Adaptation of machine translation for multilingual information retrieval in the medical domain

Pavel Pecina[a,*], Ondřej Dušek[a], Lorraine Goeuriot[b], Jan Hajič[a], Jaroslava Hlaváčová[a], Gareth J. F. Jones[b], Liadh Kelly[b], Johannes Leveling[b], David Mareček[a], Michal Novák[a], Martin Popel[a], Rudolf Rosa[a], Aleš Tamchyna[a], Zdeňka Urešová[a]

[a] *Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Malostranské nám. 25, 118 00 Prague 1, Czech Republic*

[b] *CNGL Centre for Global Intelligent Content, School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland*

Abstract

Objective. We investigate machine translation (MT) of user search queries in the context of cross-lingual information retrieval (IR) in the medical domain. The main focus is on techniques to adapt MT to increase translation quality; however, we also explore MT adaptation to improve effectiveness of cross-lingual IR.

Methods and Data. Our MT system is Moses, a state-of-the-art phrase-based statistical machine translation system. The IR system is based on the BM25 retrieval model implemented in the Lucene search engine. The MT techniques employed in this work include in-domain training and tuning, intelligent training data selection, optimization of phrase table configuration, compound splitting, and exploiting synonyms as translation variants. The IR methods include morphological normalization and using multiple translation variants for query expansion. The experiments are performed and thoroughly evaluated on three language pairs: Czech–English, German–English, and French–English. MT quality is evaluated on data sets created within the Khresmoi project and IR effectiveness is tested on the CLEF eHealth 2013 data sets.

Results. The search query translation results achieved in our experiments are outstanding – our systems outperform not only our strong baselines, but also Google Translate and Microsoft Bing Translator in direct comparison carried out on all the language pairs. The baseline BLEU scores increased from 26.59 to 41.45 for Czech–English, from 23.03 to 40.82 for German–English, and from 32.67 to 40.82 for French–English. This is a 55% improvement on average. In terms of the IR performance on this particular test collection, a significant improvement over the baseline is achieved only for French–English. For Czech–English and German–English, the increased MT quality does not lead to better IR results.

Conclusions. Most of the MT techniques employed in our experiments improve MT of medical search queries. Especially the intelligent training data selection proves to be very successful for domain adaptation of MT. Certain improvements are also obtained from German compound splitting on the source language side. Translation quality, however, does not appear to correlate with the IR performance – better translation does not necessarily yield better retrieval. We discuss in detail the contribution of the individual techniques and state-of-the-art features and provide future research directions.

Keywords: Statistical machine translation, Domain adaptation of statistical machine translation, Intelligent training data selection for machine translation, Compound splitting, Cross-language information retrieval, Medical query translation

1. Introduction

The development of health information search and retrieval techniques is an important research topic. Indeed, it has been found that almost 70% of search engine users in the US have conducted a web search for information about a specific disease or health problem [1]. Given that much medical content is written in the English language, research to date in the medical space has predominantly focused on monolingual English retrieval. However, given the large number of non-English speaking users of the Internet and the lack of content in their native language, support for them to search and utilize these English sources is required if the value of the information available on the Internet is to be fully realized [2]. In a recent study, Lopes and Ribeiro [3] assessed the effect of translating health queries for users with different levels of English language proficiency. Their results confirmed that users with even basic competence of English can benefit from a system which automatically retrieves English content based on a non-English query, or at least suggests English translations of the non-English queries.

Support for search of English language content by non-native English speakers is one of the major goals of the large in-

*Corresponding author
Email address: pecina@ufal.mff.cuni.cz (Pavel Pecina)

tegrated EU-funded Khresmoi project[1]. Among other goals, including joint text and image retrieval of radiodiagnostic records, the Khresmoi project aims to develop technology for transparent cross-lingual search of medical sources, for both professionals and laypeople, with the emphasis primarily on publicly available web sources. While a sophisticated search interface is being developed for the needs of medical professionals, the final application for the general public should be as simple as possible to operate and similar to the well-known interfaces of web search engines in use today with the addition of cross-lingual functionality.

The languages supported by the Khresmoi project are English (EN), Czech (CS), French (FR), and German (DE). Queries come from Czech, German, and French and are machine-translated to English. This reflects the real availability of data, which is predominantly available in English, and query translation needs of non-native speakers of English. Our focus in this paper is on the machine translation (MT) part of the cross-lingual search and retrieval task, while using a standard information retrieval (IR) technique for the search and retrieval part, in order to pinpoint contributions and problems with using MT for query translation from the three languages selected (Czech, German, and French) into English and its influence on the resulting quality of retrieved sets of documents.

Our MT system is based on Moses [4], a state-of-the-art statistical MT system. The IR experiments are performed using the Lucene search engine[2] on the CLEF eHealth 2013 dataset for the languages specified above, directed towards retrieving English documents only. Since MT is only an intermediate component of the whole system pipeline, we proceed in two steps. We first independently tune MT to produce the best possible translations of queries (Section 2) and then use various techniques to modify and expand the translated queries for improved IR performance (Section 3). The methods applied in Section 2 include: in-domain training and tuning, intelligent training data selection, optimization of phrase table configuration, exploiting synonyms to construct translation variants, and decompounding (splitting) of complex German words on the source language side, which normally appear as unknown words. For evaluation of translation quality itself, we use BLEU – the de facto standard automatic evaluation metric [5], which compares MT output against manual reference translation and accounts both for adequacy and fluency (word order) of the machine translation. We also report inverse position-independent word error rate [6], called PER, another automatic evaluation metric which compares words in the MT output and the reference translation but without taking the word order into account and thus might be better suited to application of MT in IR, where word order is often ignored. In selected experiments, the automatic evaluation is supplemented by manual assessment of the results performed by medical professionals.

The results of our MT for experiments for queries show that we are able to outperform results of Google Translate, the best freely available MT service on the web. We also find that using synonyms to enrich training data with translation variants does not improve the MT performance; however, decompounding of complex German words slightly improves the translation, at least according to BLEU. In Section 3, we evaluate query translation in a cross-lingual IR setting using standard methods on the CLEF eHealth 2013 Task 3 test collection. Here, despite achieving superior performance on the query MT task, as described in Section 2, we do not outperform the retrieval results obtained by using queries translated by Google Translate. In the last section, we perform a summary analysis of the overall results, the results of the individual techniques for improving MT performance and their integration into an IR system, and give suggestions for further work.

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

[2] http://lucene.apache.org/

2. Machine translation for medical queries

In this section, we describe the application of phrase-based statistical machine translation (SMT) to the translation of medical queries with the goal of producing accurate and fluent translations. This task differs from typical MT applications in two aspects: the *domain* and the *genre* of the input text. The domain, which reflects what the text is about, is very specific, characterized by a large and specialized vocabulary which does not occur in general texts. The genre, which indicates the general style, is also very distinctive. The input text is generally not in the traditional form of complete and coherent sentences, but rather in a form of short sequences of more or less independent terms. Such a situation requires application of special techniques to adapt the SMT system, including training data selection, model configuration, and parameter optimization. We also apply some standard additional methods to improve SMT quality in this task, including morphological normalization of the input text, splitting of complex compounds in German input, and exploitation of synonyms obtained from in-domain lexicons and dictionaries.

This section continues with a brief introduction to SMT and an overview of related research, followed by a detailed description of the data and the translation system used in this work. We then present details of the MT experiments carried out with details of results and a detailed analysis of our findings.

2.1. State-of-the-art and related work

In this section, we describe basic principles of phrase-based SMT (the most widely used paradigm in SMT) and review other related works to provide a complete background for our experiments.

2.1.1. Phrase-based statistical machine translation

In phrase-based SMT (e.g., the Moses system [4]), an input sentence is split into phrases (sequences of consecutive words) that are translated one-by-one and eventually reordered to produce the output translation. As there are typically many ways to split a sentence into phrases, and many possibilities for translation and reordering, the system searches for the best translation variant $\hat{\mathbf{e}}$ by maximizing the probability of the target sentence $\mathbf{e}$

given source sentence $\mathbf{f}$ in a log-linear combination of feature functions h_i with associated weights λ_i:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e}, \mathbf{f})$$

The computational complexity of this decoding approach is reduced by pruning the space of translation hypotheses using a heuristic beam-search algorithm [7] that explores the space represented as a graph by expanding the most promising nodes only. The feature functions include predictions of the *phrase translation model*, which captures probabilistic relations of source phrases to target phrases, thus ensuring that the individual phrases correspond to each other, the *target language model*, which estimates the fluency of the output sentence, the *reordering model* to capture different phrase order in the two languages, and *word penalty* to penalize translations that are too long or too short.

The phrase translation model and reordering model are trained using probabilistic word alignment [8] in parallel (i.e. bilingual pairs of) sentences. The target language model is trained on (typically) larger amounts of monolingual data. The feature weights λ_i are usually optimized using minimum error rate training (MERT) [9], a method which minimizes a given error measure (e.g., BLEU) [5] on a development set of parallel sentences using a coordinate ascent approach. The algorithm is guaranteed to converge to a local optimum only, but this usually leads to good results [10].

2.1.2. Domain adaptation

The quantity and also the quality of parallel and monolingual data is absolutely essential for SMT. Unless an SMT system is trained on data of the same nature (distribution) as the test data, it is not guaranteed to translate optimally. The most extensive and commonly available resources of SMT training data include legally required parallel parliamentary proceedings [11, 12], legislation documents [13], or news stories [14], which typically cover a number of different topics and are understood as general-domain data [see e.g., 15]. Training resources for specific domains are typically much scarcer, or not available at all. Therefore, special domain adaptation techniques are applied to adapt an SMT system trained on general-domain data to improve translation of text within a specific domain.

Much work on domain adaptation examines the usage of available in-domain data to directly improve in-domain performance of SMT. Some authors attempt to combine the predictions of two separate (in-domain and general-domain) translation models [16–19] or language models [20]. Wu and Wang [21] use in-domain data to improve word alignment in the training phase. Carpuat et al. [22] explore the possibility of using word sense disambiguation to discriminate between domains.

Other approaches concentrate on the acquisition of larger in-domain corpora. Some of them exploit existing general-domain corpora by selecting data that resemble the properties of in-domain data (e.g., using cross-entropy), thus building a larger *pseudo-in-domain* training corpus. This technique is used to adapt language models [23, 24] as well as translation models [25, 26] or their combination [27]. Similar approaches to domain adaptation are also applied in other tasks, e.g., automatic speech recognition [28].

Other possibilities for acquiring in-domain data are pursued as well. Translations of in-domain terms can be mined from comparable corpora, i.e. texts that are not strictly parallel but deal with the same topic [29, 30]. Bertoldi and Federico [31] exploit large amounts of in-domain monolingual data to create synthetic parallel training corpora. In-domain data for training can also be obtained automatically by crawling the web [15, 32]. In this work, we investigate methods combining the different kinds of data: general-domain, in-domain, and pseudo-in-domain to investigate what the optimal approach is.

2.1.3. Genre adaptation

While domain adaptation deals mainly with the problem of lexical coverage (lack of domain-specific terms and expressions), genre adaptation is mostly concerned with changes in syntax, which are very common and diverse in modern means of communication, such as SMS messages, Internet chats, discussion forums, and social network communication (e.g., unusual sentence length, ungrammatical constructions, missing punctuation, letter casing). Although most of domain adaptation techniques for SMT overviewed in the last section can be applied to genre adaptation as well, genre adaptation has not been studied extensively in SMT. However, some recent work has focused on SMT adaptation to specific genres targeted e.g., patents and patent applications [33], short text messages [34], user-generated forum content [35], public conference talks [36], and movie subtitles [37]. The methods used are generally similar to domain adaptation techniques.

One highly relevant study which explicitly deals with genre adaptation is by Nikoulina et al. [38]. They adapt the Moses SMT toolkit to translate queries for cross-lingual IR applied on the CLEF Ad Hoc TEL 2009 test collection of bibliography entries. They use two techniques: SMT tuning on genre-specific data and discriminative re-ranking of n-best lists optimizing retrieval effectiveness.

2.1.4. Statistical machine translation in the medical domain

Most work applying SMT for the medical domain relates to cross-lingual IR; a review of this work is contained in Section 3.1. In this section, we review the features of this work relating to SMT itself.

Eck et al. [39] employ an SMT system for the translation of dialogues between doctors and patients and show that a dictionary extracted from the Unified Medical Language System (UMLS) Metathesaurus [40] and its semantic type classification significantly improves translation quality from Spanish to English when applied to generalize the training data (measured by standard automatic evaluation metrics BLEU and NIST). Wu et al. [41] analyze MT quality on PubMed[3] titles and whether it is sufficient for patients. The conclusions are very positive especially for languages with large training resources (English,

[3] `http://www.ncbi.nlm.nih.gov/pubmed/`

Spanish, German) – the average fluency and content scores (based on human evaluation) are above four on a 5-point scale. In automatic evaluation, their systems substantially outperform Google Translate. However, the SMT systems are specifically trained, tuned, and tested on the domain of PubMed titles, and it is not evident how they would perform on other medical texts. Costa-jussà et al. [42] are less optimistic regarding the quality of SMT in the medical domain. They analyze and evaluate the quality of public web-based MT systems (such as Google Translate) and conclude that in both automatic and manual evaluation (reported for 7 language pairs), the performance of these systems is still not good enough to be used in daily routines of medical doctors in hospitals. Jimeno Yepes et al. [43] propose a method for obtaining in-domain parallel corpora from titles and abstracts of publications in the MEDLINE[4] database. The acquired corpora contain from 30,000 to 130,000 sentence pairs (depending on the language pair) and are reported to improve translation quality when used for SMT training, compared to a baseline trained on out-of-domain data.

2.1.5. Splitting German compound words

One of the source languages in our experiments is German. Written German tends to freely form compounds consisting of multiple regular words which relate to a complex concept. For example, the word *Raucherentwöhnungsprogramm* consists of three individual parts: *Raucher* (*smoker*), *entwöhnung* (*withdrawal*), and *programm* (*program*). However, standard tokenizers treat such compound words as single tokens since their parts are not separated by a space or hyphen. Such long expressions pose a specific problem for MT and IR. They increase the vocabulary size and are prone to be out-of-vocabulary, harming the overall output quality. In general, increasing the amount of training data cannot solve the problem because there is no upper bound of the number of possible compounds and new ones are created as needed.

Splitting compounds into separate tokens (which are in-vocabulary) and treating those as regular words usually reduces the problem – if such expressions are compositional, handling them word by word (e.g., in machine translation) can lead to better results (compared to the situation when the compounds are left untranslated). Chen [44] uses this approach in cross-lingual IR. He employs a monolingual dictionary containing uncompounded German words and splits compounds into the minimal number of components (words) which are present in this dictionary. If there are more possible decompositions, he selects the alternative with the highest probability using frequency analysis. Koehn and Knight [45] propose a simple corpus-based unsupervised method which maximizes the geometric mean of individual parts of the compounds. They show that when applied to phrase-based translation, such a simple method gives better BLEU scores than more complex methods using parallel corpora or part-of-speech tags. This is caused by the ability of phrase-based MT systems to group overaggressively split words back into phrases and translate them correctly. Popović et al. [46] compare this method with the linguistically oriented approach of Niessen and Ney [47] and conclude that both the methods yield similar improvements in DE–EN translation. Alfonseca et al. [48] combine several decompounding metrics originally proposed for German and show a substantial improvements for the total of six European languages, with respect to other state-of-the-art systems. They also show that a system trained on one language can be successfully used for splitting compounds in other languages too.

2.1.6. Exploiting synonyms in statistical machine translation

Medical terminology is very extensive; many of these terms are rare and therefore will not be present in parallel training data, even if very large amounts of material are available. In SMT, such terms will remain untranslated. In such a situation, an SMT system could benefit from additional data resources to find synonymous terms that the system is able to translate correctly. Similar ideas have already been explored for example by Wu and Zhou [49], Jones et al. [50], and Han et al. [51]; however, most work focuses on the acquisition of such resources rather than on their employment in applications.

Wu and Zhou [49] explore three common resources for automatic extraction of synonyms: monolingual dictionaries, parallel corpora, and large monolingual corpora. The authors show that a combination of all three of these types of resources yields the best results, with the parallel corpus being the most valuable of them.

A natural source of synonyms for the medical domain is the Unified Medical Language System (UMLS) [40], which implicitly defines synonym sets through its notion of concepts. A recent study of Griffon et al. [52] explores the extraction of synonyms from UMLS for a mildly related task of query expansion, concluding that proper employment of synonyms can lead to an improvement in performance. As noted, e.g., by Nakayama et al. [53], Wikipedia is another good source of information for automatic synonym extraction. Jones et al. [50] extract synonyms from the same source using the redirect pages for cultural heritage domain.

Han et al. [51] use synonyms extracted from dictionary-like data, namely English WordNet and Chinese Tongyicicilin, in SMT. However, they are trying to solve a rather different issue. They use synonyms of common words to detect sentences that are literal translations of each other, improving their training data by filtering out those that are not. We, on the other hand, need to extend our training data by using synonyms of words that are not very common.

2.2. Data description

This section provides an overview of the sources of data used in our SMT experiments. These sources are classified based on their relevance to the medical domain (in-domain vs. general-domain), nature of the data (dictionary vs. corpus), and language content (parallel vs. monolingual). The data sources used for training data selection are described in Sections 2.2.1–2.2.5. Statistics of the training data after cleaning in data preprocessing, as described in Section 2.2.6, are presented in Tables 1 and

[4] `http://www.nlm.nih.gov/pubs/factsheets/medline.html`

	Czech–English			German–English			French–English		
source	pairs	src	tgt	pairs	src	tgt	pairs	src	tgt
UMLS	70	218	224	86	303	317	80	301	256
DBpedia	69	141	151	306	685	718	375	895	893
EMEA	319	5,400	5,598	347	5,567	5,947	354	7,202	6,068
MuchMore	–	–	–	2	141	148	–	–	–
PatTR	–	–	–	1,594	55,070	58,458	–	–	–
COPPA	–	–	–	–	–	–	1,190	33,729	27,149
Com-Crawl	161	3,542	3,976	2,395	55,989	59,782	3,236	94,040	82,170
EuroParl	627	14,815	17,387	1,866	50,372	52,987	1,958	64,258	55,502
JRC-Acquis	593	18,030	20,737	773	24,347	26,233	781	29,762	25,979
News-Com	140	3,219	3,580	177	4,654	4,635	157	5,080	4,151
OJEU	1,859	44,573	50,176	1,715	41,933	44,851	2,031	64,589	54,776
DBpedia	148	333	360	681	1,562	1,712	745	1,979	1,942
CzEng	10,282	147,549	169,669	–	–	–	–	–	–
PatTR	–	–	–	7,979	290,184	321,412	–	–	–
Linguee	–	–	–	52	70	92	–	–	–
Hansard	–	–	–	–	–	–	837	21,622	18,042
MultiUN	–	–	–	–	–	–	10,267	375,337	310,649
COPPA	–	–	–	–	–	–	7,320	205,735	166,142

Table 1: Statistics of parallel training data sources including number of parallel sentence (pairs), source language (src) and target language (tgt, i.e., English) tokens. The first part includes in-domain dictionaries, the second part in-domain corpora, and the last part general-domain data (dictionaries and corpora). All figures are in thousands.

2. The data set used for development (system tuning) and testing (performance evaluation) are described in Section 2.2.7 and summarized in Table 3.

2.2.1. In-domain dictionary data

Parallel data in the form of a term-to-term dictionary is very valuable for terminology translation. Our main source of dictionary data is the UMLS Metathesaurus of health and biomedical vocabularies and standards [40]. The translation dictionaries for our experiments were constructed by selecting the UMLS concepts having translations in the respective languages (CS–EN, DE–EN, FR–EN). The number of dictionary entries ranges from 70,000 to 86,000 depending on the language pair (see Table 1).

Additional in-domain dictionaries were acquired from DBpedia [54], which contains structured information extracted from Wikipedia articles. We exploited the *owl:sameAs* links, which relate localized Wikipedia articles and their English equivalents through bijective inter-language links, to construct bilingual dictionary entries using the titles of selected articles and categories.

Since there is no straightforward strategy to identify Wikipedia content related to a particular domain, we employ the following heuristics.

For Czech, English, and French, selecting articles which transitively belong to the category *Medicine* covers almost the entire Wikipedia, which is not very useful. For instance, in the French Wikipedia, we can get from the category *Medicine* to *James Bond* in six steps: *Médecine* (*Medicine*) → *Histoire de la médecine* (*History of medicine*) → *Cas médical* (*Medical case*) → *Malade de fiction* (*Disease in fiction*) → *Drogué de fiction* (*Drug addict in fiction*) → *Fumeur de fiction* (*Smoker in fiction*) → *James Bond*. For German, however, the categorization of articles seems to be more strict and the result of this approach more precise and beneficial. Selecting subcategories and articles subordinated to the German categories *Biologie* (*Biology*) and *Gesundheit* (*Health*) and their equivalents in the other languages produced reasonably large dictionaries, more or less relevant to the domain of our interest.

As part of preprocessing, parenthesized texts were removed from the titles because they usually cannot be considered part of a dictionary entry. Moreover, their use in different languages is not consistent, as shown in the following examples of DE–EN pairs of terms: [*Krebs (Medizin)*, *Cancer*], [*Magnesiummangel*, *Magnesium deficiency (medicine)*].

Manual investigation of the dictionaries confirmed relatively high precision but lower recall. For this reason, the dictionaries were augmented by adding titles (and their equivalents in the other languages) from categories containing at least two already selected articles (based on the German categorization) or at least two articles for the UMLS concepts. For instance, the Czech UMLS contains concepts such as *Uhthoffův fenomén* (*Uhthoff's phenomenon*) and *Demence* (*Dementia*), but not *Abarognóza* (*Abarognosis*), *Agnózie* (*Agnosia*), or *Migréna* (*Migraine*). The latter concepts were added to the list based on their membership in the same category of Wikipedia, namely *Symptomy poruch nervové soustavy* (*Symptoms of nervous system diseases*), as the former ones mentioned above. This process resulted in 70,000 entries for CS–EN and more than 300,000 for DE–EN and FR–EN each. The dictionaries produced contain a certain amount of noise not directly relevant to medicine (e.g., names of persons and geographical locations), but this is generally not a problem for SMT as long as the dictionaries contain the domain-relevant material as well

(the non-relevant material, which does not typically occure in the input, is not used to construct the translation hypotheses and thus does not effect the system). A thorough evaluation of this approach will be the subject of further work.

2.2.2. In-domain parallel corpora

In-domain parallel corpora have the traditional form of a set of aligned sentences. In this study we use two established in-domain medical corpora. The EMEA corpus is an in-domain parallel corpus of documents from the European Medicines Agency, automatically processed and aligned on sentence level by Tiedemann [55]. It is publicly available for all the language pairs and comprises more than 300,000 sentence pairs for each language pair. The MuchMore Springer Corpus is a parallel corpus of approximately 6,000 DE–EN abstracts from medical journals published by Springer [56].

Two additional in-domain parallel data sets are extracted from patents and patent applications. The DE–EN set of 1.5 million sentence pairs is extracted from PatTR, a parallel corpus extracted from the MAREC patent collection [57]. The complete corpus contains more than 20 million DE–EN sentence pairs from all patent text sections. For the in-domain subset, we only consider text from titles, abstracts, and claims indicated to be from the medical domain (categories A61, C12N, and C12P). The FR–EN set is extracted in a similar way from the Corpus of Parallel Patent Applications (COPPA) provided by World Intellectual Property Organization [58], a total of 8 million sentence pairs from Patent Cooperation Treaty applications (titles and abstracts). The resulting in-domain subset consists of 1.2 million parallel sentences.

2.2.3. General-domain parallel data

In this work, we also exploit a wide variety of bilingual resources not explicitly associated with the medical domain. They cover a large number of genres and topics, and we denote them as general-domain data. All these data sets are publicly available and include both dictionary and corpus resources.

The parallel corpora exploited for all the language pairs include: Common Crawl mined from the public web crawl hosted on Amazon's Elastic Cloud [59], EuroParl version 6 extracted from the proceedings of the EU Parliament [11], JRC-Acquis Multilingual Parallel Corpus version 3.0 extracted from Acquis Communautaire, the total body of European Union law [13], the News Commentary corpus of news analysis from the Project Syndicate [14], and the OJEU corpus with texts from the Official Journal of the European Union including legislation documents, information notices, and public procurements, made available by the Apertium project [60]. We also make use of the dictionary data extracted from DBpedia and not identified as medical-domain, see Section 2.2.1.

In addition to the resources mentioned above, we employ CzEng 1.0 for the CS–EN experiments, a compilation of sentence-aligned parallel data from various sources: legislation, fiction, news articles, parallel web pages, movie subtitles, and technical documentation of software [61] (the Navajo section was excluded due to its low translation quality). The DE–EN general-domain data is supplemented by a subset of the PatTR corpus, sentence pairs extracted from the sections not identified as relevant to the medical domain, as described in Section 2.2.2, and translation pairs obtained from Linguee,[5] an online dictionary service. Additional FR–EN general-domain data comes from three sources: the Hansard Corpus extracted from proceedings of the Canadian Parliament [12], the MultiUN parallel corpus extracted from the United Nations website [62], and the COPPA corpus [58] – sentences from the sections not detected as medical-domain, see Section 2.2.2.

source	sentences	tokens
Cochrane	2,120	58,454
DrugBank	23	826
GREC	1	62
GENIA	18	557
FMA	150	884
UMLS	321	2,003
PIL	20	567
HON	44,285	1,145,384
MultiUN	14,077	422,008
WMT News	48,370	1,188,277
Gigaword	22,197	854,493

Table 2: Number of sentences and tokens in English monolingual data sources. All figures are in thousands.

2.2.4. In-domain monolingual corpora

Analogously to the parallel data, we also make use of monolingual data resources to train language models of English as the target language in all our experiments. The in-domain monolingual corpora include (cf. Table 2): the Cochrane database of reviews of primary research in human health care and health policy [63], DrugBank – a bioinformatics and cheminformatics resource describing drugs [64], Gene Regulation Event Corpus (GREC) – a semantically annotated English corpus of abstracts of biomedical texts [65], the GENIA corpus of biomedical literature compiled and annotated within the GENIA project [66], the Foundational Model of Anatomy Ontology (FMA) – a knowledge source for biomedical informatics concerned with symbolic representation of the phenotypic structure of the human body [67], English texts extracted from the UMLS Metathesaurus [40], the Patient Information Leaflet Corpus (PIL) – a collection of documents giving instructions to patients about their medication [69], and finally, a large set of texts extracted from HONcode-certified sites (HON) that have been identified by language-detection libraries [70, 71] to be English-language [72].

2.2.5. General-domain monolingual corpora

In addition to the in-domain monolingual corpora, we also use general-domain monolingual corpora including: the English part of the MultiUN corpus extracted from the United Nations Website [62], News articles from 2009–2012 provided for the Workshop on statistical Machine Translation (WMT)

[5] `http://www.linguee.de/`

shared translation task [14], and the English Gigaword containing newswire texts [73].

2.2.6. Data preprocessing

All the training data sets were preprocessed in order to get clean and reliable data without duplicities and noise. We discarded parallel sentences with an empty source and/or target side, removed non-UTF8 characters, and ignored duplicate sentences (or sentence pairs) on a corpus level. Parallel sentences containing more than 100 words on either side were removed (because our system cannot process longer sentences in training data). All the data (including training, development, and test sets) was tokenized, tagged for part-of-speech, and lemmatized using the Treex NLP framework [74].

In morphological tagging, all tokens were processed using a part-of-speech tagger that provides each token with an appropriate part-of-speech tag, which also includes various morphological properties. We applied the Morče tagger [75] for Czech and English and TreeTagger [76] for German and French. In addition, the Czech tagger employed a morphological analyzer [77] to prune impossible part-of-speech tags for a given token. During *lemmatization*, all tokens are also assigned a *lemma* – the base form of the word, e.g., infinitive for verbs and nominative singular form for nouns. For Czech, German, and French, lemmatization is done jointly with tagging, i.e., the tagger selects the appropriate lemma along with the word form. For English, we used a rule-based lemmatizer implemented by Popel and Žabokrtský [78].

Table 1 shows statistics for the parallel data sources and Table 2 presents similar statistics for the monolingual data sources, both after the cleaning and preprocessing steps described in this section have been performed.

2.2.7. Development and test data

For parameter optimization and evaluation of translation quality in our experiments, we employ data sets from three sources; details are shown in Table 3. For the general domain (denoted as *gen*), we use the test sets provided for the WMT translation task 2012 and 2013 [14] as development and test sets, respectively. This data contains sentences extracted from news stories in various languages (including Czech, German, English, and French) and manually translated into all other languages.

For the medical domain (denoted as *med*), we exploit a translation memory produced by the European Centre for Disease Prevention and Control (ECDC)[6]. This is a collection of sentences mostly on health-related topics and their professional translations into 25 languages. We randomly split the set of sentences having translations in the four relevant languages into development and test sets.

The primary development and test sets used in this study consist of user queries from the medical domain, referred to as *query*. These queries were originally in English, sampled from two sources: 50% from the general-public query logs provided by the Health On The Net (HON) Foundation [79] and 50% from the Trip database containing queries by medical professionals [80]. We hired human translators (not necessarily native speakers but fluent in the relevant languages) to manually translate all the queries into Czech, German, and French, and then medical experts to verify the accuracy of the translations. The resulting sets were randomly split into development and test sets and made publicly available via the LINDAT/Clarin repository.[7] Some examples of English language general-public queries are: *diabetic ulcer*; *cancer breast*; *disease*; *access*; *recuperation*. Examples of queries by medical professionals are: *nsaids osteoarthritis*; *meningitis and penicillin*; *asthma children*; *common cold*; *prison dermatology*.

While the *gen* and *med* data sets contain complete sentences extracted from longer texts (documents), the *query* sets consist of short expressions used as real user search queries. This difference is evident from the average length, which is about 23 words per sentence on the English side for the *gen* sets, 17 for the *med* sets, and only slightly above 2 for the *query* sets – this is comparable to the observation that user queries to search engines usually consist of 2–3 words [81]. Detailed statistics of the individual data sets are provided in Table 3.

2.3. System description

Our translation system is based on Moses [4], an open-source phrase-based SMT project providing a complete set of tools for training, parameter optimization (tuning), and decoding (translation). In this section, we provide technical details of our setup.

Word alignment is computed on *pseudo-stems*, words trimmed to 5 characters, using *fast_align* [82], which features competitive results of end-to-end MT evaluation and is faster compared to the traditional tools [e.g., 8]. The resulting alignments are symmetrized by the *grow-diag-final-and* heuristic and phrases are extracted using the standard tools bundled with Moses, with the length limit set to 7 words. All dictionary data in our experiments is used the same way as regular parallel data (with alignment and phrase pair extraction). Language models of order 5 are estimated using SRILM (Stanford Research Institute Language Modeling toolkit) [83] with modified Kneser-Ney smoothing [84]. KenLM (Kenneth Heafield's Language Model toolkit) [85] is used in decoding for querying the models. We do not employ lexicalized reordering [86] and rely on the standard distortion penalty feature instead, similarly to the state-of-the-art system of Bojar et al. [87]. MERT [9] is used for tuning the model parameters towards BLEU [5] on the development sets of parallel sentences.

2.4. Machine translation experiments

This section presents our experimental study of the adaptation of SMT towards the medical domain. First, we describe baseline MT systems trained and tuned on general-domain data and evaluate their performance. Then, we adapt the systems to the medical domain by exploiting various training resources and optimizing the configuration of the SMT system and its

[6] http://ipsc.jrc.ec.europa.eu/

[7] http://hdl.handle.net/11858/00-097C-0000-0022-D9BF-5

domain	source	type	pairs	Czech	German	French	English	len_{EN}
gen	WMT	*dev*	3,003	65,657	73,722	86,977	74,383	24.77
		test	3,000	57,411	64,555	79,013	66,222	22.07
med	ECDC	*dev*	751	11,654	12,605	15,523	12,346	16.44
		test	1,400	23,073	24,985	30,968	24,441	17.46
query	Khresmoi	*dev*	508	1,128	1,041	1,335	1,084	2.13
		test	1,000	2,121	1,951	2,490	2,067	2.07

Table 3: Statistics of development and test data sets including domain (*gen* – general, *med* – medical text, *query* – medical query), number of parallel sentences (pairs), total number of tokens per language, and average number of tokens on English side (len_{EN}).

parameters. Further, we employ more advanced linguistic preprocessing of the training data, such as morphological normalization and decompounding, and analyze their effect on system performance. Finally, we attempt to exploit a terminological thesaurus to improve translation quality in this very specific domain. Each technique is described in a separate subsection and provided with detailed analysis of its contribution. The overall results are then compared and discussed in Sections 2.5 and 2.6 and summarized in Section 2.7.

Our experiments are carried out on the CS–EN, DE–EN, and FR–EN language pairs using Eman, an experiment manager by Bojar and Tamchyna [88]. We evaluate our systems using BLEU [5] and PER (position-independent word error rate) [6]. PER is similar to word error rate known as the Levenshtein distance [89] computed on words (not characters), but it does not penalize word reordering; this might better fit IR systems which typically ignore query word order.

BLEU scores are reported as percentage and PER is reported as $100 \times (1-\text{PER})$, so that both metrics are in the range 0–100 where higher scores indicate better translations. In the tables presenting results in this section, the best scores for each language pair are marked with a $\star$ symbol and those which are statistically indistinguishable from the best ones are typed in bold. To test statistical significance, we use paired bootstrap resampling for BLEU [90] and the standard paired t-test for PER, both with $p < 0.05$. Results of selected systems are also compared in Section 2.5 using human expert evaluation where the best score for each language pair is again marked with a $\star$.

2.4.1. Baseline translation systems

Current state-of-the-art SMT systems for commonly spoken language pairs and general domains are trained on data comprising millions of parallel sentences and tens of millions of monolingual sentences [14]. We decided to limit the amount of data in each experiment to 10 million parallel sentence pairs and 30 million monolingual sentences. These numbers are quite comparable to current state-of-the-art SMT systems and define a strong baseline – a larger training data set would probably not bring substantial improvement of translation quality, especially when the available in-domain data is much smaller and additional data would have to be taken from out-of-domain sources.

To simulate a typical real-world scenario where no in-domain data is available for training an SMT system for a specific domain, we train our baseline systems using general-domain resources only. For each language pair, the baseline system is trained on a mixture of 10 million parallel sentences randomly taken from all general-domain sources, outlined in Section 2.2.3, and 30 million monolingual sentences sampled from all general-domain monolingual sources, outlined in Section 2.2.5, and tuned on 3,003 sentence pairs of general-domain (*gen*) development data, as described in Section 2.2.7. These configurations of parallel and monolingual training data are referred to as *P0* and *M0*, respectively. The performance of the resulting systems measured on various test sets is presented in Table 4.

The BLEU scores for the *gen* domain range from 24.13 to 29.62 depending on the language pair. Such scores are comparable to the state-of-the-art results reported recently on the same data sets [91]. The scores measured on the *med* domain are consistently higher by about 2 BLEU points (27.06–31.45). This might seem unexpected, since we are translating domain-specific data using a general-domain system, but the data used for training the baseline systems is large (10 million sentence pairs), taken at random from various sources and the chances that it provides translation evidence for some medical terms are high. Moreover, the *med* test sentences are much shorter, on average (about 17 words for *med* vs. 23 words for *gen*, see Table 3), and MT of shorter sentences is typically easier. Note also the wider confidence intervals (about ± 1.25 BLEU for *med* vs. ± 0.65 BLEU for *gen*), which are mainly caused by the smaller data size (in terms of number of sentence and words), see Table 3. A similar effect is evident from the results obtained on the main *query* test sets. Although domain-specific terminology is very frequent in this test data, the fact that the queries are very short (only 2 words on average, see Table 3) gives rise to relatively high BLEU scores (29.50–37.84) and even wider confidence intervals (about ± 4.48 BLEU, see Table 4).

Note. As a direct consequence of such wide confidence intervals, the tests of statistical significance in the experiments presented later in this section show insignificant differences even for results with relatively high differences in BLEU. Therefore, it is difficult to confirm a positive contribution of some methods despite the substantial increase of BLEU.

2.4.2. System tuning with in-domain data

Optimization of the SMT model parameters has been shown to have a substantial impact on model performance [e.g., 9]. In order to obtain optimal translations, it is necessary to tune the system on data of a similar nature to the data on which it will be applied [92]. The effect of varying development data domain (*gen*, *med*, *query*) in our experiments is illustrated in

test	Czech–English		German–English		French–English	
	BLEU	1-PER	BLEU	1-PER	BLEU	1-PER
gen	25.76 ± 0.63	61.04 ± 0.50	24.13 ± 0.60	60.24 ± 0.55	29.62 ± 0.71	63.06 ± 0.66
med	28.48 ± 1.25	57.76 ± 1.43	27.06 ± 1.21	58.67 ± 1.31	31.45 ± 1.29	61.91 ± 1.35
query	26.59 ± 4.42	55.25 ± 3.38	23.03 ± 3.87	54.76 ± 3.52	32.67 ± 5.17	65.73 ± 3.23

Table 4: Performance of the baseline systems trained and tuned on general-domain data and tested on general (*gen*), medical (*med*), and *query* test sets. The scores are provided with empirical 95% confidence intervals.

config	dev	test	Czech–English		German–English		French–English	
			BLEU	1-PER	BLEU	1-PER	BLEU	1-PER
P0 M0	*gen*	*query*	26.59	55.25	23.03	54.76	32.67	65.73
P0 M0	*med*	*query*	30.84	60.76	**27.44**	**59.78**	35.60	**68.87**
P0 M0	*query*	*query*	⋆**35.73**	⋆**66.21**	⋆**29.50**	⋆**60.40**	⋆**37.84**	⋆**71.78**

Table 5: The baseline systems tested on medical queries and tuned on development sets of different domains.

Table 5. Tuning on in-domain data (*med* and *query*) gives impressive improvements of translation quality measured by both BLEU and PER. Using the *med* development sets improves the baseline BLEU scores by 3.86 absolute on average and tuning on the *query* sets boosts the scores by an additional 3.06 BLEU absolute. Given the fact that the only changes in the SMT systems are the weights of the feature functions, this improvement is remarkable. All subsequent experiments are thus tuned on the *query* sets.

2.4.3. System training with parallel in-domain data

Training resources for the domain of medicine are not as scarce as for other specific domains (see Section 2.2) and provide enough data to train a complete SMT system. Parallel in-domain data include both dictionaries and corpora (see Table 6). We train three systems using different combinations of these resources to assess their relative contribution: one solely based on in-domain dictionaries described in Section 2.2.1 (denoted as *P1*), one based only on in-domain corpora described in Section 2.2.2 (denoted as *P2*), and one trained on a mixture of both the in-domain dictionaries and corpora (denoted as *P3*). Table 6 presents results of the three systems compared with the systems trained on a random general-domain sample and tuned on the *query* development sets (*P0*). The monolingual training data is the same as in the previous experiments (*M0*).

We can conclude that the systems trained on both types of in-domain resources (*P3 M0*) outperform the general-domain ones. The in-domain training data better covers the test set vocabulary and achieves higher scores with less training data, though the improvement in BLEU is statistically significant only for DE–EN. In terms of PER, the improvement is statistically significant also for FR–EN but not for BLEU – despite the large difference of 3.23 points absolute (this is a typical example of the situation discussed in the note in Section 2.4.1). The inconclusive result for CS–EN is caused by the limited availability of in-domain parallel data for Czech and English, especially by the absence of patent data. The other translation pairs benefit from larger in-domain training data, which reduces the out-of-vocabulary problem especially for DE–EN, caused by linguistic properties of German. For DE–EN and FR–EN, the relative contribution of in-domain dictionaries vs. in-domain corpora is higher for the latter (compare *P1 M0* vs. *P2 M0*), but this is caused by the larger amounts of training material available in the corpora. For CS–EN, where in-domain corpora data is available in lesser quantities, the BLEU scores for *P1* and *P2* are equal (29.00). Given the much smaller size of the dictionaries (see Table 8), results of *P1* are notable for all language pairs.

cfg	Czech–English		German–English		French–English	
P0	168,005	192,504	299,977	327,208	315,543	263,212
P1	358	375	988	1,034	1,196	1,148
P2	5,400	5,597	60,778	64,554	40,985	33,260
P3	5,758	5,972	61,766	65,588	42,181	34,408
P4	166,075	189,664	291,719	316,407	294,636	244,985
P5	177,723	203,078	332,578	362,493	285,619	236,873
P6	177,865	203,275	335,200	365,489	289,644	240,330
P7	179,523	205,882	346,202	379,732	312,685	260,132

Table 8: Number of tokens (in thousands) in each side of all configurations (cfg) of parallel training data.

2.4.4. Intelligent selection of training data

In the previous experiments, the in-domain data was selected based on explicit information about its sources (e.g., the European Medicines Agency). Such information, however, might not be completely reliable (not every piece of data from the relevant providers is expected to be related to medicine). At the same time, in-domain data can also appear in other resources (not explicitly known to be in-domain). In this section, we construct the training data for SMT from sentences found to be really similar to the language of the medical domain.

We follow an approach originally proposed for selection of monolingual sentences for language modeling [24] and its modification applied to selection of parallel sentences [26]. This technique assumes two language models for sentence scoring, one trained on (true) in-domain text and one trained on (any) general-domain text in the same language (e.g., English). Each sentence in a pool of all available data (regardless of domain) is then scored by a difference of its cross-perplexity given the in-

config	dict	med	gen	Czech–English		German–English		French–English	
				BLEU	1-PER	BLEU	1-PER	BLEU	1-PER
P0 M0			○	**35.73**	**66.21**	29.50	60.40	**37.84**	71.78
P1 M0	●			29.00	60.58	28.27	58.41	34.95	71.90
P2 M0		●		29.00	60.51	32.87	61.15	36.21	74.00
P3 M0	●	●		⋆**36.35**	⋆**67.12**	⋆**36.64**	⋆**64.87**	⋆**41.07**	⋆**77.12**

Table 6: Performance of the systems trained on various combinations of data from in-domain dictionaries (dict), in-domain corpora (med), and general-domain data (gen) selected as complete (●) and random (○) samples.

config	dict	med	gen	Czech–English		German–English		French–English	
				BLEU	1-PER	BLEU	1-PER	BLEU	1-PER
P3 M0	●	●		**36.35**	67.12	⋆**36.64**	⋆**64.87**	**41.07**	⋆**77.12**
P4 M0	●	●	○	33.13	**68.60**	32.87	**63.00**	⋆**43.00**	**76.41**
P5 M0	●	●	⦸	**35.53**	⋆**69.36**	**34.78**	**63.20**	**41.83**	**75.71**
P6 M0	●	⦸	⦸	⋆**36.65**	**68.23**	**34.67**	**64.03**	**42.74**	**76.47**
P7 M0			⦸	32.27	64.80	30.18	60.08	39.26	74.63

Table 7: Performance of the systems trained on combinations of data from in-domain dictionaries (dict) and corpora (med), and general-domain data (gen) selected as complete (●), random (○), or intelligent (⦸) samples.

domain language model and cross-perplexity given the general-domain language model (in this order). Sentences with the lowest scores (i.e., those more similar to the language of the specific domain) are selected as *pseudo-in-domain* data and used for training. The two language models for sentence scoring are trained with a restricted vocabulary extracted from the in-domain training data as words occurring at least twice (singletons and other words are treated as out-of-vocabulary).

Motivated by Moore and Lewis [24], Axelrod et al. [26] apply this approach to selection of parallel data. They scored both the source and target language sides of parallel sentence (independently) and define the selection criterion as the average of the source side score and the target side score.

In our experiments, we apply this technique to select both monolingual data for language models and parallel data for translation models. Selection of parallel data is based on the target language (English) only – so we only need two scoring models for all experiments (both English): the in-domain one is trained on the HON data set and the general-domain one on the WMT News data (the resources are described in Section 2.2.4). Compared to the approach of Moore and Lewis [24] and Axelrod et al. [26], we prune the model vocabulary more aggressively – we discard not only the singletons, but also all words with non-Latin characters, which helps clean the models from noise introduced by the automatic process of data acquisition by web crawling.

Parallel data selection. For parallel data selection, we experiment with three configurations which differ in the proportion of in-domain and pseudo-in-domain material in our training data, always summing up to 10 million sentence pairs so the training data size is comparable to other experiments. The first system is trained on all data from the in-domain sources (dictionaries and corpora) plus the intelligent selection of pseudo-in-domain data from the general-domain sources (*P5*). The second system is trained on the in-domain dictionaries plus pseudo-in-domain selection from the in-domain corpora and the general-domain sources (*P6*), and finally, the third system is based on intelligent selection of data from general-domain data only (*P7*). The effectiveness of this technique is compared to the systems trained on in-domain data only (*P3*) and with a trivial baseline which adds random selection from the general-domain data so the total sum of parallel training data is 10 million sentence pairs (*P4*). The language model and development data are the same as in the previous experiments.

The results for all configurations are presented in Table 7, with statistics of the corresponding parallel data in Table 8. The results and the observed trends differ depending on language pair. For CS–EN and DE–EN, using an additional randomly selected general-domain data (*P4*) does not help and performance of the systems decreases (compared to *P3*). For FR–EN, however, the BLEU score increases and the system outperforms all other configurations. This can probably be explained as a coincidence due to the random sampling as the result is not significantly better than *P3–P6*. Intelligent data selection gives slightly better results when applied on all corpus data (*P6*) regardless of domain (compare with *P5*) and will be used in our further experiments. Using pseudo-in-domain data exclusively from the general domain (*P7*) outperforms the dictionary-based systems (*P1*) only and confirms the importance of true in-domain corpora (*P3–P6*) for training a domain-specific SMT system.

Monolingual data selection. We perform experiments with five different configurations of monolingual training data for language modeling (*M0–M1*), always using the same parallel training data (*P6*), see Table 9.

So far, all the experiments presented employed the baseline language model trained on a random sample of 30 million sentences from the general-domain data (*M0*, see Section 2.4.1). Substituting this data with an intelligent sample of the same size and from the same source (*M1*) improves BLEU for CS–EN

config	med	gen	Czech–English		German–English		French–English		size
			BLEU	1-PER	BLEU	1-PER	BLEU	1-PER	
P6 M0		○	36.65	68.23	34.67	**64.03**	**42.74**	**76.47**	873,654
P6 M1		⦸	37.82	**68.88**	33.62	**65.10**	41.84	**76.72**	925,962
P6 M2	○		37.26	69.65	37.07	**64.09**	**44.54**	⋆**77.30**	772,532
P6 M3	⦸		36.58	**69.25**	37.92	**64.40**	⋆**45.20**	**77.15**	828,898
P6 M4	⦸	⦸	⋆**41.45**	⋆**71.61**	⋆**40.65**	⋆**65.43**	**44.50**	**77.24**	825,617

Table 9: Results for MT systems trained with varying monolingual data taken from in-domain (med) and general-domain (gen) corpora as random (○) or intelligent (⦸) samples. Size refers to the size in thousands of tokens.

domain	source	*M0*	*M1*	*M2*	*M3*	*M4*
med	Cochrane	–	–	4.52	4.32	4.22
	DrugBank	–	–	0.05	0.07	0.07
	FMA	–	–	0.32	0.42	0.42
	GENIA	–	–	0.04	0.06	0.06
	GREC	–	–	<0.01	<0.01	<0.01
	UMLS	–	–	0.68	1.00	1.00
	PIL	–	–	0.04	0.04	0.04
	HON	–	–	94.34	94.08	92.44
gen	MultiUN	16.63	41.37	–	–	1.50
	WMT News	57.15	31.84	–	–	0.12
	Gigaword	26.23	26.79	–	–	0.13

Table 10: Distribution of the monolingual data sources in the language model data configurations (*M0–M5*).

only. For DE–EN and FR–EN, the BLEU scores drop – the general-domain sources probably do not contain enough material related to the medical domain. Taking language model training data from in-domain sources (*M2–M4*) turns out to be more beneficial – for CS–EN, the difference is not significant, but for DE–EN and FR–EN, the improvement is about 3 BLEU points for the random sample (*M2*) and 4 BLEU points for the intelligent selection (*M3*). The most efficient configuration is the intelligent selection from all sources (*M4*). This approach brings an additional 5 BLEU points for CS–EN and 3 BLEU points for DE–EN. In the case of FR–EN, we observe a slight degradation, but the difference is not statistically significant.

The distribution of the monolingual data sources in the individual configurations is illustrated in Table 10. It is evident that the general-domain MultiUN corpus is more relevant to the medical domain than WMT News and Gigaword (compare *M0* and *M1*), and the distribution of *M2* and *M3* does not differ a lot. Unsurprisingly, the best-performing configuration (*M4*) contains most material from the HON data and other in-domain sources, but the general-domain corpora are present too and their contribution to translation quality is quite substantial (compare *M3* and *M4* in Table 9, especially for CS–EN and DE–EN). Naturally, the *M4* language model is used in further experiments.

2.4.5. Optimization of phrase table configuration

In the previous experiments with data selection, the MT system has no information about whether a particular translation option comes from in-domain or out-of-domain data. To make this information explicitly available, we follow the approach of Koehn and Schroeder [20] and train two independent translation models (phrase tables). The first experiment is based on the *P5* configuration of parallel data: one phrase table is trained using in-domain sections of the training data (dictionaries and medical corpora) and the second using data selected from the general domain. In the second experiment, based on *P6*, the first table is trained on the medical dictionaries only and the second one on the intelligent selection from the in-domain and general-domain corpora. We use *M4* (the best-performing option) as the source for monolingual data in both experiments to make our results comparable with the best systems reported so far.

Each phrase table has its own set of parameters in the log-linear combination. Setting their weights during model optimization allows the MT system to balance its confidence in the individual models (i.e., out-of-domain translation options can be down-weighted). During translation, both phrase tables are used simultaneously. By splitting the data, we impose a hard division which may not always be advantageous: while some data from the general domain may be very similar to the test set, it is weighted identically as sentences which are entirely out-of-domain. The division into two tables also makes the data sparser: since no statistics are shared between the corpora, both phrase probabilities and lexical smoothing can become less accurate. Table 11 shows that indeed we are not able to improve translation performance using this technique in either case.

2.4.6. Morphological normalization in the source language

Several previous studies have shown that translation from morphologically rich languages can be improved by performing some kind of morphological normalization on the source language side. Two of the three source languages in our experiments can be considered quite complex in terms of morphological variability: Czech with its multitude of inflection patterns in nouns, adjectives, and verbs, and German with similar grammar complexity emphasized by frequent usage of word compounds.

Morphological normalization reduces vocabulary size by mapping word forms to their morphological classes. It can be realized in various ways ranging from simple heuristics stripping suffixes, advanced rule-based [93] or stochastic [94] stemmers, to lemmatization substituting words by their linguistic base forms. We investigate whether morphological normalization performed on the source language side improves translation quality in our specific domain of medical queries. We take

config	p-tables	Czech–English		German–English		French–English	
		BLEU	1-PER	BLEU	1-PER	BLEU	1-PER
P5 M4	*mix*	**38.69**	**69.68**	**40.00**	**65.44**	**43.55**	**77.88**
P5 M4 T2	*med+gen*	**39.71**	**70.40**	36.89	⋆**66.11**	**43.08**	**76.15**
P6 M4	*mix*	⋆**41.45**	⋆**71.61**	⋆**40.65**	**65.43**	⋆**44.50**	⋆**77.24**
P6 M4 T2	*med+gen*	37.32	**69.61**	37.46	**66.80**	**42.95**	**76.40**

Table 11: Performance of the systems exploiting separate translation tables (*T2*) for in-domain and out-of-domain data compared with the single phrase table configurations.

config	method	Czech–English		German–English		French–English	
		BLEU	1-PER	BLEU	1-PER	BLEU	1-PER
P6 M4	*none*	⋆**41.45**	⋆**71.61**	⋆**40.65**	⋆**65.43**	⋆**44.50**	⋆**77.24**
P6 M4 N1	*prefix*	35.66	58.65	29.46	49.51	34.32	62.05
P6 M4 N2	*snowball*	29.62	65.53	35.91	62.49	38.09	70.23
P6 M4 N3	*lemma*	36.87	67.19	**40.53**	65.30	35.27	73.18

Table 12: The effect of morphological normalization performed on the source side. In *N1*, the tokens are cut after the first 5 characters, *N2* employs the Snowball stemmer, and *N3* is based on proper lemmatization.

config	splits	OOV	% OOV	BLEU	1-PER
P3 M0	0	133	6.8	36.64	**64.87**
P3 M0 C3	159	79	3.8	**37.63**	**66.94**
P6 M4	0	100	5.1	**40.65**	**65.43**
P6 M4 C3	159	61	2.9	⋆**40.82**	67.75
P6 M4 C6	156	47	2.3	**40.68**	⋆**67.75**

Table 13: Performance of DE–EN systems with splitting German compounds. The splitting models *C3* and *C6* are trained on the same parallel data as *P3* and *P6*, respectively. OOV refers to the absolute and relative (%) number of out-of-vocabulary words, and splits refers to the number of splits (in 1,951 words).

the *P6 M4* configuration as a baseline and apply three stemming methods on the source language side of the parallel training data: *N1* trimming words to 5 characters (used also for word alignment, Section 2.3), *N2* realized by Snowball stemmer [95], and *N3* based on lemmatization by the Treex NLP framework [74]. As shown in Table 12, the results are not affirmative. None of the three methods leads to better translation quality in any translation direction. Moreover, the observed degradation of translation quality is statistically significant in almost all the experiments, both in terms of BLEU and PER. The only exception is the DE–EN experiment with lemmatization, where the achieved BLEU score is lower but the difference is not statistically significant. We explain these results by the very specific nature of our domain and genre. Most queries are formulated using medical terms in their base forms which also occur frequently in the training data (e.g., in the in-domain dictionaries). Morphological normalization then increases translation ambiguity (number of translation variants), making it difficult to resolve (select the best one) in the very limited context of a typical query, which is in our case about 2 words, on average (see Table 3).

2.4.7. Splitting German compound words

In Section 2.1.5, we outlined the issue of German compounds, which increases the vocabulary size and leads to out-of-vocabulary problems in MT (especially in specific domains with rich terminology, such as medicine). The technique proposed to reduce the problem is based on splitting the compounds into components which are then treated as regular (in-vocabulary) words.

In this work, we employ a simple unsupervised frequency-based method for splitting compound words introduced by Koehn and Knight [45] and implemented in the script *compound-splitter.perl* as a part of the Moses toolkit [4]. This method is easy to use and does not require any annotated data. For a given word, it finds a split $S = p_1, \ldots, p_n$ with the highest geometric mean of word frequencies of its parts p_i:

$$S = \arg\max_{S} \left(\prod_{p_i \in S} \text{count}(p_i) \right)^{\frac{1}{n}},$$

where $count(p_i)$ indicates the number of times the potential part of the compound p_i occurs as a single token in the corpus and n is a number of parts in the particular split S.

For most words, the averaged frequency of all possible partitionings is lower than the frequency of the word itself and thus no split is performed. Since German compounds often contain filler letters between its parts, we allow *-s-* and *-es-* (which cover the most frequent cases) as linking elements. For example, the compound *Transfusionsmedizin* consists of two parts *Transfusion* and *Medizin* and uses the filler *-s-* between them. The splitter is trained on the German side of parallel training data by simply collecting frequencies of all words in the corpus. Splitting is then performed as a preprocessing step on the source (German) side of the paralell training data and also on the input text to be translated.

We experiment with two configurations; the *C3* splitter is trained on *P3* (in-domain dictionaries and corpora) and the *C6* splitter is trained on *P6* (in-domain dictionaries and intelligent selection from in-domain and general-domain corpora). Statistics comparing the number of words that were split and the number of out-of-vocabulary words in the test data are shown

in Table 13. Comparing the translations with and without using the compound splitter, a statistically significant improvement of BLEU is only observed for the model trained on smaller in-domain data (*C3*) applied with the *P3 M0* configuration. In the other settings, the contribution of this method is not that evident, but compound splitting still substantially reduces the out-of-vocabulary rates and potentially also improves word alignment.

2.4.8. Exploiting synonyms as translation variants

As highlighted previously, one of the issues that MT of texts from professional domains has to deal with is the presence of domain-specific terms which are rare or absent in the training data and can lead to out-of-vocabulary words. In our work, we try to address the issue by introducing the notion of synonyms, mined from structured data, as several researchers have already shown this approach to be useful (see Section 2.1.6). While we often lack bilingual pairs of some terms, we may still be able to find synonymous or nearly-synonymous terms for which we do have bilingual pairs in our data. We explore two sources of synonymy information – the UMLS Metathesaurus and DBpedia described in Section 2.2.1.

Both of the data sources can be viewed as providing a mapping from various synonymous terms to one canonical term. In UMLS, the canonical term is the UMLS heading, such as *Mandible* in English or *Mandibule* in French, and the synonyms are the alternative headings, such as *Jaw*, *Lower jaw bone* and *Inferior Maxillary Bone* for English, or *Maxillaire inférieur* for French. In DBpedia, the canonical term is the article title and the synonyms are titles redirected to it. The redirected titles are not always true synonyms – often, they can be also hyponyms or other closely related terms, as is the case with the English article titled *Vitamin A*, to which many names of drugs containing vitamin A are redirected (*Disatabs*, *Myvpack*, *Testavol*, etc.). However, following Jones et al. [50], we believe that for the sake of IR, treating all such terms as synonyms is usually appropriate, and can even be beneficial. Conveniently, both UMLS and DBpedia also provide translations of the canonical terms. This allows us to easily extract additional sets of bilingual pairs.

Let S_c be a canonical term in the source language (e.g., *Mandible*); $\{S_i\}$ be synonyms of S_c in the source language (e.g., *Jaw*, *Lower jaw bone*, *Inferior Maxillary Bone*); and T_c be the canonical translation of S_c in the target language (*Mandibule*). For each S_i, we create a new bilingual pair $[S_i, T_c]$, which we add to the *P6* training data set ([*Jaw*, *Mandibule*], [*Lower jaw bone*, *Mandibule*], [*Inferior Maxillary Bone*, *Mandibule*]), replacing a part of the general domain data to keep the data set size constant; we also add $[S_c, T_i]$ pairs ([*Mandible*, *Maxillaire inférieur*]), i.e., target-side synonyms. The resulting numbers of tokens are presented in Table 14. This configuration is further denoted as *P6 M4 S1*.

Table 14 summarizes the results of MT with the mined synonyms added to the training data. They are generally unfavourable – while the differences in PER are less than one percentage point and even a slight improvement was observed for DE–EN, BLEU score decreases in all cases by more than 1 point. Still, the decrease in MT metrics does not necessarily mean a decrease of overall performance, since the employment of synonyms might hurt MT quality while improving the performance of subsequent IR.

Manual inspection of the translation outputs on the development data set shows that in many cases, the *target* term is substituted by its synonym or near-synonym, which is most probably caused by the fact that there are far more synonyms for English than for other languages in our data, as our source data sets are much larger for English. In some cases, this can be regarded as canonicalization, while in other cases, it is probably closer to term expansion. Unfortunately, we have observed very few cases where our approach led to avoidance of the out-of-vocabulary words – their frequency decreased from approximately 5% to approximately 4.6%. Still, we believe that exploiting synonyms from UMLS and DBpedia has a potential to improve the performance of our cross-lingual IR system.

2.5. Overview of main translation results

So far, we have focused on optimizing the translation quality of medical queries. The main achievements are summarized in Table 15, which compares performance to two freely available MT systems on the web: Google Translate[8] and Microsoft Bing Translator[9].

The general-domain baselines trained on 10 million parallel sentence pairs and 30 million monolingual sentences and tuned on general-domain data have been improved greatly for all language pairs. The improvement comes incrementally from in-domain tuning, careful selection of parallel and monolingual training data, and word decompounding (for DE–EN).

In terms of BLEU, we have added a total of 14.86 points for CS–EN, 17.62 points for DE–EN, and 11.83 points for FR–EN. All these results are statistically significant ($p < 0.05$). The relative improvements are remarkable at 55.89%, 76.51%, and 36.21%, respectively. The most substantial improvement is observed for DE–EN, but this is also the language pair with the lowest scores for the baseline system. In comparison with Google Translate, the results are quite competitive. In terms of BLEU, our system performs better – although not by a statistically significant difference (ranging between 0.80–2.18 points).

In terms of PER, the results are similar. The (inverse) PER scores of the baseline systems have increased by 16.36 points for CS–EN, 12.99 points for DE–EN, and 11.51 points for FR–EN. The respective relative improvements are 29.61%, 23.72%, and 17.51%. Google Translate has been outperformed by 1.11–1.62 absolute points.

In MT, automatic evaluation measures (BLEU and PER, in our case) may not correlate well with human judgements [e.g., 14]. In order to verify the automatic evaluation results, we carried out a human evaluation of the following systems: *P0 M0*, *P6 M4*, Google Translate, and *P6 M4 C3* for DE–EN. For each language pair and each query in the MT test set, a human expert was asked to rank outputs of the systems and the reference translations for 100 randomly sampled queries (presented in a

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

[9] `http://www.bing.com/translator/`

config	Czech–English		German–English		French–English	
	BLEU	1-PER	BLEU	1-PER	BLEU	1-PER
P6 M4	⋆**41.45**	⋆**71.61**	⋆**40.65**	**65.43**	⋆**44.50**	⋆**77.24**
P6 M4 S1	**40.38**	**70.86**	**37.52**	⋆**65.52**	**42.53**	**76.61**
size	1,841	1,994	2,640	3,265	2,928	2,760

Table 14: The effect of exploiting synonyms (*P6 M4 S1*) in the best-performing configuration (*P6 M4*). The last row refers to the size of the synonym data sets, measured in thousands of tokens on each side.

config	Czech–English			German–English			French–English		
	BLEU	1-PER	HUM	BLEU	1-PER	HUM	BLEU	1-PER	HUM
P0 M0 gen	26.59	55.25	23.91	23.03	54.76	29.31	32.67	65.73	17.05
P0 M0 query	35.73	66.21	–	29.50	60.40	–	37.84	71.78	
P6 M0	36.65	68.23	–	34.67	**64.03**	–	**42.74**	**76.47**	–
P6 M4	⋆**41.45**	⋆**71.61**	45.83	**40.65**	**65.43**	37.63	⋆**44.50**	⋆**77.24**	⋆56.06
P6 M4 C3	–	–	–	⋆**40.82**	⋆**67.75**	37.78	–	–	–
P6 M4 S1	**40.38**	**70.86**	–	**37.52**	**65.52**	–	**42.53**	**76.61**	–
Google	**40.65**	**70.50**	⋆56.47	**38.64**	**66.13**	⋆54.39	**42.95**	**76.01**	45.45
Microsoft	27.54	51.25	–	35.25	61.88	–	36.44	71.39	–
reference	–	–	74.16	–	–	80.29	–	–	84.34

Table 15: Comparison of translation quality in selected experiments. The main results are also compared with the translations by public web-based systems and a reference translation in a human evaluation (HUM).

random order) according to descending translations quality (ties allowed). As proposed by Bojar et al. [96], the output was transformed to pairwise comparison and is presented as a percentage of cases when the translation of a particular system is judged as better than outputs of the other systems, ties ignored (see the columns denoted as HUM in Table 15). Formally, let $\{S_j\}$ be a set of systems to be compared and $win(A, B)$ be the number of times system A is ranked better than system B.

$$HUM(S_i) = \frac{\sum_{i,i\neq j} win(S_i, S_j)}{\sum_{i,i\neq j} win(S_i, S_j) + win(S_j, S_i)} \cdot 100\%$$

The HUM score does not accord with the automatic measures, showing Google Translate being outperformed by *P6 M4* only for FR–EN. However, these differences should not be considered significant, since the proportion of ties in all pairwise contests is more than 72% (i.e., for each pair of systems, the two systems were judged of equal quality in more than 72% ofthe queries).

Further, we also investigate the final reduction of out-of-vocabulary words (both in terms of tokens and types) in the best systems compared to the baseline. As shown in Table 16, the reduction ranges between 42% and 69%, depending on language pair. The most substantial decrease is observed for DE–EN, where the effect of exploiting in-domain training data is emphasized by word decompounding.

2.6. Manual analysis of translation results

In addition to the human evaluation presented in the previous subsection, we also performed a detailed manual analysis of the results achieved by the best-performing systems (*P6 M4* for CS–EN and FR–EN, *P6 M4 C3* for DE–EN) in comparison with the baselines (*P0 M0*). We hired medical experts to judge the quality of 50% samples of the *query* test set translations that were produced by the two systems (our best performing one and the baseline) for all the language pairs, using the following five-point scale:

4 – perfect translation, identical to the reference;
3 – perfect translation, different from the reference;
2 – acceptable translation, errors allowed in morphology, word order, and stopwords;
1 – bad translation, no untranslated words;
0 – bad translation, some words untranslated.

This scale allows us to easily quantify the translation quality of all the systems and also to analyze the improvement of the best systems over the baselines. For a complete overview for all the language pairs, see Table 17. For simplicity, we describe the findings for CS–EN only; however, the results for DE–EN and FR–EN are very similar.

The total of 45% of Czech test queries are translated by the baseline system to match the reference translations. However, an additional 23% are also judged as perfect translations, although different from the reference translation. Such cases are not (fully) matched by the automatic measures and their scores are undervalued (but this is a traditional problem in MT whenever test sets with a single reference translation are used). A further 16% of the translations cannot be perceived as fully correct, but are considered adequate for querying in IR. For example: *potravinová alergie* (*food allergy*) translated as *food allergies* (error in number), *chirurgické odnětí dělohy* (*hysterectomy*) translated as *surgical womb removed* (error in syntax), *rodičovský* (*parental*) translated as *parent* (error in part-of-speech), *růstový faktor hepatocytů* (*hepatocyte growth factor*) translated as *growth factor hepatocytes* (error in word order). IR systems typically remove stopwords, ignore the word order, and perform some kind of morphological normalization.

config	Czech–English		German–English		French–English	
	token	type	token	type	token	type
P0 M0	7.97	11.29	9.53	13.10	3.78	6.71
P6 M4 / C3	4.53	6.51	2.90	4.05	1.69	2.96
reduction (%)	−43.16	−42.34	−69.57	−69.08	−55.29	−55.89

Table 16: Relative OOV rates (%) of types and tokens for the baseline *P0 M0* and the best system *P6 M4 / C3*.

Czech–English						
C	4	3	2	1	0	Σ
4	38.8	*3.6*	*1.7*	*0.4*	*0.1*	44.6
3	**4.4**	14.5	*1.6*	*1.6*	*0.5*	22.6
2	**4.9**	**3.7**	5.6	*1.2*	*0.2*	15.7
1	**1.1**	**1.4**	**0.0**	1.1	*0.0*	3.6
0	**3.7**	**2.1**	**0.5**	**0.5**	4.7	11.4
Σ	52.9	25.3	9.4	4.7	5.5	

German–English						
C	4	3	2	1	0	Σ
4	35.9	*2.3*	*1.1*	*2.9*	*0.1*	42.3
3	**4.1**	11.3	*1.4*	*1.7*	*0.0*	18.5
2	**3.7**	**2.3**	2.8	*1.1*	*0.0*	9.9
1	**3.7**	**2.6**	**0.6**	2.3	*0.9*	9.9
0	**4.5**	**3.4**	**2.6**	**2.3**	1.6	14.3
Σ	51.8	21.8	8.5	10.3	2.6	

French–English						
C	4	3	2	1	0	Σ
4	47.0	*1.3*	*1.7*	*1.2*	*0.1*	51.3
3	**4.2**	15.5	*2.0*	*2.2*	*0.0*	24.0
2	**3.9**	**2.8**	4.7	*0.9*	*0.0*	12.3
1	**2.1**	**0.9**	**0.7**	2.3	*0.2*	6.3
0	**1.9**	**0.9**	**0.9**	**0.4**	1.8	6.0
Σ	59.1	21.4	10.1	7.1	2.1	

Table 17: Results of manual translation-error analysis of the baseline (*P0 M0*) and best systems (*P6 M4 / C3*). The figures (in %) represent joint and marginal (Σ) distributions of categories 0–4 (*C*) observed in the baseline (rows) and best system (columns) translations. The bold and italics fonts denote improvement and degradation, respectively.

Therefore, translation errors in such phenomena do not usually harm retrieval performance. Finally, a total of 15% of the translations are completely wrong, and 3/4 of them contain one or more words that remained untranslated.

When we compare the results of the best system with the baseline, we observe an improvement in 22% of the translations (sum of the figures in bold for CS–EN Table 17) and a degradation in 11% (sum of the figures in italics). The quality of the remaining translations does not change. The distribution of all types of changes is depicted in Table 17 (improvements in bold and degradations in italics): 55% of the bad translations (category 1 and 0) are improved in such a way that they are judged as perfect (category 4 and 3). However, in less than 3% of cases, we also observe the opposite behaviour, which can be explained either by the change of training data where the correct translation is not present anymore and the system is not able to generate it, or the system is able to generate the correct hypothesis but it is not scored as the highest, or the system fails to find it because of pruning the search space. The best system for CS–EN is estimated to produce perfect results (category 4 and 3) in 78% and only 11% are judged as bad (category 1 and 0). Such results seem very promising for the application in IR investigated in the Section 3 of this paper.

2.7. Summary

In this section, we described a series of experiments focused on increasing quality of machine translation of medical queries. Substantial improvements were obtained by tuning and training on in-domain data. Even better results were observed in experiments using pseudo-in-domain training data (both parallel and monolingual) acquired by intelligent selection from large pool of data irrespective of domain. For the DE–EN translation direction, the translation quality was further improved by splitting compound words on the source language side (German). Other techniques investigated were not shown to have such a positive effect: they either did not bring any significant improvement (optimization of phrase table configuration, exploiting synonyms) or led to a significant degradation of translation quality (morphological normalization on the source language side). Translation quality was evaluated by automatic comparisons against reference translations and by human experts. In automatic evaluation, our best system even outperformed the on-line translation systems of Google and Microsoft, although in manual evaluation, this was not fully confirmed. In the thorough manual analysis of translation errors of our best systems, we observed that about 70–80% of translations (depending on translation direction) are perfect and only about 9–12% are not acceptable. The remaining 10% (approximately) is expected to be acceptable in cross-lingual IR as it only contains minor errors.

3. Optimizing query translation for cross-lingual information retrieval

In a standard MT scenario, the MT system is optimized to produce an output aimed to be read by a human. However, if used in a cross-lingual IR (CLIR) system, a consumer of the MT output is a computer system performing IR. Such systems usually do not require the input to be linguistically fluent or grammatically correct. The ordering of words can be loose and function words and the accuracy of other words deemed to be IR-irrelevant (traditionally called stopwords) does not matter. On the other hand, inclusion of synonymous words or words related to the query typically can have a positive effect by encouraging matching with these terms in relevant documents. Moreover, over the years, people have become accustomed to communicating with IR systems in the language of keywords. Thus, we can assume a human input to a CLIR system will often be of this non-linguistic form. Although there are tasks where the expected queries may be longer (e.g., in patent retrieval, the query is the complete text of a patent proposal [97]), typical queries tend to be much shorter than a standard MT input [81].

The average sentence length in the *gen* and *query* domains (see Table 3) shows that our data is not an exception. These assumptions allow us to introduce techniques that would be harmful in a standard MT scenario, but may be beneficial to the quality of the CLIR system if the MT output serves as a query.

This section continues with an overview of related work followed by a description of our experimental CLIR setup, the experiments conducted, their results, and an analysis of our findings.

3.1. State-of-the-art in CLIR and related work

IR has been studied for several decades now. The first attempts to design "auto-indexing" machines date back to the 1950's [98], but it has now evolved into a large research field covering a wide range of tasks and problems [99]. We first present work related to traditional monolingual IR applied to medical-domain data. Then, we review approaches to cross-lingual IR with a focus on query translation and continue with an overview of CLIR work targeting the medical domain.

3.1.1. Monolingual information retrieval in the medical domain

Given that much medical content is written in the English language, research to date in the medical space has predominantly focused on monolingual English retrieval. A number of evaluation campaigns dedicated to this task have taken place.

The first such evaluation campaign using medical data for evaluation of IR was OHSUMED [100]. The test collection contained around 350,000 abstracts from medical journals taken from the MEDLINE database over a period of five years and two sets of topics: a manually created one and another one based on the controlled vocabulary thesaurus of the Medical Subject Headings (MeSH) [68].

The TREC Genomics Track ran between 2003 and 2007. The test collection comprised publications from medical journals and clinical reports related to genes and genomics. The track included tasks ranging from ad-hoc retrieval to document categorization, passage retrieval, and entity-based question-answering [101]. More recently, the TREC Medical Records Track ran in 2011 and 2012 [102]. This track was based on a collection of anonymized medical records and queries that resembled eligibility criteria of clinical studies. The goal was to find patient cohorts that are relevant to the given criteria for recruitment as populations in comparative effectiveness studies.

These past campaigns for health IR technique development have focused on physicians and other health care professionals. However, a new evaluation campaign introduced in 2013, CLEF eHealth [103], considers types of queries posed by laypeople searching the web for medical information. Given the realistic nature of these (English) queries, and the very real need for translation of non-English medical queries for laypeople searching for medical information on the web, we use this test collection for the cross-lingual IR evaluations described in this article. For these queries, which mostly contain no more than two terms (items indexed by the system, typically single content words), this type of translation is very challenging. The IR test collection is described in greater detail in Section 3.2.

3.1.2. Cross-lingual information retrieval and related techniques

In contrast to standard IR, the query and the set of documents in CLIR are not in the same language. This issue is generally dealt with by translating queries into the language of the documents, or translating the documents into the language of queries. An alternative approach is to translate both queries and documents into another language or a language-independent semantic representation [e.g., 104, 105]. We follow the first approach based on translation of queries rather than documents. However, most of the techniques described can be applied to document translation as well. A detailed overview of CLIR can be found in Nie [106], Peters et al. [107], or Zhou et al. [108].

The categorization of CLIR systems is traditionally based on the method for translating queries: using a dictionary, MT, or corpora. Dictionary-based methods employ machine-readable bilingual dictionaries to map the query language onto the language of documents. These methods were investigated, e.g., by Ballesteros and Croft [109], Maeda et al. [110], and Gao and Nie [111]. With the advent of modern SMT, the difference between machine-translation-based and corpora-based CLIR has faded away. SMT systems can be adapted to act similarly to dictionary-based methods, e.g., by disabling phrase reordering, or by incorporating a human-made dictionary into the translation model. Unlike the traditional dictionary-based approaches, SMT is of a stochastic nature and exploits additional sources of information, such as a target-language model and context of individual terms in the query, which are combined using advanced machine learning techniques. These advantages resulted in the current dominance of SMT-based approaches to CLIR [108, pp. 23–24], exploiting especially online services such as Google Translate or Microsoft Bing Translator. Nonetheless, these systems lack some means of adaptation to specific domains and are usually constrained to limit their use, e.g., by imposing a maximum number of translations per day.

The major issues one has to face in query translation are ambiguity and low coverage [108]. Ambiguity can arise in both the source language and the target language. For example, the German noun *Kanne* can be translated into English as *can* or *canister*. The former translation would correspond to the auxiliary verb *can* which is typically considered as a stopword for IR (and thus, removed from queries and documents). The issue of ambiguity is usually alleviated by enriching the query with multiple translation options. A common application is *structured query translation* introduced by Pirkola [112] who implemented a synonym operator to group the translation alternatives for individual words. Darwish and Oard [113] extend his work by weighting the translation candidates by translation probabilities. Federico and Bertoldi [114] employ a query-translation model based on a Hidden Markov Model and a language-model-based query-document model within a single statistical framework. Integration of the two models is ensured over the weighted n-best list of possible translations of the query.

Several methods address the issue of coverage, i.e., of handling out-of-vocabulary words during translation. Stemming is

a standard method used in MT as well as in IR to effectively cluster words by removing their inflectional and derivational affixes. Both rule-based [93] and statistical [94] stemmers have been applied. Lemmatization is an alternative approach of substituting various forms by the canonical form of a given word. So far, however, the best application in IR by Hollink et al. [115] produce results of only moderate quality.

Another technique for handling the out-of-vocabulary words is *query expansion*, which works by enriching the query with synonymous or related expressions. It can be achieved by a widely-used approach known as *pseudo-relevance*, *blind-relevance*, or *local feedback* [116] – the query in the source language is used to retrieve the top-ranked documents from the collection in the same language. The high-weighted terms (assumed to be related) are then extracted from these documents and added to the original query. This technique was applied to CLIR e.g., by Ballesteros and Croft [117]. If performed also on the target side, it can mitigate the effects caused by picking wrong translation alternatives [see summary in 107].

Another technique closely tied with IR but rarely performed in standard MT is *stopword removal*. Prepositions, articles, pronouns, conjunctions, and other non-significant words are typically not indexed in IR document collections and can be removed from the queries. Magdy and Jones [97] remove stopwords even from the MT training data (along with stemming) and shown significant speed-up of the translation process that follows in the CLIR setup, and requires less training data for the MT system.

3.1.3. Cross-lingual information retrieval in the medical domain

Cross-lingual retrieval in the medical domain has been addressed in several previous works. The majority of them employ the UMLS Metathesaurus as the main source of health-related information. For instance, Eichmann et al. [118] use UMLS to translate Spanish and French queries into English following several strategies: full and partial phrase match, dictionary-based look-up, and simple adding of the source language query words. Volk et al. [119] identify the UMLS terms and their semantic relations in both queries and documents which is reported to improve performance in both cross-lingual (German–English) and monolingual IR. Tran et al. [120] show that UMLS-based translation mixed with hybrid translation (combining pattern-based module with morpho-syntactic conversion rules) outperforms the two translation components used separately in French–English CLIR. Déjean et al. [121] focus on extraction of bilingual lexicons from parallel and comparable corpora to enrich monolingual or bilingual medical thesauri (such as UMLS). They show that using such improved lexicons in CLIR significantly improves the performance and outperforms using the existing ones acquired in the traditional way.

The structured queries proposed in Pirkola [112] are evaluated on data from the medical domain. To better deal with the issues of coverage and ambiguity, the authors build a Finnish–English health dictionary containing more than 60 thousand entries. They show that CLIR systems based on dictionary-based translation could achieve the performances of a monolingual system if the queries structured and both general and domain terminologies are available. Rosemblat et al. [122] use medical queries from the Clinical Trials website[10] to compare two main approaches in CLIR – query translation and document translation. Their results favour the former approach. The MorphoSaurus multilingual retrieval system for medical documents [123, 124] adopts the approach of translating both queries and documents to a morpho-semantic representation. Its central component is a dictionary, whose entries constitute equivalence classes of morpho-semantically minimal units, capturing interlingual as well as intra-lingual synonymy. The MorphoSaurus system was shown to outperform standard IR and CLIR approaches for languages such as German, where decompouding words into smaller lexical units can greatly benefit IR performances.

3.2. Data description

The IR evaluation described in this article was carried out on the CLEF eHealth 2013 Task 3 test collection [103]. It consists of a set of around one million web pages covering medical topics related to general public and general practitioners collected by the EU FP7 Khresmoi project[11] and 50 English medical queries with corresponding relevance assessments generated from the pooled set of results submitted to the task.

The documents are predominantly health and medicine websites that have been certified by the Health on the Net (HON) Foundation as adhering to the HONcode principles[12] (approx. 60–70% of the collection) as well as other commonly used health and medicine websites such as DrugBank[13], Diagnosia[14], and Trip Answers[15]. The documents are provided in the data set in their raw HTML format along with their uniform resource locators (URLs).

The queries in the collection aim to model those used by laypeople (i.e., patients, their relatives, or other representatives) to find out more about their disorders in a specific situation, after they have examined their discharge summary. The discharge summaries used for the task originate from the anonymized clinical free-text notes of the MIMIC II database, version 2.5[16]. The queries are intended to be representative of real patients' information needs and statements, and as such are relatively short with average length of no more than two terms (words). The generated queries consist of a topic title (text of the query), description (longer description of what the query means), and narrative (expected content of the relevant documents). The query set (topic titles) has been manually translated for the purpose of this work by medical professionals into German, French, and Czech (and double reviewed). The titles of the original English topics include, for example: *facial cuts and scar tissue*;

[10] http://www.clinicaltrials.gov/
[11] http://khresmoi.eu/
[12] http://www.hon.ch/HONcode/Patients-Conduct.html
[13] http://www.drugbank.ca/
[14] http://www.diagnosia.com/
[15] http://www.tripanswers.org/
[16] http://mimic.physionet.org/

asystolic arrest; *nausea and vomiting and hematemesis*; *sinus tachycardia*; *chills and gallstones*.

Relevance assessment was performed by domain experts and IR experts on documents obtained by pooling the top ten documents from three runs submitted by each of 9 participants to the CLEF 2013 eHealth Task 3, which resulted in a pool of 6,391 documents. A total of 1,878 documents were assessed as relevant, which is 37.56 per topic on average. Given that only the top ten documents were assessed for each selected run, the resulting pools are rather shallow. Therefore, relevant documents may have been missed in the assessment process. For details on this process see Goeuriot et al. [125].

3.3. System description

Our experimental IR system is based on the BM25 retrieval model [126, 127] implemented in Lucene 3[17]. The quality of this model for IR in the medical domain has been previously demonstrated by Leveling et al. [128].

Given a query q, a document d in this model is scored by the following formula:

$$score(d,q) = \sum_{t \in q} w^{(1)} \cdot \frac{(k_1+1)\mathrm{tf}_{td}}{k_1(1-b+b\cdot(L_d/L_{\mathrm{ave}}))+\mathrm{tf}_{td}} \cdot \frac{(k_3+1)\mathrm{tf}_{tq}}{k_3+\mathrm{tf}_{tq}} \cdot L_d \cdot L_{\mathrm{ave}} + L_d$$

where t ranges over the terms in q, k_1, k_3, and b are model parameters, tf_{td} is the frequency of t in d, tf_{tq} is the frequency of t in q, L_d is the length of d, L_{ave} is the average document length, and $w^{(1)}$ is the Robertson and Spärck Jones [129] weight of t in q defined as:

$$w^{(1)} = \frac{(R_t+0.5)/(R_q-R_t+0.5)}{(\mathrm{df}_t-R_t+0.5)/(N-\mathrm{df}_t-R_q+R_t+0.5)},$$

where N is the number of documents in the collection, R_q is the number of documents known or presumed to be relevant for q, R_t is the number of the relevant documents containing t, and df_t is the document frequency for t. The default values for the model parameters used are $b = 0.75$, $k_1 = 1.2$, and $k_3 = 7$ [126].

In the pseudo-relevance feedback experiments, we set the parameters after empirical experimentation on the test queries, varying the range of feedback terms and documents between 5 and 50. For the test queries, the parameters selected yielded the best results (e.g., MAP). The expansion terms are thus chosen by taking the top $R_q = 10$ documents from the initial retrieval step presumed to be relevant and selecting the top $T = 10$ terms in the documents ranked by the term selection value [130] defined as:

$$TSV = \frac{R_d}{R_q} \cdot w^{(1)}.$$

The documents were preprocessed by stripping out HTML tags and the content flattened into a single index. We applied the JSOUP[18] processing libraries to extract textual content of web pages. This removed HTML markup and JavaScript, leaving raw textual content. Standard Lucene modules were employed to tokenize the text and to fold upper case characters to lower case. A stopword list for English (571 words) by Salton [131] was used to identify stopwords. Stemming of topics and documents was performed using the English Snowball stemmer provided in Lucene, which is based on the Porter algorithm [95].

3.4. Experiments

We first evaluate the translation quality of the selected MT systems as applied to the IR test sets described in Section 3.2 and then analyze the effect of using the retrieval techniques described in Section 3.3.

3.4.1. Machine translation quality

Most of the translation experiments described in Section 2 improved the baseline results measured on the MT (*query*) test sets quite substantially (see Sections 2.5–2.7). The best systems even outperformed the state-of-the-art publicly available on-line services (Google Translate and Microsoft Bing Translator). In this section, we analyze the translation quality of the best-performing MT systems applied on the IR test sets consisting of the titles of the 50 CLEF eHealth 2013 Task 3 test topics in English and their translations. We also introduce some new MT systems designed specifically for CLIR. The evaluation in this section is *intrinsic* – based on comparison of translation quality and realized by the standard automatic measures (BLEU and PER) as well as human evaluation (for *P0 M0*, *P6 M4*, and Google only).

The results are shown in Table 18: *P0 M0* refers to the baseline systems tuned on general-domain data, *P6 M4* to the best configuration of parallel and monolingual training data, *P6 M4 comp* denotes the configuration *C3* exploiting German decompounding, *P6 M4 syn* the configuration *S1* exploiting synonyms, and *P6 M4 per* is a new configuration based on *P6 M4*, tuned on the *query* development sets by MERT [9] optimizing PER instead of BLEU. As explained in Section 2.4, PER ignores word order, which implies more focus on translation adequacy and less focus on fluency compared to BLEU.

In contrast to the tables presented in the previous section, the bold font indicates those scores that are significantly better than the baseline *P0 M0*. The tests are performed in the same way: by the paired bootstrap resampling for BLEU [90] and by the paired t-test for PER, both with $p < 0.05$. There were no BLEU or PER results significantly worse than the baseline. The best scores for each language pair are again indicated by the $\star$ symbol.

The 95% confidence intervals for BLEU observed on the IR test sets are much larger (21–29 points) than in the case of the MT test sets (7–12 points). This is not surprising because the IR test sets contain just 50 queries compared to the 1,000 queries in the MT test sets and the sample variance is much higher. Therefore, the results presented in Table 18 (IR test sets) cannot be considered as reliable as those presented in Table 15 (MT test sets). Also, the MT and IR test sets cannot be considered completely comparable – although both comprise medical queries,

[17] http://lucene.apache.org/
[18] http://www.jsoup.org/

config	Czech–English			German–English			French–English		
	BLEU	1-PER	HUM	BLEU	1-PER	HUM	BLEU	1-PER	HUM
P0 M0	47.01	66.41	26.56	39.52	62.47	17.95	39.20	71.48	12.09
P6 M4	40.91	70.26	28.57	42.95	**64.19**	36.36	**52.96**	**76.69**	55.56
P6 M4 comp	–	–	–	43.42	**66.41**	–	–	–	–
P6 M4 syn	47.60	71.66	–	40.19	**62.58**	–	**54.50**	**77.19**	–
P6 M4 per	52.28	**75.06**	–	**50.24**	**68.91**	–	**51.01**	**80.05**	–
Google	⋆**56.02**	⋆**77.30**	⋆75.93	**54.53**	**75.78**	⋆71.43	⋆**61.99**	⋆**83.02**	⋆66.67
Microsoft	47.46	67.06	–	⋆**59.72**	⋆**76.54**	–	**58.34**	80.79	–
reference	–	–	72.31	–	–	78.87	–	–	76.92

Table 18: Performance of selected MT systems applied on the IR test data. The scores in bold font are significantly better than the baseline (*P0 M0*). The ⋆ symbol indicates the best scores for each language pair.

the IR test sets were created in a more controlled setting. In addition to the automatic evaluation using BLEU and PER, we also conduct a human evaluation (column HUM in Table 18) using the same method as described in Section 2.5 but performed on the entire IR test sets.

The overall translation results on the IR test sets do not confirm the results from Section 2 (Table 15) achieved on the MT test sets, where our best system significantly improved baselines for all language pairs and even outperformed Google Translate and Microsoft Bing Translator.

Although most of our systems are able to outperform the baseline, the improvement of BLEU is statistically significant for DE–EN and FR–EN only. For CS–EN, the baseline scores are surprisingly high and significantly outperformed only in terms of PER by the PER-tuned system (*P6 M4 per*). This configuration performs best also for DE–EN and FR–EN measured by PER – in terms of BLEU, it wins for CS–EN and DE–EN. The winner for FR–EN is *P6 M4 syn* but not by a statistically significant margin (compared with *P6 M4* and *P6 M4 per*).

Despite these findings, Google Translate dominates all our systems for all language pairs measured by all three measures. Although for DE–EN, the absolute winner is Microsoft Bing Translator, none of these results are significantly better than those achieved by our best systems. Interestingly, the HUM score for CS–EN suggests that the output of Google Translate is better than the reference. However, this is caused by including comparison with the other systems (*P0 M0* and *P6 M4*), in which Google Translate is judged as better more often than the reference. The pairwise score reveals that in 9 queries, the reference translation was judged to be better, in 8 queries, Google Translate outperformed the reference, and in 33 cases, both systems were judged equally, thus confirming their comparable quality.

3.4.2. Information retrieval quality

The translations of the IR test sets produced by the MT systems presented in the previous subsection are now employed in the IR setup described in Section 3.3. The results of these experiments are reported in the first part of Table 19, using the standard IR evaluation measures: precision at a cut-off of 10 documents (P@10), normalized discounted cumulative gain [132] at 10 documents (N@10), and mean average precision (MAP) [133]. The cross-lingual MAP scores are also compared with the monolingual ones, i.e., those obtained by using the reference (English) translations of the test topics to see how the system would perform if the queries were translated perfectly (see columns denoted as MAP^{rel}_{EN}). Our monolingual scores are quite comparable to the best results achieved by the CLEF eHealth 2013 task participants [125].

The performance metrics for the IR experiments are computed with the standard TREC evaluation tool.[19] We also indicate significance of the results using the standard Wilcoxon signed rank test ($p < 0.05$) [134]. For comparison with the MT experiments presented in previous section, the IR results are also tested against the *M0 P0* baseline (of the respective language pair); those that are significantly better are typed in bold and those which are significantly worse are typed in italics. The best cross-lingual results are marked with a ⋆ symbol.

The first important observation is that using pseudo-relevance feedback (surprisingly) does not improve retrieval performance. This finding can be attributed to the fact that for this evaluation collection, only all documents up to rank 10 in the runs used to generate the pool (see Section 3.2 for greater details) have been assessed for relevance/non-relevance due to relevance assessor availability constraints. This means that some relevant documents may not have been assessed. Pseudo-relevance feedback as a form of query expansion could retrieve additional documents at top ranks which were not included in the initial retrieval results. These additional documents likely have not been assessed for relevance (and thus count as not relevant even if they are relevant). A similar behaviour of BM25 with pseudo-relevance feedback can be observed in the result of the organizers' baseline experiment for the CLEF eHealth 2013 Task 3 [135]. Hence, in our analysis, we focus on the runs which do not use pseudo-relevance feedback, and correspondingly, no results using pseudo-relevance feedback are included in Table 19.

Unsurprisingly, translated queries do not perform as well in retrieval as the original English queries. Interestingly, the highest baseline scores (*P0 M0*) are observed for the CS–EN translation direction. Czech is not usually easier to translate than German and especially French; however, on this test set, we

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

	Czech–English				German–English				French–English			
Run ID	P@10	N@10	MAP	MAP^{rel}_{EN}	P@10	N@10	MAP	MAP^{rel}_{EN}	P@10	N@10	MAP	MAP^{rel}_{EN}
reference	**47.0**	**42.1**	**30.35**	**100.0**	**47.0**	**42.1**	**30.35**	**100.0**	**47.0**	**42.1**	**30.35**	**100.0**
P0 M0	34.8	31.1	24.28	80.00	29.4	25.6	19.02	62.67	31.0	27.1	21.87	72.06
P6 M4	37.2	32.3	23.67	77.99	32.6	28.6	20.39	67.18	**38.4**	**34.5**	**26.33**	86.75
P6 M4 comp	–	–	–	–	32.8	29.0	21.85	71.99	–	–	–	–
P6 M4 syn	35.0	30.5	23.11	76.14	30.6	26.7	19.79	65.21	38.0	33.2	25.17	82.93
P6 M4 per	35.4	30.7	23.16	76.31	35.0	30.8	22.62	74.53	**38.4**	**34.1**	**25.94**	**85.47**
P6 M4 stem	28.2	*24.0*	20.27	66.79	23.4	19.8	16.29	53.67	32.0	26.5	20.33	66.99
P6 M4 n5	31.0	27.5	22.42	73.87	24.0	21.4	*16.60*	*54.70*	29.2	26.6	20.94	69.00
P6 M4 n10	31.4	27.7	22.71	74.83	21.8	19.8	*16.19*	*53.34*	29.6	27.6	21.44	70.64
Google	⋆38.4	⋆34.4	⋆**25.97**	⋆**85.57**	**37.0**	**33.2**	**23.22**	**76.51**	⋆**40.6**	**36.1**	**26.74**	**88.11**
Microsoft	32.6	28.7	22.76	74.99	⋆**38.8**	⋆**34.2**	⋆**25.09**	⋆**82.67**	**40.2**	⋆**36.3**	⋆**27.57**	⋆**90.84**

Table 19: IR results for query translations produced by various MT systems compared with the original (reference) queries in English. The scores typed in bold, normal, and italics are significantly better, equal, and worse (respectively) than the baseline (*P0 M0*). MAP^{rel}_{EN} refers to MAP relative to the monolingual performance (reference). All figures are displayed as percentages.

experience the opposite (see also the translation results in Table 18). This behaviour can be explained by the randomness of training data selection for the baseline system, which for this particular language pair must have contained more material relevant to this data set than for the other language pairs. With the exception of CS–EN, the best MT systems (measured by IR performance) are those tuned for PER (*P6 M4 per*). For CS–EN, the winner is the plain *P6 M4* configuration with no additional enhancements.

In the overall comparison with the commercial on-line translation systems, the single winner would be Google Translate – it beats Microsoft Bing Translator on CS–EN and FR–EN and is on par with it on DE–EN (see the third part of Table 19). It performs best translating French to English, with only a 14% reduction in P@10 compared to the monolingual baseline. Microsoft Bing Translator also outperforms our translation technique for German and French queries. However, for Czech queries, better P@10 results are obtained using our translation technique than using Microsoft Bing Translator. Here a 21% decrease in P@10 is noted relative to the monolingual baseline. Relative to this, queries translated from French using our technique yield an 18% decrease in P@10, and queries translated from German yield a 31% decrease.

In addition to the MT systems employed so far, we introduce two new configurations of MT systems aiming at improving retrieval quality – one for producing stemmed translations (stemming) and one for exploiting multiple translation options. Their results are presented in the second part of Table 19.

Stemming. Stemming is a standard technique used in IR. In the traditional CLIR setup, this step is applied ex-post – on the MT output which is in the traditional human readable form. In our configuration (denoted as *P6 M4 stem*), we produce stemmed output directly during translation. This is achieved by stemming the target language side of the parallel training data as well as the monolingual data for language models. In this experiment, we employ stemming also on the source language side to reduce the morphological complexity (which is important especially for Czech). We use the Porter's Snowball stemmer for the source side languages (Czech, German, French) and the original Porter's stemmer for English [93]. However, the results are not very optimistic. Similar to the experiments with morphological normalization described in Section 2.4.6 which focused on translation quality, we also observe degradation in the retrieval performance when stems are used instead of full word forms on both source and target side. Training MT on stemmed words probably introduces too much ambiguity, which hurts not only MT quality but also IR performance.

Query expansion by multiple translation options. Our MT system can produce multiple translation variants for each query, which can be easily incorporated in the IR setup by using the entire n-best list as a translation of the query. We experimented with several values of n but present results for $n = 5$ and $n = 10$ only. The scores of *P6 M4 n5* and *P6 M4 n10* in Table 19 show that the resulting performance is even lower than the baseline *P0 M0* (although the difference is significant only for MAP on DE–EN). A possible reason for the decrease is that the translation variants differ to such an extent that they cannot be considered good translations and therefore, the queries are expanded by non-relevant terms.

4. Conclusions

In this work, we explored cross-lingual IR in the domain of medicine and focused on machine translation as a key component introducing the possibility to search in a multilingual environment. We translate queries in Czech, German, and French to English and perform search on a collection of English documents from CLEF eHealth 2013 Task 3. Such a task is especially challenging when applied to a specific domain, such as medicine, because traditional MT systems are not generally tuned to translate short expressions (queries) in specific domains.

The experiments described in this paper were conducted within the Khresmoi project. In the first phase of the work, we focused on improving translation quality of a baseline general-domain system by means of domain adaptation. Most of the

adaptation techniques (including in-domain tuning and training, selection of pseudo-in-domain data) substantially outperformed the baseline and even the state-of-the-art on-line MT systems (Google Translate, Microsoft Bing Translator). For DE–EN, the translation quality was also improved by automatic decomposition of complex German compound words. We also explored some more advanced methods, such as exploiting synonyms extracted from domain-specific resources or morphological normalization on the source language side. These techniques, however, did not bring additional improvement of translation quality.

In the second phase of our experiments, we focused on the impact of various MT (adaptation) techniques on retrieval quality. The retrieval experiments were conducted on a set of 50 topics from the CLEF eHealth 2013 Task 3 and their relevance assessments. In addition to the techniques used to improve the traditional translation quality, we conducted experiments with SMT tuning for PER (position-independent word error rate), stemming on the target language side, and query expansion by using multiple translation options. The overall results do not correlate with the findings of the MT experiments. The IR system using our baseline translations was significantly improved by using the adapted translations for the FR–EN translation only. However, none of our translations did outperform the system using Google Translate for query translation; our CS–EN system did outperform Microsoft Bing Translator in this CLIR setting, though.

This is the first comprehensive attempt to rigorously assess the contribution of domain-based MT adaptation as well as IR-targeted MT adaptation to real user queries about health and health-related problems. Our overall results are very positive in terms of MT quality, even though also report some negative results to illustrate that some traditional techniques for MT do not improve results in this specific domain and for the purpose of short query translation. When applied to the cross-language IR task, the positive MT results have not directly translated to improvements over the state-of-the-art in MT; we have shown in detail which techniques have a certain potential and which seem to lead to a dead end. The bottom line of this work can be expressed in two points: first, adapting MT systems using in-domain data can lead to major performance improvements and results that even surpass large-scale commercial systems, and second, MT quality and IR quality do not correlate in a straightforward way. We consider these findings promising for our future work. The highest potential seems to be in query expansion through analysis of multiple translation options, especially in combination with synonyms incorporated directly in the translation models as translation variants.

5. Acknowledgments

This work was supported by the EU FP7 project Khresmoi (contract no. 257528), the Czech Science Foundation (grant no. P103/12/G084), the Science Foundation Ireland (grant no. 07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University, and by the ESF project ELIAS.

The work described herein uses language resources hosted by the LINDAT/CLARIN repository[20], funded by the project LM2010013 of the MEYS of the Czech Republic.

References

[1] S. Fox, Health Topics: 80% of internet users look for health information online, Technical Report, Pew Research Center, 2011.

[2] R. J. W. Cline, K. M. Haynes, Consumer health information seeking on the internet: the state of the art, Health Education Research 16 (2001) 671–692.

[3] C. T. Lopes, C. Ribeiro, Measuring the value of health query translation: an analysis by user language proficiency, Journal of the American Society for Information Science and Technology 64 (2013) 951–963.

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

[5] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, PA, USA, 2002, pp. 311–318.

[6] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, H. Sawaf, Accelerated DP based search for statistical translation, in: G. Kokkinakis, N. Fakotakis, E. Dermatas (Eds.), Proceedings of the Fifth European Conference on Speech Communication and Technology, International Speech Communication Association, Rhodes, Greece, 1997, pp. 2667–2670.

[7] F. Jelinek, Statistical methods for speech recognition, MIT Press, Cambridge, MA, USA, 1997.

[8] F. J. Och, H. Ney, A systematic comparison of various statistical alignment models, Computational linguistics 29 (2003) 19–51.

[9] F. J. Och, Minimum error rate training in statistical machine translation, in: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Sapporo, Japan, 2003, pp. 160–167.

[10] N. Bertoldi, B. Haddow, J.-B. Fouet, Improved minimum error rate training in Moses, Prague Bulletin of Mathematical Linguistics 91 (2009) 7–16.

[11] P. Koehn, Europarl: a parallel corpus for statistical machine translation, in: Conference Proceedings: the tenth Machine Translation Summit, Asia-Pacific Association for Machine Translation, Phuket, Thailand, 2005, pp. 79–86.

[12] S. Roukos, D. Graff, D. Melamed, Hansard corpus of parallel English and French, 1995. Linguistic Data Consortium, Philadelphia, PA, USA.

[13] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, et al., The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, D. Tapias (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation, European Language Resources Association, Genoa, Italy, 2006, pp. 2141–2147.

[14] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, L. Specia, Findings of the 2012 workshop on statistical machine translation, in: Proceedings of the Seventh Workshop on Statistical Machine Translation, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 10–51.

[15] P. Pecina, A. Toral, V. Papavassiliou, P. Prokopidis, J. van Genabith, Domain adaptation of statistical machine translation using web-crawled resources: A case study, in: M. Cettolo, M. Federico, L. Specia, A. Way (Eds.), EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Trento, Italy, 2012, pp. 145–152.

[16] P. Langlais, Improving a general-purpose statistical translation engine by terminological lexicons, in: COLING-02 on COMPUTERM 2002: second international workshop on computational terminology,

[20] http://lindat.cz/

volume 14, Association for Computational Linguistics, Taipei, Taiwan, 2002, pp. 1–7.

[17] G. Sanchis-Trilles, F. Casacuberta, Log-linear weight optimisation via bayesian adaptation in statistical machine translation, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Beijing, China, 2010, pp. 1077–1085.

[18] A. Bisazza, N. Ruiz, M. Federico, Fill-up versus interpolation methods for phrase-based SMT adaptation, in: Proceedings of the International Workshop on Spoken Language Translation, International Speech Communication Association, San Francisco, CA, USA, 2011, pp. 136–143.

[19] P. Nakov, Improving English–Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing, in: Proceedings of the Third Workshop on Statistical Machine Translation, Association for Computational Linguistics, Columbus, OH, USA, 2008, pp. 147–150.

[20] P. Koehn, J. Schroeder, Experiments in domain adaptation for statistical machine translation, in: Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 224–227.

[21] H. Wu, H. Wang, Improving domain-specific word alignment with a general bilingual corpus, in: R. E. Frederking, K. B. Taylor (Eds.), Machine Translation: From Real Users to Research, volume 3265 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2004, pp. 262–271.

[22] M. Carpuat, H. Daumé III, A. Fraser, C. Quirk, F. Braune, A. Clifton, et al., Domain adaptation in machine translation: Final report, in: 2012 Johns Hopkins Summer Workshop Final Report, Johns Hopkins University, 2012, pp. 61–72.

[23] M. Eck, S. Vogel, A. Waibel, Language model adaptation for statistical machine translation based on information retrieval, in: M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva (Eds.), Proceedings of the International Conference on Language Resources and Evaluation, European Language Resources Association, Lisbon, Portugal, 2004, pp. 327–330.

[24] R. C. Moore, W. Lewis, Intelligent selection of language model training data, in: Proceedings of the ACL 2010 Conference Short Papers, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 220–224.

[25] A. S. Hildebrand, M. Eck, S. Vogel, A. Waibel, Adaptation of the translation model for statistical machine translation based on information retrieval, in: Proceedings of the 10th Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Budapest, Hungary, 2005, pp. 133–142.

[26] A. Axelrod, X. He, J. Gao, Domain adaptation via pseudo in-domain data selection, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, United Kingdom, 2011, pp. 355–362.

[27] S. Mansour, J. Wuebker, H. Ney, Combining translation and language model scoring for domain-specific data filtering, in: International Workshop on Spoken Language Translation, International Speech Communication Associati, San Francisco, CA, USA, 2011, pp. 222–229.

[28] W. Byrne, D. S. Doermann, M. Franz, S. Gustman, J. Hajič, D. W. Oard, et al., Automatic recognition of spontaneous speech for access to multilingual oral history archives, Speech and Audio Processing, IEEE Transactions on 12 (2004) 420–435.

[29] D. S. Munteanu, D. Marcu, Improving machine translation performance by exploiting non-parallel corpora, Computational Linguistics 31 (2005) 477–504.

[30] H. Daumé III, J. Jagarlamudi, Domain adaptation for machine translation by mining unseen words, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics and Human Language Technologies, Short Papers, Association for Computational Linguistics, Portland, OR, USA, 2011, pp. 407–412.

[31] N. Bertoldi, M. Federico, Domain adaptation for statistical machine translation with monolingual resources, in: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Athens, Greece, 2009, pp. 182–189.

[32] P. Pecina, A. Toral, A. Way, V. Papavassiliou, P. Prokopidis, M. Giagkou, Towards using web-crawled data for domain adaptation in statistical machine translation, in: Proceedings of the 15th Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Leuven, Belgium, 2011, pp. 297–304.

[33] A. Ceausu, J. Tinsley, J. Zhang, A. Way, Experiments on domain adaptation for patent machine translation in the PLuTO project, in: Proceedings of the 15th Annual Meeting of the European Association for Machine Translation, European Association for Machine Translation, 2011, pp. 21–28.

[34] C. Callison-Burch, P. Koehn, C. Monz, O. Zaidan, Findings of the 2011 Workshop on Statistical Machine Translation, in: Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Edinburgh, Scotland, 2011, pp. 22–64.

[35] P. Banerjee, S. K. Naskar, J. Roturier, A. Way, J. van Genabith, Domain adaptation in SMT of user-generated forum content guided by OOV word reduction: Normalization and/or supplementary data?, in: Proceedings of the 16th Annual Meeting of the European Association for Machine Translation, European Association for Machine Translation, Trento, Italy, 2012, pp. 169–176.

[36] A. Bisazza, M. Federico, Cutting the long tail: Hybrid language models for translation style adaptation, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 439–448.

[37] M. Fishel, Y. Georgakopoulou, S. Penkale, V. Petukhova, M. Rojc, M. Volk, A. Way, From subtitles to parallel corpora, in: Proceedins of the 16th Annual Conference of the European Association for Machine Translation EAMT'2012, European Association for Machine Translation, Trento, Italy, 2012, pp. 3–6.

[38] V. Nikoulina, B. Kovachev, N. Lagos, C. 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, Association for Computational Linguistics, Avignon, France, 2012, pp. 109–119.

[39] M. Eck, S. Vogel, A. Waibel, Improving statistical machine translation in the medical domain using the Unified Medical Language System, in: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, Association for Computational Linguistics, Geneva, Switzerland, 2004, pp. 792–798.

[40] U.S. National Library of Medicine, UMLS reference manual, 2009. Metathesaurus. Bethesda, MD, USA.

[41] C. Wu, F. Xia, L. Deleger, I. Solti, Statistical machine translation for biomedical text: are we there yet?, AMIA Annual Symposium proceedings (2011) 1290–1299.

[42] M. R. Costa-jussà, M. Farrús, J. S. Pons, Machine translation in medicine. A quality analysis of statistical machine translation in the medical domain, in: Proceedings of the 1st Virtual International Conference on Advanced Research in Scientific Areas, Žilinská univerzita, Žilina, Slovakia, 2012, pp. 1995–1998.

[43] A. Jimeno Yepes, É. Prieur-Gaston, A. Névéol, Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text, BMC Bioinformatics 14 (2013) 1–10.

[44] A. Chen, Cross-language retrieval experiments at CLEF 2002, in: C. Peters, M. Braschler, J. Gonzalo (Eds.), Advances in Cross-Language Information Retrieval, volume 2785 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Rome, Italy, 2002, pp. 28–48.

[45] P. Koehn, K. Knight, Empirical methods for compound splitting, in: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Budapest, Hungary, 2003, pp. 187–193.

[46] M. Popović, D. Stein, H. Ney, Statistical machine translation of German compound words, in: Proceedings of the 5th international conference on Advances in Natural Language Processing, Springer-Verlag, Turku, Finland, 2006, pp. 616–624.

[47] S. Niessen, H. Ney, Improving SMT quality with morpho-syntactic analysis, in: Proceedings of the 18th International Conference on Computational Linguistics, volume 2, Association for Computational Linguistics, Saarbrücken, Germany, 2000, pp. 1081–1085.

[48] E. Alfonseca, S. Bilac, S. Pharies, Decompounding query keywords from compounding languages, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Short Papers, Association for Computational Linguistics, Columbus, OH, USA, 2008, pp. 253–256.

[49] H. Wu, M. Zhou, Optimizing synonym extraction using monolingual

and bilingual resources, in: Proceedings of the second international workshop on Paraphrasing, volume 16, Association for Computational Linguistics, Sapporo, Japan, 2003, pp. 72–79.

[50] G. J. Jones, F. Fantino, E. Newman, Y. Zhang, Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia, in: CLIA 2008 - 2nd International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, Hyderabad, India, 2008, pp. 34–41.

[51] X. Han, H. Li, T. Zhao, Train the machine with what it can learn: corpus selection for SMT, in: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora, Association for Computational Linguistics, Singapore, 2009, pp. 27–33.

[52] N. Griffon, W. Chebil, L. Rollin, G. Kerdelhue, B. Thirion, J.-F. Gehanno, et al., Performance evaluation of Unified Medical Language System's synonyms expansion to query PubMed, BMC medical informatics and decision making 12 (2012) 12.

[53] K. Nakayama, M. Pei, M. Erdmann, M. Ito, M. Shirakawa, T. Hara, S. Nishio, Wikipedia mining – Wikipedia as a corpus for knowledge extraction, in: Proceedings of Annual Wikipedia Conference (Wikimania), Alexandria, Egypt, 2008.

[54] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia – a crystallization point for the web of data, Web Semantics: Science, Services and Agents on the World Wide Web 7 (2009) 154–165.

[55] J. Tiedemann, News from OPUS – a collection of multilingual parallel corpora with tools and interfaces, in: Recent Advances in Natural Language Processing, volume 5, John Benjamins, Borovets, Bulgaria, 2009, pp. 237–248.

[56] P. Buitelaar, B. Sacaleanu, Špela Vintar, D. Steffen, M. Volk, H. Dejean, E. Gaussier, D. Widdows, O. Weiser, R. Frederking, Multilingual Concept Hierarchies for Medical Information Organization and Retrieval, Public deliverable, MuchMore project, 2003.

[57] K. Wäschle, S. Riezler, Analyzing parallelism and domain similarities in the MAREC patent corpus, in: M. Salampasis, B. Larsen (Eds.), Multidisciplinary Information Retrieval, volume 7356 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2012, pp. 12–27.

[58] B. Pouliquen, C. Mazenc, COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent barrier at WIPO, in: Proceedings of the Thirteenth Machine Translation Summit, Asia-Pacific Association for Machine Translation, Xiamen, China, 2011, pp. 24–30.

[59] J. R. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, A. 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), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 1374–1383.

[60] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, F. M. Tyers, Apertium: a free/open-source platform for rule-based machine translation, Machine Translation 25 (2011) 127–144.

[61] O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, A. Tamchyna, The joy of parallelism with CzEng 1.0, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation, European Language Resources Association, Istanbul, Turkey, 2012, pp. 3921–3928.

[62] A. Eisele, Y. Chen, MultiUN: A multilingual corpus from United Nation documents, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC-2010), European Language Resources Association, Valletta, Malta, 2010, pp. 2868–2872.

[63] K. Dickersin, E. Manheimer, S. Wieland, K. A. Robinson, C. Lefebvre, S. McDonald, Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials, Evaluation & the Health Professions 25 (2002) 38–64.

[64] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, D. S. Wishart, DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs, Nucleic acids research 39 (2011) D1035–D1041.

[65] P. Thompson, S. Iqbal, J. McNaught, S. Ananiadou, Construction of an annotated corpus to support biomedical information extraction, BMC bioinformatics 10 (2009) 349.

[66] J.-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, GENIA corpus – a semantically annotated corpus for bio-textmining, Bioinformatics 19 (2003) i180–i182.

[67] C. Rosse, J. L. V. Mejino Jr., The foundational model of anatomy ontology, in: A. Burger, D. Davidson, R. Baldock (Eds.), Anatomy Ontologies for Bioinformatics, volume 6 of *Computational Biology*, Springer London, 2008, pp. 59–117.

[68] F. Rogers, Medical subject headings, Bulletin of the Medical Library Association 51 (1963) 114–116.

[69] N. Bouayad-Agha, D. R. Scott, R. Power, Integrating content and style in documents: A case study of patient information leaflets, Information Design Journal 9 (2000) 161–176.

[70] N. Shuyo, Language detection library for Java, 2010. http://code.google.com/p/language-detection/ (accessed 1 January, 2014).

[71] M. Majliš, Yet another language identifier, in: Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 46–54.

[72] C. Boyer, W. Belle, N. Pletneva, N. Lawson, M. Samwald, A. Hanbury, Prototype of a first search system for intensive tests (D8.3), Public deliverable, Khresmoi EU project, 2012.

[73] R. Parker, D. Graff, J. Kong, K. Chen, K. Maeda, English Gigaword fifth edition, 2011. Linguistic Data Consortium, Philadelphia, PA, USA.

[74] M. Popel, Z. Žabokrtský, TectoMT: Modular NLP framework, in: H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.), Advances in Natural Language Processing, volume 6233 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Reykjavik, Iceland, 2010, pp. 293–304.

[75] D. Spoustová, J. Hajič, J. Votrubec, P. Krbec, P. Květoň, The best of two worlds: Cooperation of statistical and rule-based taggers for Czech, in: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 67–74.

[76] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: Proceedings of international conference on new methods in language processing, volume 12, Manchester, United Kingdom, 1994, pp. 44–49.

[77] J. Hajič, Disambiguation of rich inflection (Computational morphology of Czech), Nakladatelství Karolinum, Prague, Czech Republic, 2004.

[78] M. Popel, Z. Žabokrtský, Improving English–Czech tectogrammatical MT, The Prague Bulletin of Mathematical Linguistics 92 (2009) 1–20.

[79] C. Boyer, M. Gschwandtner, A. Hanbury, M. Kritz, N. Pletneva, M. Samwald, A. Vargas, Use case definition including concrete data requirements (D8.2), Public deliverable, Khresmoi project, 2012.

[80] E. Meats, J. Brassey, C. Heneghan, P. Glasziou, Using the Turning Research Into Practice (TRIP) database: how do clinicians really search?, Journal of the Medical Library Association 95 (2007) 156–163.

[81] A. Spink, D. Wolfram, M. B. J. Jansen, T. Saracevic, Searching the web: The public and their queries, Journal of the American Society for Information Science and Technology 52 (2001) 226–234.

[82] C. Dyer, V. Chahuneau, N. 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: Human Language Technologies, Association for Computational Linguistics, Atlanta, GA, USA, 2013, pp. 644–648.

[83] A. Stolcke, SRILM – an extensible language modeling toolkit, in: Proceedings of International Conference on Spoken Language Processing, International Speech Communication Association, Denver, CO, USA, 2002, pp. 901–904.

[84] R. Kneser, H. Ney, Improved backing-off for N-gram language modeling, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, volume 1, IEEE, Detroit, MI, USA, 1995, pp. 181–184.

[85] K. Heafield, KenLM: faster and smaller language model queries, in: Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Edinburgh, United Kingdom, 2011, pp. 187–197.

[86] P. Koehn, A. Axelrod, A. Birch, C. Callison-Burch, M. Osborne, D. Tal-

bot, Edinburgh system description for the 2005 IWSLT speech translation evaluation, in: Proceedings of the International Workshop on Spoken Language Translation 2005, Hong Kong, 2005, pp. 78–85.

[87] O. Bojar, R. Rosa, A. Tamchyna, Chimera – three heads for English-to-Czech translation, in: Proceedings of the Eighth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 92–98.

[88] O. Bojar, A. Tamchyna, The design of Eman, an experiment manager, Prague Bulletin of Mathematical Linguistics 100 (2013) 39–58.

[89] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady 10 (1966) 707–710.

[90] P. Koehn, Statistical significance tests for machine translation evaluation, in: D. Lin, D. Wu (Eds.), Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 388–395.

[91] O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, et al., Findings of the 2013 Workshop on Statistical Machine Translation, in: Proceedings of the Eighth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 1–44.

[92] P. Pecina, A. Toral, J. van Genabith, Simple and effective parameter tuning for domain adaptation of statistical machine translation, in: M. Kay, C. Boitet (Eds.), Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), Coling 2012 Organizing Committee, Mumbai, India, 2012, pp. 2209–2224.

[93] M. F. Porter, An algorithm for suffix stripping, Program: electronic library and information systems 14 (1980) 130–137.

[94] D. W. Oard, G.-A. Levow, C. I. Cabezas, CLEF experiments at Maryland: Statistical stemming and backoff translation, in: C. Peters (Ed.), Cross-Language Information Retrieval and Evaluation, volume 2069 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2001, pp. 176–187.

[95] M. F. Porter, Snowball: A language for stemming algorithms, 2001. http://snowball.tartarus.org/ (accessed 1 January, 2014).

[96] O. Bojar, M. Ercegovčević, M. Popel, O. F. Zaidan, A grain of salt for the WMT manual evaluation, in: Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Edinburgh, Scotland, 2011, pp. 1–11.

[97] W. Magdy, G. J. F. Jones, An efficient method for using machine translation technologies in cross-language patent search, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, Glasgow, United Kingdom, 2011, pp. 1925–1928.

[98] C. E. Mooers, Coding, information retrieval, and the rapid selector, American Documentation 1 (1950) 225–229.

[99] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.

[100] W. Hersh, C. Buckley, T. Leone, D. Hickam, OHSUMED: An interactive retrieval evaluation and new large test collection for research, in: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc., Dublin, Ireland, 1994, pp. 192–201.

[101] P. M. Roberts, A. M. Cohen, W. R. Hersh, Tasks, topics and relevance judging for the TREC Genomics Track: five years of experience evaluating biomedical text information retrieval systems, Information Retrieval 12 (2009) 81–97.

[102] E. M. Voorhees, R. M. Tong, Overview of the TREC 2011 Medical Records Track, in: The Eleventh Text Retrieval Conference (TREC 2002), National Institute of Standards and Technology, Gaithersburg, MD, USA, 2011, pp. 1–11.

[103] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, et al., Overview of the ShARe/CLEF eHealth evaluation lab 2013, in: P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative, CLEF 2013, volume 8138 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Valencia, Spain, 2013, pp. 212–231.

[104] M. Ruiz, A. Diekema, P. Sheridan, D. C. Plaza, CINDOR conceptual interlingua document retrieval: TREC-8 evaluation, in: The Eighth Text REtrieval Conference (TREC 8), National Institute of Standards and Technology, Gaithersburg, MD, USA, 1999, pp. 597–605.

[105] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[106] J.-Y. Nie, Cross-language information retrieval, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2010.

[107] C. Peters, M. Braschler, P. Clough, Multilingual information retrieval: From research to practice, Springer Berlin Heidelberg, 2012.

[108] D. Zhou, M. Truran, T. Brailsford, V. Wade, H. Ashman, Translation techniques in cross-language information retrieval, ACM Computing Surveys 45 (2012) 1:1–1:44.

[109] L. Ballesteros, W. B. Croft, Resolving ambiguity for cross-language retrieval, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Association for Computing Machinery, Melbourne, Australia, 1998, pp. 64–71.

[110] A. Maeda, F. Sadat, M. Yoshikawa, S. Uemura, Query term disambiguation for Web cross-language information retrieval using a search engine, in: Proceedings of the fifth international workshop on on Information retrieval with Asian languages, Association for Computing Machinery, Hong Kong, China, 2000, pp. 25–32.

[111] J. Gao, J.-Y. Nie, A study of statistical models for query translation: finding a good unit of translation, in: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Association for Computing Machinery, Seattle, Washington, USA, 2006, pp. 194–201.

[112] A. 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, Association for Computing Machinery, Melbourne, Australia, 1998, pp. 55–63.

[113] K. Darwish, D. W. Oard, Probabilistic structured query methods, in: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Association for Computing Machinery, Toronto, Canada, 2003, pp. 338–344.

[114] M. Federico, N. Bertoldi, Statistical cross-language information retrieval using n-best query translations, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Association for Computing Machinery, Tampere, Finland, 2002, pp. 167–174.

[115] V. Hollink, J. Kamps, C. Monz, M. de Rijke, Monolingual document retrieval for European languages, Information Retrieval 7 (2004) 33–52.

[116] R. Attar, A. S. Fraenkel, Local feedback in full-text retrieval systems, Journal of Association for Computing Machinery 24 (1977) 397–417.

[117] L. Ballesteros, W. B. Croft, Phrasal translation and query expansion techniques for cross-language information retrieval, SIGIR Forum 31 (1997) 84–91.

[118] D. Eichmann, M. E. Ruiz, P. Srinivasan, Cross-language information retrieval with the UMLS metathesaurus, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Association for Computing Machinery, Melbourne, Australia, 1998, pp. 72–80.

[119] M. Volk, B. Ripplinger, Š. Vintar, P. Buitelaar, D. Raileanu, B. Sacaleanu, Semantic annotation for concept-based cross-language medical information retrieval, International Journal of Medical Informatics 67 (2002) 97–112.

[120] T. D. Tran, N. Garcelon, A. Burgun, P. L. Beux, Experiments in cross-language medical information retrieval using a mixing translation module, Studies in Health Technology and Informatic 107 (2004) 946–949.

[121] H. Déjean, E. Gaussier, J.-M. Renders, F. Sadat, Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval, Artificial Intelligence in Medicine 33 (2005) 111–124.

[122] G. Rosemblat, D. Gemoets, A. C. Browne, T. Tse, Machine translation-supported cross-language information retrieval for a consumer health resource, AMIA Annual Symposium proceedings (2003) 564–568.

[123] K. Markó, S. Schulz, U. Hahn, MorphoSaurus–design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain, Methods of information in medicine 44 (2005) 9.

[124] K. G. Markó, P. Daumke, S. Schulz, R. Klar, U. Hahn, Large-scale evaluation of a medical cross-language information retrieval system, in: K. A. Kuhn, J. R. Warren, T.-Y. Leong (Eds.), Proceedings of the 12th World Congress on Health (Medical) Informatics – Building Sustain-

able Health Systems, volume 129 of *Studies in Health Technology and Informatics*, IOS Press, Brisbane, Australia, 2007, pp. 392–396.

[125] L. Goeuriot, G. J. F. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, et al., ShARe/CLEF eHealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports, in: D. T. Pamela Forner, Roberto Navigli (Ed.), CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, Valencia, Spain, 2013.

[126] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: Overview of the Third Text Retrieval Conference (TREC-3), National Institute of Standards and Technology, Gaithersburg, MD, USA, 1995, pp. 109–126.

[127] S. E. Robertson, S. Walker, M. Beaulieu, Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track, in: D. K. Harman (Ed.), The Seventh Text REtrieval Conference (TREC-7), National Institute of Standards and Technology, Gaithersburg, MD, USA, 1998, pp. 253–264.

[128] J. Leveling, L. Goeuriot, L. Kelly, G. J. Jones, DCU@TRECMed 2012: Using ad-hoc baselines for domain-specific retrieval, in: Text Retrieval Conference (TREC) 2012, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2012, pp. 1–9.

[129] S. E. Robertson, K. Spärck Jones, Relevance weighting of search terms, Journal of the American Society for Information Science 27 (1976) 143–160.

[130] S. E. Robertson, On term selection for query expansion, Journal of Documentation 46 (1990) 359–364.

[131] G. Salton, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[132] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems 20 (2002) 422–446.

[133] E. M. Voorhees, D. K. Harman (Eds.), TREC: Experiment and evaluation in information retrieval, volume 63 of *Digital libraries and electronic publishing series*, MIT press Cambridge, Cambridge, MA, USA, 2005.

[134] D. Hull, Using statistical testing in the evaluation of retrieval experiments, in: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '93, Association for Computing Machinery, Pittsburgh, PA, USA, 1993, pp. 329–338.

[135] L. Goeuriot, L. Kelly, G. J. F. Jones, G. Zuccon, H. Suominen, A. Hanbury, et al., Creation of a new evaluation benchmark for information retrieval targeting patient information needs, in: R. Song, W. Webber, N. Kando, K. Kishida (Eds.), Proceedings of the 5th International Workshop on Evaluating Information Access (EVIA), National Institute of Informatics/Kijima Printing, Tokyo/Fukuoka, Japan, 2013, pp. 29–32.