT-form	***51.52***	24.11	36.70
	Google 2016	**51.21**	***25.46***	37.11
	Google 2017	48.94	24.89	37.47
	Bing 2016	50.00	**25.31**	37.09
	Bing 2017	50.15	**25.15**	***38.56***
German	Baseline (QT-SMT-form)	44.24	22.51	35.38
	QT-NMT-form	50.30	22.53	36.62
	Google 2016	***51.67***	***26.02***	***39.07***
	Google 2017	47.42	24.86	37.67
	Bing 2016	48.18	23.77	36.51
	Bing 2017	45.61	21.37	35.45
Spanish	Baseline (QT-SMT-form)	46.97	24.11	37.24
	QT-NMT-form	49.09	22.69	36.87
	Google 2017	***53.33***	**26.61**	**39.18**
	Bing 2017	**53.18**	***26.96***	***40.34***
Hungarian	Baseline (QT-SMT-form)	45.91	23.60	37.08
	QT-NMT-form	***50.76***	**24.08**	**37.84**
	Google 2017	47.27	***24.33***	***38.14***
	Bing 2017	41.67	20.58	33.73
Polish	Baseline (QT-SMT-form)	42.12	21.00	33.77
	QT-NMT-form	47.27	22.38	35.55
	Google 2017	***50.00***	***24.40***	**37.62**
	Bing 2017	48.03	**24.26**	***38.40***
Swedish	Baseline (QT-SMT-form)	40.00	20.94	33.24
	QT-NMT-form	***50.15***	***23.89***	***37.83***
	Google 2017	42.58	20.54	32.99
	Bing 2017	46.67	**23.60**	35.97

Table 6.7: Results of CLIR systems based on query translation using the presented QT-SMT-form, NMT and public MT systems against the test set

tackles this issue by using a fixed dictionary for the most frequent sub-words in the training data and deals with translation as an open-vocabulary approach (as we presented in Section 5.3.3). This helps to significantly reduce the effect of OOVs in the translated queries and boosts the CLIR performance eventually.

Query: 2013.38 (Czech)

ref: mi and hereditary (0)

SMT: im and hereditary (0)

NMT: hereditary myocardial infarction (10)

Query: 2015.17 (Czech)

ref: scaly rash (10)

SMT: scaly rash (10)

NMT: rash squamous (80)

Query: 2015.61 (French)

ref: fingernail bruises (40)

SMT: bruising under the nail (10)

NMT: nail hematoma (60)

Query: 2014.19 (Swedish)

ref: l common carotid aneurysm (60)

SMT: l aneurysm in halspulsåder (0)

NMT: carotid artery aneurysm (100)

Query: 2015.34 (Hungarian)

ref: cavity problem (60)

SMT: caries problem (90)

NMT: tooth decay disorder (100)

Query: 2015.61 (Spanish)

ref: fingernail bruises (40)

SMT: bruising in toe nail (20)

NMT: nail hematoma (60)

Table 6.8: Comparison of query translation by two CLIR systems and the monolingual IR system. Query translation is done using SMT (QT-SMT-form) and NMT (QT-NMT-form) models presented in this work. Numbers in parentheses represent P@10 performance (in percentage) of retrieval when using the translation as a query

6.2.4 Query Translation using Public MT Systems

In this experiment, we use two public MT systems for query translation, namely Bing Translator, and Google Translate.

We translated the queries into English two times; the first time was in 2016 when these two MT systems were using phrase-based SMT, and in 2018 after both deployed NMT systems in their public translation services. This allows us to compare the effect of moving from phrase-based SMT into neural MT on the medical CLIR performance.

At the time (2016) when we translated the Czech, French and German queries using the public MT systems into English, we did not have the queries translated manually from English into Spanish, Hungarian, Polish and Swedish; thus; we do not report the performance of CLIR systems for these languages, however, in 2017, we could do that because we performed full manual translations expanding our CLIR systems into seven languages.

Interestingly, we can observe in Tables 6.9 and 6.7 that the performance of the CLIR systems (QT using Google Translate) degrades from 2016 to 2017 significantly in both training and test sets for German and French.

The values marked in bold and italic refer to the best system in each language that significantly outperforms all CLIR systems for that language, and bold only value refers to the system that is not statistically different than the best performing one.

Notably, the Czech CLIR system using Google Translate in 2017 significantly outperforms the same system in 2016. According to Wu et al. [2016], the NMT approach that Google adopted reduced the translation errors by 60% on average compared to the previously used SMT approach. This shows that improving the translation quality does not necessarily improve the CLIR performance.

The same case appears for Bing Translator. The worst-case appears in the German CLIR system, which degrades from 48.18% in 2016 to 45.61% in 2017.

For further analysis of the results, we apply per query comparison between the translations of the same system in 2016 and 2017. In 2016, the systems tended to have more untranslated terms (OOV) in the output. For example, the Czech query 2015.63 (*krustovitá ložiska na kůži*) was translated by Google Translate into *krustovitá bearing skin* resulting in 00.00% of P@10, while in 2017, it was translated into *crusty bearings on the skin* increasing P@10 to 30.00%.

We can explain the outperformance of SMT over NMT because NMT showed to perform poorly in a domain-specific task (medical query translation in our case) comparing to SMT [Dowling et al., 2018]. This happens because, in SMT, monolingual text from domain-specific data is usually used to train the language model component, which helps eventually to make the translations more fluent in the target language.

There are some proposed solutions to overcome this weakness of NMT, such as leveraging monolingual data through iterative back-translation or fusing an RNN-based language model that is trained on domain-specific monolingual data together with the NMT model [Gulcehre et al., 2015]. However, we are not sure exactly if these commercial MT systems (Google Translate and Bing Translator) are using any techniques for domain adaptation.

Lang	MT System/Year	P@10	MAP	BPREF
Mono	-	50.70	26.14	37.76
Czech	Baseline (QT-SMT-form)	49.10	23.76	35.97
	Google 2016	***52.90***	***25.15***	***37.45***
	Bing 2016	46.80	23.35	36.88
	Google 2017	51.00	23.75	36.46
	Bing 2017	47.50	23.39	36.55
French	Baseline (QT-SMT-form)	48.00	23.34	35.78
	Google 2016	***52.50***	***25.56***	***38.53***
	Google 2017	50.20	24.97	36.79
	Bing 2016	***51.60***	***25.89***	***39.02***
	Bing 2017	48.30	24.10	36.19
German	Baseline (QT-SMT-form)	44.90	20.04	32.60
	Google 2016	***49.10***	**23.04**	**35.40**
	Google 2017	47.70	22.53	35.39
	Bing 2016	***49.10***	***23.51***	***36.23***
	Bing 2017	47.10	21.14	34.48
Spanish	Baseline (QT-SMT-form)	48.00	23.05	35.97
	Google 2017	**49.20**	**23.68**	**36.76**
	Bing 2017	***49.30***	***23.73***	***37.08***
Hungarian	Baseline (QT-SMT-form)	**45.80**	**21.88**	34.95
	Google 2017	***46.10***	***22.04***	***35.78***
	Bing 2017	45.10	21.35	34.64
Polish	Baseline (QT-SMT-form)	45.40	21.72	35.24
	Google 2017	44.90	20.31	33.50
	Bing 2017	***50.40***	***24.21***	***36.96***
Swedish	Baseline (QT-SMT-form)	44.60	20.54	33.22
	Google 2017	48.00	22.50	35.93
	Bing 2017	***49.90***	***23.62***	***36.47***

Table 6.9: Results of CLIR systems that are based on query translation using QT-SMT-form, Google Translate and Bing Translator in 2016 and 2017 on the training set

6.2.5 Conclusions

In this section, we presented our approaches to CLIR using two MT systems (SMT and NMT). We followed the query translation approach to translate the queries into the document language (English) using QT-SMT-form and QT-NMT-form.

The QT-NMT-form CLIR system significantly outperformed the QT-SMT-form. The way we employed back translation in training and selection the final model helped to boost the CLIR results.

We found empirically that MT systems did not often produce the best translation from IR perspective. Motivated by this finding, we presented a machine learning model to rerank translation hypotheses towards better IR performance. Our presented feature set aimed at detecting the translation hypotheses that were more useful for retrieval, and reranked those translations accordingly.

Our reranking of translation hypotheses work was published as a long paper in the CLEF 2016 main conference [Saleh and Pecina, 2016a], and we published the approach of adapting the reranker to new languages as a short paper in MedIR SIGIR 2016 (the Medical IR workshop organized during the Special Interest Group on Information Retrieval conference 2016) [Saleh and Pecina, 2016b]. We followed the reranker approach during our participation in the 2016 and 2017 CLEF eHealth CLIR tasks [Saleh and Pecina, 2016c, 2017]

6.3 Document Translation

The document translation approach to CLIR is the process of translating the document collection into the query language as a reduction of the CLIR task into a monolingual IR task. The general architecture of a DT system is shown in Figure 6.10, where MT can be SMT or NMT.

We presented in Section 4.1 the studies that have been conducted in comparing DT and QT approaches. However, we realized that there is no recent research comparing the two methods. Moreover, some recent work [Khiroun et al., 2018] assumed that DT is better than QT relying on old studies from 1998 [Oard, 1998].

The main argument for this hypothesis is that text in the DT approach is translated in a larger context (sentences, documents) than short isolated queries in the QT approach, and the larger context should help in translation disambiguation and better lexical selection during translation, which should subsequently lead to better retrieval results.

This hypothesis needs to be revised, taking into consideration the significant improvement of machine translation quality in recent years, despite the strong practical disadvantages of DT over QT: DT is computationally expensive and hard to scale (all documents need to be translated into each supported language and indexed) while QT is performed in query time and only a short text (query) is translated into the document language only when needed, and the index does not change.

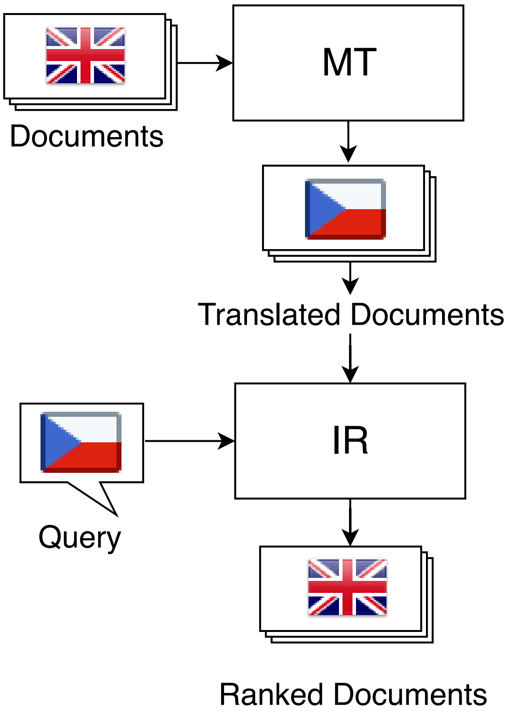


Figure 6.10: In the document translation approach, documents are translated into the query language and then indexed. The IR system takes as an input query in the source language and returns a ranked list of documents. Retrieved documents can be either translated or kept in their original language, so the user has the choice to use their own MT system

6.3.1 Approaches

Our DT approaches focus mainly on two aspects. The first aspect is to investigate the CLIR performance when using SMT versus NMT for the translation process. The second aspect is to study the effect of morphological processing on retrieval quality. To achieve this, we propose the following systems.

- **DT-SMT-form** In this experiment, we employ the SMT system for sentence translation, which is presented in Section 5.1, to translate the documents from English into the seven languages.

- **DT-NMT-form** This system employs NMT that we deployed for this purpose (Section 5.5.2).

- **DT-SMT-pre-lem** In this system, the documents are translated into the target language in the lemmatised form, because the SMT-pre-lem was trained to do so as we presented in Section 5.5.1.

- **DT-SMT-post-lem** First, we translate the documents into the target languages using SMT-form, then we lemmatise them.

- **DT-SMT-post-stem** Translated documents (using SMT-form) are stemmed using Snowball stemmer.

6.3.2 Results

Table 6.10 shows results of our experiments categorised into monolingual systems, CLIR query-translation approach and document-translation approach using both SMT and NMT system.

We can explain the outperformance of QT over DT because of the domain of our dataset. This means that queries include symptoms and health conditions where linguistics and contextual information do not play a significant role in solving ambiguity in the translation process.

For example, the Czech query *clef2015.test.33*, which is "*bílá infekce hltanu*", is translated into English as "*white infection of pharynx*". The reference translation for that query is "*white infection in pharynx*". We can see that the CS→EN SMT system fails in translating prepositions ("*of*" instead of "*in*"), which are considered stopwords in our setting; hence, this does not affect the CLIR performance.

A total of 5 queries are improved in the Czech *SMT-pre-lem* CLIR system over the *SMT-form* system, because lemmatization helped to reduce the search space by mapping all the morphological variants into one word. For example, the English query *clef2013.test.18*: *"aspiration pneumonia and pharyngeal dysphagia"* is "*aspirační pneumonie a dysfágie hltanu*" in Czech. The world "*hltanu*", which means "*pharyngeal*" is lemmatised in the training data of the SMT system and the Czech query into "*hltan*", which means "*pharynx*". When translating the English documents into Czech, "*pharynx*" and "*pharyngeal*" are translated back into "*hltan*". This helped to retrieve more relevant documents, increasing P@10 to 0.9 in DT-SMT-pre-lem from 0.7 in the monolingual systems (mono, mono-lem and mono-stem), 0.6 in QT-SMT-form and 0 in DT-SMT-form. Although *DT-SMT-post-lem* outperforms *DT-SMT-pre-lem* in most languages in terms of

SMT evaluation as we showed in Section 5.6, *DT-SMT-pre-lem* improved the CLIR performance as shown in *DT-SMT-pre-lem)* over DT-SMT-post-lem in four languages, and degraded the results in Hungarian, Swedish, and Polish.

DT-SMT-pre-lem in Spanish is the only DT system that outperforms the QT system. While for DT-NMT-form, we can observe a significant drop in performance in all languages comparing to SMT, except for Czech and French. This can be explained because it is known that NMT performs poorly when translating long sentences (which is the case in the document translation) comparing to SMT [Koehn and Knowles, 2017]. The source of the monolingual text in back translation in QT-NMT-form is the document collection (Section 5.5.2). To investigate the effect of monolingual text source on CLIR, we train another CS-EN NMT system by choosing a different source, which is the English side of the parallel text, and the rest remains exactly the same. Then, we used the trained NMT system to translate Czech queries into English, and then performing CLIR. Interestingly, P@10 for this system is 54.2% versus 57.2% when using the document collection (significantly higher). This shows that employing the document collection in back-translation helps to produce translations that are more adapted to the collection domain.

6.3.3 Conclusions

Our results showed that a well-tuned QT system outperforms DT, which is a positive result with an important impact on practical applications. So far, the QT approach has been preferred mainly for efficiency reasons (less space and computation needed). Our experiments suggest that this approach is even more effective (better retrieval results).

We also investigated the effect of using NMT, which is now considered state of the art in various domains. This completely new paradigm in MT tends to improve the fluency of the generated output.

In our experiments, NMT improved retrieval results in both QT and DT, but the QT approach is still superior, so the results are consistent with the findings from the SMT experiments. However, we want to emphasize that the way we trained our MT systems is very domain-specific (medical domain), and we made use of a vast amount of medical data (monolingual and parallel). This makes our comparative study very domain-oriented.

When dealing with general domain test collection, some search terms might have a different meaning in different domains. For example, the word "development" probably in most cases means in medicine the growth or spread of a disease (or a tumor), while in a general domain we can not say without a context, and in that case, the need for linguistics information in the queries will be more important to solve translation ambiguity. This should be considered when comparing QT and DT approaches; thus, the reader should be careful when drawing the same conclusion of this work when working on a different domain.

	Czech		**French**		**German**		**Hungarian**		**Spanish**		**Swedish**		**Polish**	
System	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
Monolingual (Oracle)														
Mono-form	53.0	28.3	53.0	28.3	53.0	28.3	53.0	28.3	53.0	28.3	53.0	28.3	53.0	28.3
Mono-lem	52.1	27.5	52.1	27.5	52.1	27.5	52.1	27.5	52.1	27.5	52.1	27.5	52.1	27.5
Mono-stem	52.1	26.4	52.1	26.4	52.1	26.4	52.1	26.4	52.1	26.4	52.1	26.4	52.1	26.4
Query translation														
QT-SMT-form	47.2	22.6	48.0	23.6	44.2	21.7	45.9	22.9	46.9	23.2	40.0	20.2	42.1	20.1
QT-NMT-form	***57.2***	***26.0***	***51.5***	***24.1***	***50.3***	***22.5***	***50.7***	***24.0***	***49.0***	22.6	***50.1***	***23.8***	***47.2***	***22.3***
Document translation														
DT-SMT-form	39.0	17.4	42.1	21.5	40.4	**22.1**	40.0	17.2	45.6	**26.9**	38.3	17.0	40.7	20.4
DT-SMT-post-stem	36.9	16.7	44.5	22.7	39.2	***22.9***	35.4	17.0	46.3	***27.3***	33.9	16.7	35.3	18.7
DT-SMT-post-lem	39.3	18.3	41.9	21.7	37.2	**22.4**	37.1	17.0	42.7	25.0	33.0	16.0	37.1	**22.2**
DT-SMT-pre-lem	42.8	21.3	43.6	20.6	42.1	19.8	36.5	16.8	47.7	22.4	30.7	12.6	34.8	19.7
NMT-form	42.1	15.6	46.0	19.8	36.6	14.0	26.0	10.5	43.9	17.5	33.9	11.6	38.9	12.3

Table 6.10: Retrieval results of the cross-lingual (and monolingual) IR experiments (Monolingual, CLIR query-translation approach and CLIR document-translation approach) employing SMT and NMT systems with different morphological processing settings and using the test set. Bold and italic results donate those which are statistically significantly better than the QT-SMT-form (the best performing CLIR systems for each language), only bold results donate those which are not statistically different than the best results

Finally, we published our finding of comparing DT and QT using SMT and NMT as a long paper in the main conference of ACL 2020 (Association of Computational Linguistics) 2020 [Saleh and Pecina, 2020].

6.4 Term Selection for Query Expansion

In this section, we present our approach for query expansion using a QE term selection method in CLIR.

The proposed method is based on a simple linear regression model that predicts the retrieval performance for each candidate expansion term when combined with a query translated by an SMT. The model features are obtained from the SMT system, external document sources (Wikipedia, PubMed), and information from the document collection. The model is used to score each term from a candidate pool, and those scored above a (pre-trained) threshold are automatically added to the translated query. As a result, the queries are expanded with strong candidates only. If no strong candidates are available, the queries remain unchanged. This prevents performance drop caused by adding irrelevant terms to the query.

The term selection method is performed in four steps. First, a set of candidate terms (candidate pool) are collected from various sources. Second, each term from the candidate pool is assigned a vector of features describing its potential to identify relevant documents. Third, the features are combined in a regression model to score each candidate term. Finally, terms with scores exceeding a given threshold are selected to expand the query. Figure 6.11 shows the architecture of our presented model in detail, and the following sections explain the term selection process.

6.4.1 Candidate Pool

In order to expand a query with one or more terms, it is needed to have a reliable source of candidate terms to choose from. Alternative translations sometimes showed to be very limited in terms of the variations of synonyms or related terms, wherein that case each different translation contains either the same terms in a different order or punctuations and stopwords. This makes the alternative translations the same as the best translation in terms of the usefulness in the context of information retrieval. To remedy the limitation of available related candidate terms, we introduce three sources of candidate terms: terms form the translation hypotheses, terms from the related articles from PubMed and terms from the related abstracts from Wikipedia, as illustrated in Figure 6.12.

The pool is created from three different resources as follows:

Machine translation (MT)

When translating a query in a source language into a target language, an SMT system produces a set of translation hypotheses (*n-best-list*), as shown in Table 6.12. Usually, only the best translation (*1-best-list*) is used for retrieval as we

source	*query id: 2014.1.cs*	*query id: 2015.11.cs*
src:	ischemická choroba srdeční	bílé povlaky v dutině ústní
ref:	coronary artery disease	white patchiness in mouth
	translation hypotheses	
1	red ischaemic heart disease	white coating mouth
2	ischemic heart disease	white coating oral
3	heart disease	white coating the mouth
4	blue coronary heart disease	oral white coating
5	ischaemic disease	white coating in oral cavity
6	ischemic blue cardiac disease	white coating in mouth
7	coronary disease	white sheets oral
8	ischaemic cardiac disease	white coatings oral
9	ischemic disease	white coating in oral
10	coronary artery disease	the white coating mouth
11	ischemic cardiac	white coating of mouth
12	cardiac disease	white sheets mouth
13	blue stroke heart	white coatings mouth
14	heart disease	mouth white coating
15	ischaemic cardiac	oral white sheets

Table 6.11: Examples of two queries showing the source query in the source language, the reference query in the target language, and the translation hypotheses of the source query in English, underlined terms are candidates for query expansion obtained by QT-SMT-form

showed in our baseline setting (QT-SMT-form), which is *ocular tremor* in this example.

However, other hypotheses might contain multiple terms that are related to the information need that is represented in the query, but they do not appear in the best translation, for example: *eyes*, *shaking* and *trembling.* To exploit these terms in the presented method, we use translation hypotheses to create a bag-of-words and added to the candidate pool. We remove from this pool duplicated terms and stopwords. For each source query, we collect all the terms from the 100 highest-scored translation hypotheses as produced by the SMT system.

English Wikipedia (Wiki)

The tendency of the queries in the consumer health search, which are posted by laypeople, is the lack of medical terminologies (as shown in Section 2.2). Liu and Nie [2015] showed that using Wikipedia articles to enrich such queries with medical terms helps to improve the medical information need that is represented in their initial query. Because disease names usually appear in the title, and their symptoms are described in the abstract.

We make use of this finding as follows: First, we index Wikipedia collection using titles and abstracts. Then, we use our baseline settings (generating queries using 1-best-list translation with the Dirichlet retrieval model) to retrieve the top 10 ranked retrieved articles. Then we add to the candidate pool the titles of

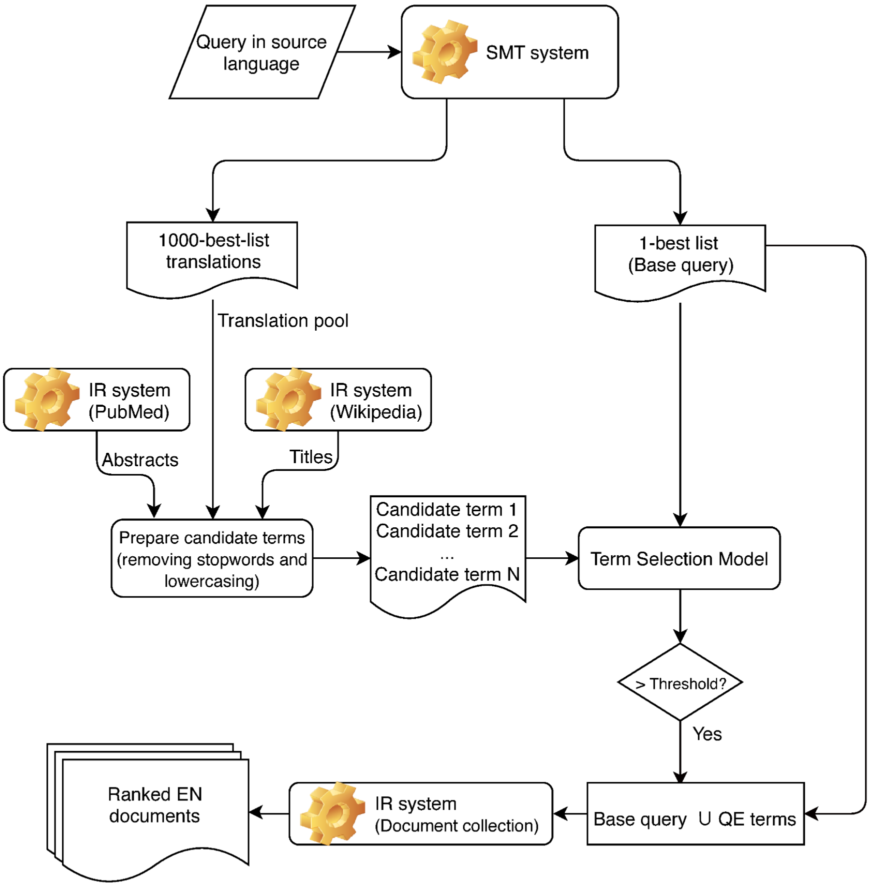
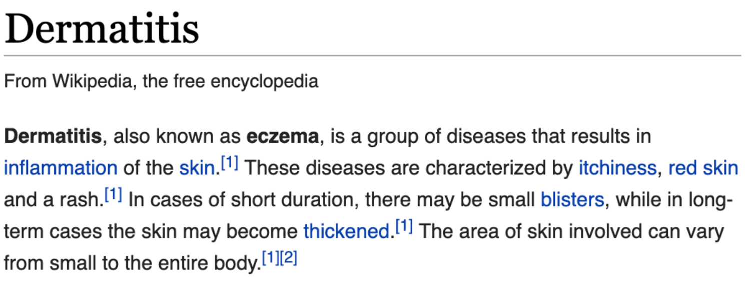


Figure 6.11: The architecture overview of our term selection method

Wikipedia articles only. This helps enrich the pool with more medical terms.

For example, when translating the Czech query *clef2015.test.39* in the training set, which is *kožní ekzém s puchýřky a hnisem*, the best translation from the Czech to English QT-SMT-form system is *skin eczema with pustules and pus*. We use this English translation to retrieve the top-ranked articles from Wikipedia. The article abstract is presented in Figure 6.13. It is shown that the title of this article (dermatitis) is a medical term, while the abstract contains an explanation of this disease in a simpler language. This helps map the medical terms that are found in Wikipedia titles to each query.



Dermatitis

From Wikipedia, the free encyclopedia

Dermatitis, also known as **eczema**, is a group of diseases that results in inflammation of the skin.[1] These diseases are characterized by itchiness, red skin and a rash.[1] In cases of short duration, there may be small blisters, while in long-term cases the skin may become thickened.[1] The area of skin involved can vary from small to the entire body.[1][2]

Figure 6.13: A snippet from the abstract of the Wikipedia article titled “Dermatitis“, source `https://en.wikipedia.org/wiki/Dermatitis`

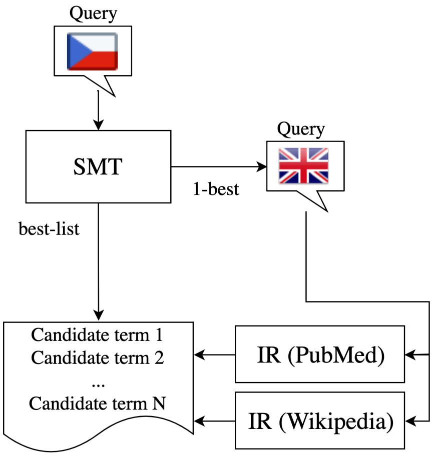


Figure 6.12: Generating candidate pool of English terms for an input Czech query

QT-SMT-form rank	English hypothesis
1	ocular tremor
2	eyes tremor
3	eyes trembling
4	eyes shivering
5	eyes shaking
6	eyes jitter
7	eyes shakiness
8	ocular trembling
9	ocular tremor
10	ophthalmic tremor
11	ocular shivering
12	eyes dithering
13	eyes tremble
14	eyes tingling
15	as ocular tremor

Table 6.12: English translation hypotheses of a German query (Augenzittern) as ranked by QT-SMT-form

PubMed

PubMed is a public digital archive for biomedical and life science journal literature in the United States. It contains more than 30 million English articles; thus, it is considered a very important linguistic resource for NLP applications in the medical domain.[2]

We enrich the candidate pool with terms from the PubMed articles [Peng et al., 2016] following the settings as the Wikipedia articles. PubMed articles (both abstracts and titles) are indexed, then the top 10 ranked articles are retrieved using the *1-best* translation as a base query and added to the candidate pool.

[2] https://www.ncbi.nlm.nih.gov/pubmed/

6.4.2 Feature Set

Our goal is to define a set of features that enable our supervised model to detect the usefulness of each term in the candidate pool for retrieval. We introduce the following features:

IDF
The IDF score of each candidate term calculated in the document collection. This feature helps statistically measure the importance of a term in the collection and distinguishes informative terms from non-informative ones (such as stopwords).

Translation pool frequency
The frequency of the term in the 100 highest-scored translation hypotheses as produced by the SMT system. When a term appears in multiple hypotheses, this means that the probability of being a relevant translation to one of the terms in the original query is high. This feature is excluded from our monolingual QE model.

Wikipedia frequency
The frequency of the term in the top 10 Wikipedia abstracts retrieved from the Wikipedia index using the *1-best* translation as a base query.

Retrieval Status Value (RSV)
RSV feature is the difference of the RSV value (the score of the Dirichlet retrieval model) of the highest-ranked document retrieved using the base query, and the RSV value of the highest-ranked document retrieved using the base query expanded by the candidate term. This feature tells us the contribution of the candidate term to the RSV score.

Query similarity
This feature represents the average similarity between a candidate term t_m and a query term obtained using a pre-trained model of *word2vec* embeddings on 25 million articles from PubMed.[3] First, we get a word embedding for each term in the original query, and we sum those embeddings to get a vector that represents the entire query. Then we take the embeddings for t_m and calculate the cosine similarity between the query vector and the t_m vector. It is important to point out here that choosing terms that are similar to each term of the query caused significant drift in the information need. For example, *mother* was suggested as a similar term to *baby*, and *white* as a similar term to *black*.

Co-occurrence frequency
The co-occurrences of a candidate term t_m and the query terms $t_i \in Q$ indicates how likely t_m is related to the original query Q, we sum up the co-occurrence frequency for each term in query Q and the candidate term t_m in all documents

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

d_j in the collection C, as shown below:

$$co(t_m, Q) = \sum_{d_j \in C, t_i \in Q} tf(d_j, t_i) TF(d_j, t_m)$$

Term frequency
First, we perform retrieval from the collection using a query that is constructed from the *1-best* translation. The translation is done using the QT-SMT-form system (SMT for query translation), then we calculate the term frequency of a candidate term t_m in the top 10 ranked documents from the retrieval result.

UMLS frequency
This feature represents how many times a term appeared in the UMLS lexicon Humphreys et al. [1998], as an attempt to give more weight to the medical terms.

6.4.3 Regression Model

The term selection model is based on linear regression. Training instances are candidate terms for the training queries after translating those queries from all seven languages into English. Each term t from a candidate pool of a given query is assigned a value computed as the difference of P@10 obtained by the baseline query (*1-best-list* translation) and P@10 obtained by the expanded baseline query with the term t. Expansion terms increasing P@10 for the given query are assigned positive values, terms decreasing P@10 are assigned negative values, and terms without any effect on the retrieval performance for that query are assigned zero. The purpose is to expand the queries with terms that can improve the performance, rather than terms that might degrade it.

The feature values are normalized using standard scaling by removing the mean and scaling them to have unit variance. This is done independently on each feature on the training set. Then we use the scaler coefficient to standardize the test set. Scaling is important since the range of the feature values varies widely.

We consider P@ difference as the objective function, and we use the proposed feature set to train the model. Linear Regression (LR) models the relationship between the dependent variable (P@10 in our case) and the regressors x (term feature values).

We use ordinary least squares linear regression as it is implemented in the *scikit* package [Pedregosa et al., 2011]. There might be one or more good candidate terms for expansion. To select these terms, we set a threshold value for the predicted score. The threshold value is tuned on the training set for all languages, as shown in Figure 6.15. All terms which have a score equal or higher than the threshold are added to the base query. This allows us to avoid expanding queries with irrelevant terms.

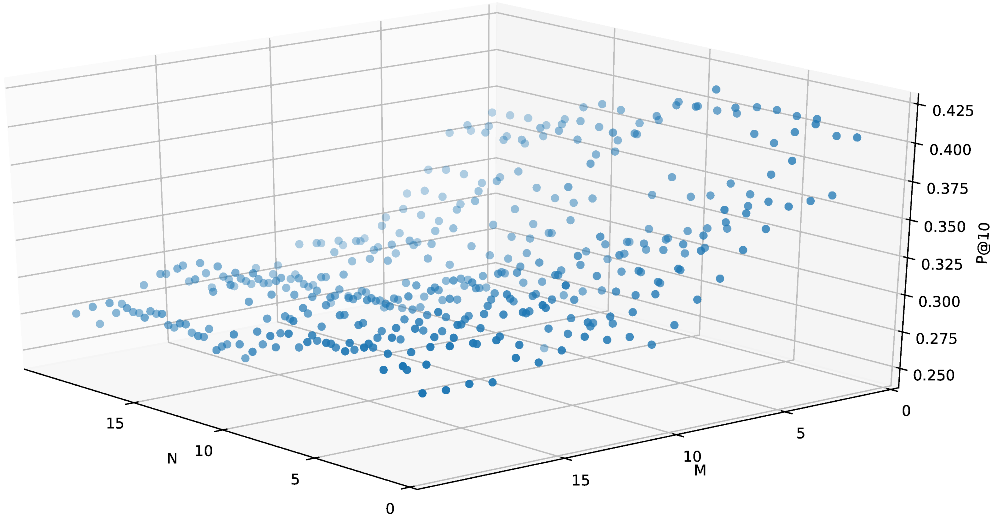


Figure 6.14: Tuning KLD parameters, number of documents (N) and number of expansion terms (M) on the monolingual queries

6.4.4 Results

Results of all experiments for the seven languages are presented in percentages in Table 6.13 (in terms of P@10) and Table 6.14 (in terms of BPREF). For each language, the underlined score denotes the best result, and the scores in bold refer to results that are not significantly different (given the Wilcoxon signed-rank test) from the best (underlined) score. TS refers to the proposed QE technique based on term selection, and the text in the brackets denote the candidate term sources: machine translation (MT) hypotheses, Wikipedia titles (Wiki), and PubMed articles (PubMed).

The monolingual experiment (exploiting the reference English queries) sets a theoretical upper-boundary for the results of the CLIR experiments. It is 53.03 in terms of P@10 and 39.94 in terms of BPREF. These values hold for all the languages since the reference translations of the source queries are the same.

P@10 scores of the CLIR baseline systems (exploiting *1-best* translation) range between 40.00 and 48.03 depending on the query language. The KLD-based expansion in CLIR brings the scores even lower (36.36–45.76), which is in line with the monolingual expansion experiments. Though, for some queries (10 on average), P@10 improved and results degraded for more queries (20 on average).

The proposed term selection experiments show consistent improvement over the baseline (QT-SMT-form). The best system uses terms from MT and Wiki for expansion. Samples of queries that are improved by this system are shown in Table 6.16.

TS(MT$\cup$Wiki) improved 21 queries in Czech, 18 in French, 14 for German and 11 in Spanish, 10 queries in Hungarian, 2 queries in Polish, and 3 queries in Swedish. While it degraded 11 queries in Czech, 12 in French, 11 in German, 4 in Spanish, 5 queries in Hungarian, 2 queries in Polish, and 2 queries in Swedish, the performance of the rest of the queries did not change. The average result in Czech is very close to the monolingual performance. Table 6.17 shows examples

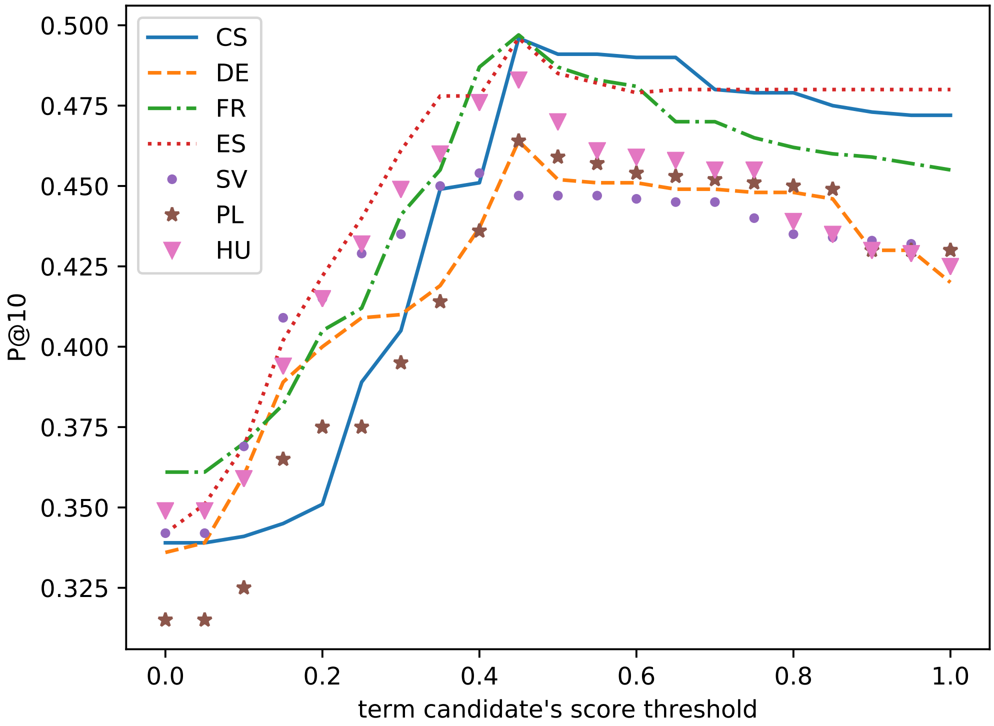


Figure 6.15: Tuning threshold for term candidate selection based on their predicted scores

of queries that degraded in the TS(MT ∪ Wiki) system.

For further analysis of the expansion quality, we report in Table 6.15 the percentage of relevant expansion terms calculated by two methods: In the first method, we provided a medical doctor with query titles, their narratives (to understand the topic for each query), and the expansion terms as suggested by the TS(MT ∪ Wiki) system. We asked them to identify the expanded terms, whether they are relevant to the topic or not.

The second method is an automatic evaluation that is done by checking if the expansion terms exist in the reference queries. For example, in the Czech system, 78.51% of the expansion terms did not appear in the reference query; however, we could not tell if they are relevant or not. In contrast, when checked by a medical doctor, it appeared that only 12.4% of them are irrelevant to the topic.

6.4.5 Term Selection for Monolingual IR

The presented method is designed to predict the usefulness of terms from a candidate pool when added base query. To adapt this method for monolingual IR, we change a few things in the model as follows:

- The initial retrieval for Wikipedia frequency, RSV and term frequency is done using the base query, which is the English query (reference) in this case, not the *1-best* translation.

- The candidate pool contains only terms from PubMed. And this is done by using the English query to retrieve the top 10 abstracts from the PubMed

system/query language	CS	FR	DE	ES	HU	PL	SV
Monolingual	53.03	53.03	53.03	53.03	53.03	53.03	53.03
+KLD	48.18	48.18	48.18	48.18	48.18	48.18	48.18
+TS(PubMed)	55.76	55.76	55.76	55.76	55.76	55.76	55.76
QT-SMT-form	47.27	48.03	44.24	46.97	45.91	42.12	40.00
+KLD	39.85	45.76	38.33	42.12	42.12	39.24	36.36
+TS(MT)	47.42	48.03	43.03	46.82	46.21	**42.42**	***41.52***
+TS(Wiki)	44.85	44.70	43.03	43.18	**47.12**	41.06	39.70
+TS(PubMed)	50.15	47.12	43.33	45.30	43.48	37.58	36.52
+TS(MT ∪ Wiki)	***52.58***	***49.55***	***47.12***	***48.33***	***47.88***	**42.42**	***41.52***
+TS(MT ∪ PubMed)	50.30	**48.79**	**45.45**	**48.03**	42.73	38.48	34.85
+TS(MT ∪ Wiki ∪ PubMed)	**52.12**	**48.94**	**45.45**	**47.42**	**47.58**	***43.18***	**41.21**

Table 6.13: Experiment results in terms of P@10 in percentages against the test set. Bold and italic numbers indicate the best systems that are statistically different from the baseline system, bold numbers only indicate systems that are not statistically different from the best systems

index. This can be done using any external resource relevant to the document collection.

- Query similarity is calculated using English queries and candidate terms.

Then, we generate the feature values as presented above for the English queries. We use this data for testing, and we train the term selection model using the CLIR data (after excluding the translation pool frequency). Results for applying the term selection method on the monolingual data is presented in the *TS(PubMed)* system.

Monolingual+KLD refers to the result of the KLD-based query expansion applied to the reference translations of the queries. In terms of P@10, the result went down substantially. This can be explained because either the indexed documents are not good enough as a source of candidate expansion terms, or because there is no criterion to prevent expanding some queries with low scored term candidates.

In terms of P@10, we observe in the expanded queries 13 improvements, and 4 queries degraded. The rest of the queries did not change due to the low scores of candidate terms as predicted by the model. In terms of BPREF, both KLD and TS bring a small improvement, which is not statistically significant.

Because we use PubMed as a source of expansion terms, we can notice the presence of medical terms in the expanded queries. This helps enrich queries with more medical information. For example, the base query clef2015.test.18 (*poor gait and balance with shaking*) is expanded by the term *parkinson*. The model suggests in this case that the symptoms that appear in the base query refer to the Parkinson disease.

system/query language	CS	FR	DE	ES	HU	PL	SV
Monolingual	39.94	39.94	39.94	39.94	39.94	39.94	39.94
+KLD	41.22	41.22	41.22	41.22	41.22	41.22	41.22
+TS(PubMed)	41.41	41.41	41.41	41.41	41.41	41.41	41.41
QT-SMT-form	36.79	35.65	35.38	37.24	37.08	33.77	20.94
+KLD	36.21	**38.34**	34.84	**39.64**	36.59	***34.33***	32.11
+TS(MT)	36.80	35.49	35.64	37.05	37.03	**33.92**	**33.38**
+TS(Wiki)	36.82	36.10	36.09	36.17	***38.77***	**33.82**	***34.23***
+TS(PubMed)	**39.16**	**38.14**	**39.15**	***39.47***	36.87	33.51	**33.78**
+TS($\mathrm{MT} \cup \mathrm{Wiki}$)	***40.49***	***38.82***	***40.86***	37.93	36.95	**33.92**	**33.38**
+TS($\mathrm{MT} \cup \mathrm{PubMed}$)	38.90	**37.63**	36.09	**38.87**	36.57	**34.16**	**33.67**
+TS($\mathrm{MT} \cup \mathrm{Wiki} \cup \mathrm{PubMed}$)	**40.21**	**37.15**	36.02	37.93	37.70	**33.86**	32.98

Table 6.14: Experiment results in terms of BPREF (percentages). Bold and italic numbers indicate the best systems that are statistically different from the baseline system, bold numbers only indicate systems that are not statistically different from the best systems

measure / query language	CS	FR	DE	ES	HU	PL	SV
Precision w.r.t. manual judgments	87.60	89.33	90.84	87.50	96.43	90.91	87.50
Precision w.r.t. reference translations	21.49	14.04	13.74	25.00	21.43	36.36	12.50

Table 6.15: Precision of selected terms manually checked by a medical expert (first raw) and with respect to the terms that appeared in the reference English queries (second raw)

Query: 2015.18 (Czech)
ref: poor gait and balance with shaking (50.00)
base: bad posture and balance with tremor (60.00)
QE: poor balanced shaking (70.00)

Query: 2014.21 (French)
ref: white patchiness in mouth (10.00)
base: renal impairment (00.00)
QE: kidney disease function dysfunction failure insufficiency deficiency poor (30.00)

Query: 2013.11 (German)
ref: chest pain and liver transplantation (50.00)
base: breast pain and liver transplantation (10.00)
QE: chest hepatic graft thoracic (40.00)

Query: 2014.11 (Spanish)
ref: Diabetes type 1 and heart problems (40.00)
base: type 1 diabetes and heart problems (40.00)
QE: cardiac disease (60.00)

Table 6.16: Examples of queries from different systems including Mono (*ref*), baseline using QT-SMT-form (*base*), and expansion terms to the baseline query (*QE*). The scores in parentheses refer to the query P@10 scores in percentages

Query: 2013.41 (Czech)
ref: right macular hemorrhage (60.00)
base: amacular bleeding right (70.00)
QE: hemorrhage haemorrhage side blood (30.00)

Query: 2013.41 (French)
ref: right macular hemorrhage (60.00)
base: macular hemorrhage right eye (80.00)
QE: eyes haemorrhage hemorrhagic bleeding (50.00)

Query: 2015.65 (German)
ref: weird brown patches on skin (10.00)
base: strange brown spots on the skin (40.00)
QE: spot patches cutaneous patch (10.00)

Query: 2014.31 (Spanish)
ref: Acute renal failure (80.00)
base: acute renal failure (80.00)
QE: kidney disease (60.00)

Table 6.17: Examples of queries degraded in the QE approach (*QE*) with respect to Mono (*ref*), Baseline using QT-SMT-form (*base*). The scores in parentheses refer to the query P@10 scores in percentages

6.4.6 Conclusions

We presented in this section our automatic QE method in CLIR. The goal of QE is to improve information need represented in user queries.

Our source of expansion (candidate pool) includes alternative translations from SMT, English Wikipedia abstracts, and the titles of PubMed articles. Our QE can be used in different domains than the medical domain by choosing a relevant candidate pool. For example, translation hypotheses can be replaced with similar synonyms, for example, using word2vec in the monolingual task. This makes our model not exclusive to the context of CLIR.

In order to find how a term is informative for expansion, we presented a rich set of features to define the contribution of each term to the CLIR performance. Then we trained a regression model that predicts the change of performance of the base query when expanded with one or more candidate terms.

We found that not every query in the test set should be expanded, especially when such a query does not have good candidate terms in the candidate pool. To achieve this, we tuned a threshold value for the term contribution score to the P@10 score, where every query is expanded with all terms that have a score greater than the threshold value. This helped us to avoid the degradation of the performance of some queries in the test set.

The evaluation of our QE method showed a significant improvement in the CLIR performance for all languages compared to the baseline system. When we applied the method on the monolingual system and using only PubMed as a source of candidates, QE significantly outperformed the monolingual baseline system, which confirms that our method can be applied on both CLIR and monolingual IR.

We published our term selection method in the European Conference on Information Retrieval (ECIR) 2019 as a long paper [Saleh and Pecina, 2019b] and applied our method during our participation in the 2018 CLEF eHealth CLIR task [Saleh and Pecina, 2018].

Conclusions

In this thesis, we studied the task of Cross-lingual Information Retrieval (CLIR) in the medical domain. The task is patient-centered, the case when a searcher does not have medical experience and prefers to use the Internet for self-diagnosis of their medical conditions, rather than consulting a medical expert because of their various reasons.

We used the existing test collections from the CLIR eHealth IR tasks 2013–2015 which allowed CLIR experiments in seven languages (Czech, German, French, Spanish, Hungarian, Polish and Swedish). However, the judgment rate of our system results was low and the manual translations of queries were incomplete; thus we conducted manual translation of the English queries to cover all the studied languages and performed an intensive relevance assessment to fully evaluate our developed systems. The extended test collection is available from the LINDAT repository and licensed under the Creative Commons - Attribution-NonCommercial 4.0 to allow further research of the task.

We presented a strong monolingual IR system using the English queries and considered this system as reference (topline) to our CLIR systems. The same monolingual retrieval settings were used in the CLIR task after reducing it into a monolingual IR task.

We replicated two Statistical Machine Translation (SMT) systems that were developed within the Khresmoi project. The first system was tuned to translate short queries and the second one to translate full sentences as can appear in documents. We thoroughly studied the main approaches to CLIR: query translation (QT) in which queries were translated to the document language, and document translation (DT) wherein documents were translated to the query language.

In QT, we employed SMT to design a baseline system that takes the best translation as produced by SMT to construct queries. Due to a loss of information during the translation process, the baseline achieved around 84% of the P@10 of the monolingual system on average

Our experiments showed that constructing queries as the best SMT translations failed in 50% cases to produce the best retrieval performance. In the other 50% of the queries, the translation that gave the best retrieval performance was in the top 15-best translations. This was a motivation for us to develop a machine-learning based model that reranked these translations towards a better retrieval performance. The model, which selected one of the 15 best translations produced by SMT system, outperformed the baseline systems that used the default 1-best translation in all languages. This can be explained because exploiting features from the collection and external resources gave the model access to information helped distinguish translations that better represented relevant documents in the collection.

As an alternative to the reranking method, we studied query expansion (QE) to reformulate base queries and eventually to improve the represented information need. We developed a post-translation QE based on supervised learning to expand the query with terms from a rich candidate pool formed from internal resources (translation pool and document collection) and external resources (Wikipedia titles and PubMed abstracts). We presented a rich set of features for each candidate term that reflects the usefulness of that term to the base query. The model predicted the contribution score for each candidate term to the retrieval performance of the base query. We set up a threshold for this score to avoid harming base queries by expanding them with irrelevant terms. Our query expansion method significantly improved the performance when using base queries in both CLIR and monolingual IR.

Neural Machine Translation (NMT) as a new paradigm in MT substantially improved translation quality in the last few years. In our work, we employed NMT-base CLIR systems too. We achieved this with two steps: in the first step, we employed back translation to create synthetic parallel data. The source of the monolingual data in the back translation involved the document collection which gave the NMT systems access to the collection vocabularies. In the second step, we presented a novel approach to define the model selection criterion. Our method ensembled two NMT models: the model that achieved the best BLEU and the model that achieved the best PER. We observed that none of them alone could achieve comparable results compared to the ensembled ones when tested on the training queries. NMT significantly outperformed our SMT systems in the QT approach in all languages. Multiple researchers had been assuming that DT is better than QT since the late 1990's, despite the fact that MT improved significantly in the recent few years. Our comparative study of the two approaches was conducted using two MT paradigms: SMT and NMT. None of the reported DT approaches could outperform the QT approach which was unexpected and contradicted what was reported earlier. Our experiments showed that our own implementation of an SMT system that produced lemma (instead of word forms) helped reduce the mismatching problem in the DT experiments but could not bring further improvements over the QT system. Although NMT outperformed SMT in the QT experiments, it did not outperform SMT in the DT experiments. This is probably because of the bad performance of NMT when translating long sentences as reported earlier by multiple researchers in the machine translation task.

We plan to continue our research in the CLIR task in the medical domain. We will investigate our reranking and term selection approaches in NMT. We will also study the use of the shared cross-lingual embedding to replace the MT systems and design one model that can translate queries for all languages into the collection language. Furthermore, we will research indexing the documents using the distributed vector representation, wherein retrieval can be conducted using similarity score between document embedding and query embedding. This might help solve the mismatching problem in the term-based matching approach.

List of Publications

Petra Galuščáková, Shadi Saleh, and Pavel Pecina. Shamus: UFAL Search and Hyperlinking Multimedia System. In *European Conference on Information Retrieval*, pages 853–856. Springer, 2016.

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*, volume 1180, pages 226–235, Sheffield, UK, 2014.

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*, pages 54–66, Évora, Portugal, 2016a. Springer.

Shadi Saleh and Pavel Pecina. Adapting SMT Query Translation Reranker to New Languages in Cross-Lingual Information Retrieval. In *Proceedings of the Medical Information Retrieval (MedIR) Workshop. A SIGIR 2016 workshop*, Pisa, Italy, 2016b.

Shadi Saleh and Pavel Pecina. Task3 Patient-Centred Information Retrieval: Team CUNI. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, Evora, Portugal, 2016c. CEUR-WS.org.

Shadi Saleh and Pavel Pecina. Task3 Patient-Centred Information Retrieval: Team CUNI. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings*, volume 1866, Dublin, Ireland, 2017.

Shadi Saleh and Pavel Pecina. CUNI Team: CLEF eHealth Consumer Health Search Task 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, pages 1–11, Avignon, France, 2018.

Shadi Saleh and Pavel Pecina. 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*, Lecture Notes in Computer Science, pages 188–195. Springer, 2019a.

Shadi Saleh and Pavel Pecina. Term Selection for Query Expansion in Medical Cross-lingual Information Retrieval. In *Advances in Information Retrieval - 41th European Conference on IR Research, ECIR 2019, Cologne, Germany, Proceedings*, Lecture Notes in Computer Science, pages 507–522. Springer, 2019b.

Shadi Saleh and Pavel Pecina. 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*, pages 6849–6860, Online, 2020. ACL.

Shadi Saleh, Feraena Bibyna, and Pavel Pecina. CUNI at the CLEF eHealth 2015 Task 2. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation Forum*, volume 1391, Toulouse, France, 2015.

List of Abbreviations

AveP Average Precision

BiRNN Bidirectional Recurrent Neural Network

BLEU BiLingual Evaluation Understudy

BP Brevity Penalty

BPE Byte-Pair Encoding

BRF Blind Relevance Feedback

CBOW Continuous Bag-of-Words

CHS Consumer Health Search

CLEF Conference and Labs of the Evaluation Forum

CLIR Cross-Lingual Information Retrieval

COVID-19 Coronavirus disease 2019

DT Document Translation

GPU Graphics Processing Units

HON Health On the Net

IDF Inverse Document Frequency

IR Information Retrieval

IRGAN Information Retrieval Generative Adversarial Network

KLD Kullback-Leiber Divergence

LETOR LEarning TO Rank

LM Language Model

LOOCV Leave-One-Out Cross-Validation

LSA Latent Semantic Analysis

LSTM Long Short-Term Memory

MAP Mean Average Precision

MERT Minimum Error Rate Training

MeSH Medical Subject Heading

MIRA Margin Infused Relaxed Algorithm

MT Machine Translation

MRD Machine-Readable Dictionary

NIST National Institute of Standards and Technology

NMT Neural Machine Translation

NLP Natural Language Processing

OOV Out-Of-Vocabulary

PBMT Phrase-Based Machine Translation

PBSMT Phrase-Based Statistical Machine Translation

PER Position-independent word Error Rate

QE Query Expansion

QT Query Translation

RNN Recurrent Neural Network

RSV Retrieval Status Value

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2

RWE Relevance-based Word Embedding

SMT Statistical Machine Translation

SVD Singular Value Decomposition

TF Term Frequency

TREC Text REtriveal Conference

TSV Term Selection Value

UMLS Unified Medical Language System

List of Figures

List of Tables

Bibliography

Alan Aronson. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The Metamap Program. *AMIA Symposium*, pages 17–21, 2001.

Alan Aronson and François-Michel Lang. An Overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

Niraj Aswani, Thomas Beckers, Erich Birngruber, Célia Boyer, Andreas Burner, and Jakub Bystroň et al. Khresmoi: Multimodal Multilingual Medical Information Search. In John Mantas, Stig Kjær Andersen, Maria Christina Mazzoleni, Bernd Blobel, Silvana Quaglini, and Anne Moen, editors, *Proceedings of the 24th International Conference of the European Federation for Medical Informatics, Quality of Life through Quality of Information, Village of the future*, Pisa, Italy, 2012. IOS Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *The third International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, San Diego, CA, USA, 2015.

Lisa Ballesteros and Bruce Croft. Resolving Ambiguity for Cross-language Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 64–71, New York, NY, USA, 1998. ACM.

Lisa Ballesteros and W. Bruce Croft. Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. *SIGIR Forum*, 31(SI): 84–91, 1997.

N. J. Belkin. The Problem of Matching in Information Retrieval. *Theory and Application of Information Research*, pages 187–197, 1980.

Ondřej Bojar and Aleš Tamchyna. The design of eman, an experiment manager. *The Prague Bulletin of Mathematical Linguistics*, 99:39–58, 2013.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August 2016. ACL.

Alessio Bosca, Matteo Casu, Mauro Dragoni, and Chiara Di Francescomarino. Using semantic and domain-based information in CLIR systems. In *The Semantic Web: Trends and Challenges*, pages 240–254. Springer, 2014.

P. Buitelaar, B. Sacaleanu, Špela Vintar, D. Steffen, M. Volk, H. Dejean, E. Gaussier, D. Widdows, O. Weiser, and R. Frederking. Muchmore: Multilingual concept hierarchies for medical information organization and retrieval. 2003.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 10–51, Stroudsburg, PA, USA, 2012. ACL.

Guihong Cao, Jianfeng Gao, Jian-Yun Nie, and Jing Bai. Extending Query Translation to Cross-language Query Expansion with Markov Chain Models. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 351–360, New York, NY, USA, 2007. ACM.

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting Good Expansion Terms for Pseudo-relevance Feedback. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 243–250, New York, NY, USA, 2008. ACM.

Ganesh Chandra and Sanjay K. Dwivedi. Query Expansion Based on Term Selection for Hindi-English Cross Lingual IR. *Journal of King Saud University - Computer and Information Sciences*, 2017.

Wei-Tsen Milly Chiang, Markus Hagenbuchner, and Ah Chung Tsoi. The WT10G Dataset and the Evolution of the Web. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, pages 938–939, New York, NY, USA, 2005. ACM.

Sungbin Choi and Jinwook Choi. 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*, volume 1180, pages 167–175, Sheffield, UK, 2014. CEUR-WS.org.

Cyril Cleverdon. ASLIB Granfield Research Project on the Comparative Efficiency of Indexing Systems. *Aslib Proceedings*, 12(12):421–431, 1960.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Koby Crammer and Yoram Singer. Ultraconservative Online Algorithms for Multiclass Problems. *The Journal of Machine Learning Research*, 3:951–991, 2003.

Nick Craswell. *Bpref*, pages 266–267. Springer US, Boston, MA, 2009.

Bruce Croft, Howard Turtle, and David Lewis. The Use of Phrases and Structured Queries in Information Retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 32–45, New York, NY, USA, 1991. ACM.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Giorgio M Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. CLEF 2007: Ad Hoc track overview. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 13–32, Berlin, Heidelberg, 2008. Springer.

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. SMT versus NMT: Preliminary comparisons for irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20, Boston, MA, March 2018. Association for Machine Translation in the Americas. URL `https://www.aclweb.org/anthology/W18-2202`.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, and et al. Machine translation of medical texts in the Khresmoi project. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 221–228, Baltimore, USA, 2014. ACL.

Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 644–648, Atlanta, GA, USA, 2013. ACL.

Efthimis Efthimiadis. Query Expansion. *Annual Review of Information Science and Technology (ARIST)*, 31:121–87, 1996.

Andreas Eisele and Yu Chen. MultiUN: A Multilingual Corpus from United Nation Documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. ELRA.

Liana Ermakova and Josiane Mothe. Query expansion by local context analysis. In *Conference francophone en Recherche d'Information et Applications (CORIA 2016)*, pages 235–250, Toulouse, France, 2016. CORIA-CIFED.

Nicola Ferro and Carol Peters. Clef 2009 Ad-hoc Track Overview: TEL and Persian tasks. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 13–35. Springer, 2010.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011.

Susannah Fox. Health Topics: 80% of internet users look for health information online. Technical report, Pew Research Center, 2011.

Atsushi Fujii and Tetsuya Ishikawa. Applying Machine Translation to Two-stage Cross-language Information Retrieval. In *Envisioning Machine Translation in the Information Future*, volume 1934, pages 13–24. Springer, Berlin, Germany, 2000.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 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, pages 674–675, New York, USA, 2009. ACM.

Evgeniy Gabrilovich, Andrei Broder, Marcus Fontoura, Amruta Joshi, Vanja Josifovski, Lance Riedel, and Tong Zhang. Classifying search queries using the web as a source of knowledge. *ACM Transactions on the Web*, 3(2):5, 2009.

Philip Gage. A New Algorithm for Data Compression. *C Users J.*, 12(2):23–38, 1994.

Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, and Changning Huang. 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*, pages 96–104. ACM, 2001.

Fredric C Gey and Aitao Chen. TREC-9 cross-language information retrieval (english-chinese) overview. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pages 15–23, Gaithersburg, US, 2000. NIST.

Fredric C Gey and Douglas W Oard. The TREC-2001 cross-language information retrieval track: Searching Arabic using English, French or Arabic queries. In *The Tenth Text REtrieval Conference (TREC 2001)*, pages 16–26, Gaithersburg, US, 2001. NIST.

M Christopher Gibbons, Renee F Wilson, Lipika Samal, Christoph U Lehman, Kay Dickersin, Harold P Lehmann, Hanan Aboumatar, Joseph Finkelstein, Erica Shelton, Ritu Sharma, et al. Impact of Consumer Health Informatics Applications. *Evidence Report/Technology Assessment*, (188):1, 2009.

Dean Giustini. Online Portals Part I. *Health Care on the Internet*, 3(4):83–87, 1999.

Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen, and Guido Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. *CLEF 2013 Online Working Notes*, 8138:1–16, 2013.

Lorraine Goeuriot, Liadh Kelly, Wei Li, Joao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth Jones, and Henning Mueller. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred Health Information Retrieval. In *Proceedings of CLEF 2014*, pages 43–61, Sheffield,UK, 2014. CEUR-WS.org.

Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Néváol, Cyril Grouin, Joao Palotti, and Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2015. In *The 6th Conference and Labs of the Evaluation Forum*, pages 429–443, Berlin, Germany, 2015. Springer.

G. H. Golub and C. Reinsch. Singular Value Decomposition and Least Squares Solutions. *Numer. Math.*, 14(5):403–420, 1970.

Gregory Grefenstette and Julien Nioche. Estimation of English and non-English Language Use on the WWW. In *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, pages 237–246, Paris, France, France, 2000.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On Using Monolingual Corpora in Neural Machine Translation. *CoRR*, 2015.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. On integrating a language model into neural machine translation. *Comput. Speech Lang.*, 45(C):137–148, September 2017.

Donna Harman. Towards Interactive Query Expansion. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, pages 321–331, New York, NY, USA, 1988. ACM.

Donna Harman. Relevance Feedback and Other Query Modification Techniques. pages 241–263. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1992.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. Improved Neural Machine Translation with SMT Features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 151–157. AAAI Press, 2016.

Benjamin Herbert, György Szarvas, and Iryna Gurevych. Combining query translation techniques to improve cross-language information retrieval. In *Advances in Information Retrieval*, volume 6611, pages 712–715. Springer, Berlin, Germany, 2011.

William Hersh. *Information Retrieval: A Health and Biomedical Perspective*. Springer Science & Business Media, 2008.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative Back-Translation for Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, 2018. ACL.

David A. Hull and Gregory Grefenstette. 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*, pages 49–57, New York, USA, 1996. ACM.

Betsy L. Humphreys, Donald A. B. Lindberg, Harold M. Schoolman, and G. Octo Barnett. The Unified Medical Language System. *Journal of the American Medical Informatics Association*, 5(1):1–11, 1998.

Amélie Imafouo and Xavier Tannier. Retrieval Status Values in Information Retrieval Evaluation. In *String Processing and Information Retrieval*, pages 224–227, Berlin, Heidelberg, 2005. Springer.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China, 2015. ACL.

Jimmy, Guido Zuccon, Joao Palotti, Lorraine Goeuriot, and Liadh Kelly. Overview of the CLEF 2018 Consumer Health Search Task. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes*, Avignon, France, 2018. CEUR-WS.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. ACL.

Jayashree Kalpathy-Cramer, Henning Muller, Steven Bedrick, Ivan Eggel, Alba G Seco De Herrera, and Theodora Tsikrika. Overview of the CLEF 2011 Medical Image Classification and Retrieval Tasks. In *CLEF 2011 - Working Notes for CLEF 2011 Conference*, volume 1177. CEUR-WS, 2011.

Noriko Kando. *NTCIR Workshop : Japanese- and Chinese-English Cross-Lingual Information Retrieval and Multi-grade Relevance Judgments*, pages 24–35. Springer, Berlin, Heidelberg, 2001.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, Joao Palotti, and Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2016. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9822, pages 255–266, Cham, 2016. Springer.

Alla Keselman, Allen C Browne, and David R Kaufman. Consumer Health Information Seeking as Hypothesis Testing. *Journal of the American Medical Informatics Association*, 15(4):484–495, 2008.

Oussama Ben Khiroun, Bilel Elayeb, and Narjes Bellamine Ben Saoud. Towards a Query Translation Disambiguation Approach using Possibility Theory. In *ICAART (2)*, pages 606–613, Portugal, 2018. SciTePress.

Ahmad Khwileh, Haithem Afli, Gareth Jones, and Andy Way. Identifying Effective Translations for Cross-lingual Arabic-to-English User-generated Speech Search. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 100–109, Valencia, Spain, April 2017. ACL.

Sun Kim, W John Wilbur, and Zhiyong Lu. Bridging the Gap: A Semantic Similarity Measure Between Queries and Documents. *arXiv preprint arXiv:1608.01972*, 2016.

David A Kindig, Allison M Panzer, Lynn Nielsen-Bohlman, et al. *Health Literacy: A Prescription to End Confusion*. National Academies Press, Washington, DC, 2004.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, 2017. ACL.

Philipp Koehn and Josh Schroeder. Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Stroudsburg, PA, USA, 2007. ACL.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. ACL, 2003.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *International Workshop on Spoken Language Translation (IWSLT) 2005*, 2005.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, 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*, pages 177–180, Stroudsburg, PA, USA, 2007. ACL.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate Detection Using Shallow Text Features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, New York, NY, USA, 2010. ACM.

Bevan Koopman and Guido Zuccon. 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*, pages 1243–1244. ACM, 2014.

Saar Kuzi, Anna Shtok, and Oren Kurland. Query Expansion Using Word Embeddings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1929–1932, New York, NY, USA, 2016. ACM.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. Unsupervised Cross-Lingual Information Retrieval Using Monolingual Data Only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1253–1256, New York, NY, USA, 2018. ACM.

Xiaojie Liu and Jian-Yun Nie. 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*, volume 1391, Toulouse, France, 2015. CEUR-WS.org.

David E. Losada, Javier Parapar, and Alvaro Barreiro. When to Stop Making Relevance Judgments? A Study of Stopping Methods for Building Information Retrieval Test Collections. *Journal of the Association for Information Science and Technology*, 70(1):49–60, 2019.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, July 2015. ACL.

Craig Macdonald, Vassilis Plachouras, Ben He, Christina Lioma, and Iadh Ounis. University of Glasgow at WebCLEF 2005: Experiments in Per-Field Normalisation and Language Specific Stemming. In *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 898–907. Springer, Berlin, Germany, 2006.

Prasenjit Majumder, Dipasree Pal, Ayan Bandyopadhyay, and Mandar Mitra. Overview of FIRE 2010. In *Multilingual Information Access in South Asian Languages*, pages 252–257. Springer, 2013.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, UK, 2008.

J. Scott McCarley. 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*, pages 208–214, College Park, Maryland, 1999. ACL.

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

Lukas Michelbacher, Florian Laws, Beate Dorow, Ulrich Heid, and Hinrich Schütze. Building a Cross-lingual Relatedness Thesaurus using a Graph Similarity Measure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. ELRA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *ICLR Workshop Papers*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013b. Curran Associates Inc.

George Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Robert C. Moore and William Lewis. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July 2010. ACL.

Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 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*, pages 109–119, Stroudsburg, PA, USA, 2012. ACL.

Rodrigo Nogueira and Kyunghyun Cho. Task-Oriented Query Reformulation with Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583. ACL, 2017a.

Rodrigo Nogueira and Kyunghyun Cho. Task-oriented query reformulation with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583, 2017b.

Henrik Nottelmann and Norbert Fuhr. From Retrieval Status Values to Probabilities of Relevance for Advanced IR Applications. *Information retrieval*, 6: 363–388, 2003.

Giorgio Maria Di Nunzio and Alexandru Moldovan. A Study on Query Expansion with MeSH Terms and Elasticsearch. IMS Unipd at CLEF eHealth Task 3. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, Avignon, France, 2018. CEUR-WS.

Douglas W Oard and Fredric C Gey. The TREC 2002 Arabic/English CLIR track. In *The Eleventh Text Retrieval Conference (TREC 2002)*, pages 1–15, Gaithersburg, US, 2002. NIST.

DouglasW. Oard. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *Machine Translation and the Information Soup*, volume 1529, pages 472–483. Springer, Berlin, Germany, 1998.

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167, Sapporo, Japan, 2003.

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on Advances in Information Retrieval Research*, ECIR'05, pages 517–519, Berlin, Heidelberg, 2005. Springer-Verlag.

Heather O'Brien, Nicola Ferro, Hideo Joho, Dirk Lewandowski, Paul Thomas, and Keith van Rijsbergen. System And User Centered Evaluation Approaches in Interactive Information Retrieval (SAUCE 2016). In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, page 337–340, New York, NY, USA, 2016. ACM.

Dipasree Pal, Mandar Mitra, and Kalyankumar Datta. Improving Query Expansion Using WordNet. *Journal of the Association for Information Science and Technology*, 65(12):2469–2478, 2014.

Joao Palotti, Guido Zuccon, Pavel Pecina Jimmy, Mihai Lupu, Lorraine Goeuriot, Liadh Kelly, and Allan Hanbury. CLEF 2017 task overview: The IR Task at the eHealth Evaluation Lab. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, pages 1–10, Dublin, Ireland, 2017. CEUR-WS.

João RM Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanbury, Gareth JF Jones, Mihai Lu pu, and Pavel Pecina. CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving information about medical symptoms. In *CLEF (Working Notes)*, pages 1–22, Berlin, Germany, 2015. Spriner.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, USA, 2002. ACL.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword Fifth Edition. Philadelphia, USA, 2011. Linguistics Data Consortium.

Kevin Patrick, S. Koss, M. J. Deering, and L. Harris. Consumer Health Information: A Federal Perspective on a Important Aspect of the National Information Infrastructure. In *Proceedings of the Second International Workshop on Community Networking 'Integrated Multimedia Services to the Home'*, pages 261–267, NY, USA, 1995. IEEE Communications Society.

Pavel Pecina, Petra Hoffmannová, Gareth J. F. Jones, Ying Zhang, and Douglas W. Oard. Overview of the CLEF-2007 Cross-language Speech Retrieval Track. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 674–686, Berlin, Heidelberg, 2008. Springer.

Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J.F. Jones, and et al. Adaptation of Machine Translation for Multilingual Information Retrieval in the Medical Domain. *Artificial Intelligence in Medicine*, 61(3):165–185, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Yifan Peng, Chih-Hsuan Wei, and Zhiyong Lu. Improving Chemical Disease Relation Extraction with Rich Features and Weakly Labeled Data. *Journal of Cheminformatics*, 8(1):53, 2016.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. ACL, 2014.

Ari Pirkola. The Effects of Query Structure and Dictionary Setups in Dictionary-based Cross-language Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 55–63, New York, USA, 1998. ACM.

Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin. Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information retrieval*, 4(3-4):209–230, 2001.

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

Martin Popel and Ondřej Bojar. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70, 2018a.

Martin Popel and Ondrej Bojar. Training tips for the transformer model. *PBML*, 110:43–70, 2018b.

Bruno Pouliquen and Christophe Mazenc. COPPA, CLIR and TAPTA: Three Tools to Assist in Overcoming the Patent Barrier at WIPO. In *proceedings of the Thirteenth Machine Translation Summit*, pages 24–30, Xiamen, China, 2011. Asia-Pacific Association for Machine Translation.

Michael Przystupa and Muhammad Abdul-Mageed. Neural Machine Translation of Low-Resource and Similar Languages with Backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235, Florence, Italy, 2019. ACL.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

Phyllis A. Richmond. Review of the Cranfield Project. *American Documentation*, 14(4):307–311, 1963.

Stephen E. Robertson and Karen Sparck Jones. Document Retrieval Systems. chapter Relevance Weighting of Search Terms, pages 143–160. Taylor Graham Publishing, London, UK, 1988.

Joseph John Rocchio. Relevance Feedback in Information Retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323, 1971.

Monica Rogati and Yiming Yang. Cross-Lingual Pseudo-Relevance Feedback Using a Comparable Corpus. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, pages 151–157, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

Dwaipayan Roy, Debasis Ganguly, Mandar Mitra, and Gareth JF Jones. 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.

Andreas Rücklé, Krishnkant Swarnkar, and Iryna Gurevych. Improved Cross-Lingual Question Retrieval for Community Question Answering. In *WWW Conference*, WWW '19, pages 3179–3186, New York , USA, 2019. ACM.

Tetsuya Sakai. Alternatives to Bpref. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 71–78, New York, NY, USA, 2007. ACM.

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*, volume 1180, pages 226–235, Sheffield, UK, 2014. CEUR-WS.org.

Shadi Saleh and Pavel Pecina. *Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval*, pages 54–66. Springer International Publishing, Évora, Portugal, 2016a.

Shadi Saleh and Pavel Pecina. Adapting SMT query translation reranker to new languages in cross-lingual information retrieval. In *Proceedings of the Medical Information Retrieval (MedIR) Workshop. A SIGIR 2016 workshop*, Pisa, Italy, 2016b.

Shadi Saleh and Pavel Pecina. Task3 patient-centred information retrieval: Team CUNI. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, Evora, Portugal, 2016c. CEUR-WS.org.

Shadi Saleh and Pavel Pecina. Task3 patient-centred information retrieval: Team CUNI. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings*, volume 1866, Dublin, Ireland, 2017.

Shadi Saleh and Pavel Pecina. CUNI Team: CLEF eHealth Consumer Health Search Task 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, pages 1–11, Avignon, France, 2018. CEUR-WS.

Shadi Saleh and Pavel Pecina. 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, April 14-18, 2019, Proceedings*, Lecture Notes in Computer Science. Springer, 2019a.

Shadi Saleh and Pavel Pecina. Term Selection for Query Expansion in Medical Cross-lingual Information Retrieval. In *Advances in Information Retrieval - 41th European Conference on IR Research, ECIR 2019, Cologne, Germany,* *Proceedings*, Lecture Notes in Computer Science, pages 507–522. Springer, 2019b.

Shadi Saleh and Pavel Pecina. 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*, pages 6849–6860, Online, 2020. ACL.

G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Sheikh Muhammad Sarwar, Hamed Bonab, and James Allan. A Multi-task Architecture on Relevance-based Neural Query Translation. Florence, Italy, 2019. ACL.

Stefan Schultz, Martin Honeck, and Udo Hahn. Biomedical Text Retrieval in Languages with a Complex Morphology. In *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, pages 61–68, Phildadelphia, Pennsylvania, USA, 2002. ACL.

Peri Schuyler, William Hole, Mark Tuttle, and David Sherertz. The UMLS Metathesaurus: Representing Different Views of Biomedical Concepts. *Bulletin of the Medical Library Association*, 81(2):217, 1993.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, Berlin, Germany, 2016a. ACL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016b. ACL.

Claude Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria, August 2013. ACL.

Michael Smith and Gavriel Salvendy. Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design. In *Proceedings of the Symposium on Human Interface 2007, Held as Part of HCI International 2007*, volume 4557, China, Beijing, 2007. Springer Science & Business Media.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece, 2009. ACL.

Artem Sokolov, Laura Jehl, Felix Hieber, and Stefan Riezler. Boosting Cross-Language Retrieval by Learning Bilingual Phrase Associations from Relevance Rankings. In *Proceedings of the Conference on Empirical Methods in NLP*, Seattle, USA, 2013.

Artem Sokolov, Felix Hieber, and Stefan Riezler. Learning to Translate Queries for CLIR. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1179–1182, New York, USA, 2014.

Amanda Spink, Dietmar Wolfram, Major BJ Jansen, and Tefko Saracevic. Searching the Web: The Public and their Queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 2006. European Language Resources Association (ELRA).

Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, CO, USA, 2002.

Milan Straka and Jana Straková. 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*, pages 88–99, Vancouver, Canada, August 2017. ACL.

Trevor Strohman, Donald Metzler, Howard Turtle, and Bruce Croft. Indri: A Language Model-based Search Engine for Complex Queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6, McLean, VA, 2005.

Ying Sun and Paul B. Kantor. Cross-Evaluation: A New Model for Information System Evaluation. *Journal of American Society of Information Science and Technology*, 57(5):614–628, 2006.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, and et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. Springer, Berlin, Germany, 2013.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Tuomas Talvensaari. Effects of Aligned Corpus Quality and Size in Corpus-Based CLIR. In *Advances in Information Retrieval*, pages 114–125. Springer, 2008.

Gregory Tassey, Brent R Rowe, Dallas W Wood, Albert N Link, and Diglio A Simoni. Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. *National Institute of Standards and Technology, Gaithersburg, Maryland*, 2010.

Goutham Tholpadi, Chiranjib Bhattacharyya, and Shirish Shevade. Corpus-based translation induction in indian languages using auxiliary language corpora from wikipedia. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(3):20:1–20:25, March 2017.

Jörg Tiedemann. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248, Borovets, Bulgaria, 2009. John Benjamins.

Ferhan Ture and Elizabeth Boschee. 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*, pages 589–599, Qatar, 2014.

Andrew Turpin and Falk Scholer. User Performance versus Precision Measures for Simple Search Tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 11–18, New York, NY, USA, 2006. ACM.

Zdeňka Urešová, Jan Hajič, Pavel Pecina, and Ondřej Dušek. Multilingual Test Sets for Machine Translation of Search Queries for Cross-Lingual Information Retrieval in the Medical Domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

Zdeňka Urešová, Jan Hajič, Pavel Pecina, and Ondřej Dušek. Multilingual test sets for machine translation of search queries for cross-lingual information retrieval in the medical domain. In *Proceedings of LREC'14*, pages 3244–3247, Reykjavik, Iceland, 2014. ERLA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

Ellen M. Voorhees and Donna Harman. Overview of the Seventh Text REtrieval conference TREC-7. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 1–24, Gaithersburg, US, 1998. NIST.

Ellen M Voorhees and Donna Harman. Overview of the Sixth Text Retrieval Conference (TREC-6). *Information Processing & Management*, 36(1):3–35, 2000a.

Ellen M Voorhees and Donna Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 1–24, Gaithersburg, US, 2000b. NIST.

Katharina Wäschle and Stefan Riezler. Analyzing parallelism and domain similarities in the marec patent corpus. In Michail Salampasis and Birger Larsen, editors, *Multidisciplinary Information Retrieval*, pages 12–27, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

Ryen W White, Douglas W Oard, Gareth JF Jones, Dagobert Soergel, and Xiaoli Huang. Overview of the clef-2005 cross-language speech retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 744–759, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

F Wissbrock. Information Need Assessment in Information Retrieval; Beyond Lists and Queries. In *Proceedings of the 27th German Conference on Artificial Intelligence*, pages 77–68, 2004.

Theodore B. Wright, David Ball, and William Hersh. Query Expansion using MeSH Terms for Dataset Retrieval: OHSU at the bioCADDIE 2016 Dataset Retrieval Challenge. *Database: The Journal of Biological Databases and Curation*, 2017, 2017.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, 2016.

Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn, and Kevin Duh. Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 12–20, 2019.

Shipeng Yu, Deng Cai, Ji-Rong Wen, and Wei-Ying Ma. Improving Pseudo-Relevance Feedback in Web Information Retrieval Using Web page Segmentation. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, page 11–18, New York, NY, USA, 2003. Association for Computing Machinery.

Hamed Zamani and Bruce Croft. Embedding-based Query Language Models. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pages 147–156, New York, NY, USA, 2016. ACM.

Hamed Zamani and Bruce Croft. Relevance-based Word Embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 505–514, New York, NY, USA, 2017. ACM.

Qing Zeng, Sandra Kogan, Nachman Ash, Robert A Greenes, and Aziz A Boxwala. Characteristics of Consumer Terminology for Health Information Retrieval. *Methods of Information in Medicine*, 41(04):289–298, 2002.

Chengxiang Zhai and John Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. Integrating and Evaluating Neural Word Embeddings in Information Retrieval. In *Proceedings of the 20th Australasian Document Computing Symposium*, page 12, Stroudsburg, PA, USA, 2015. ACL.