Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain

Shadi Saleh and **Pavel Pecina**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague, Czech Republic
{saleh,pecina}@ufal.mff.cuni.cz

Abstract

We present a thorough comparison of two principal approaches to Cross-Lingual Information Retrieval: document translation (DT) and query translation (QT). Our experiments are conducted using the cross-lingual test collection produced within the CLEF eHealth information retrieval tasks in 2013–2015 containing English documents and queries in several European languages. We exploit the Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) paradigms and train several domain-specific and task-specific machine translation systems to translate the non-English queries into English (for the QT approach) and the English documents to all the query languages (for the DT approach). The results show that the quality of QT by SMT is sufficient enough to outperform the retrieval results of the DT approach for all the languages. NMT then further boosts translation quality and retrieval quality for both QT and DT for most languages, but still, QT provides generally better retrieval results than DT.

1 Introduction

Multilingual content has been growing significantly in the last few years simultaneously with rapid internet access growth all over the world. Monolingual information retrieval task allows users to find information in documents that are written in the language that they use to write their queries. This ignores a vast amount of information that is represented in other languages. Cross-Lingual Information Retrieval (CLIR) breaks this language barrier by allowing users to look up information that is represented in documents written in languages different from the language of the query.

We reinvestigate the effectiveness of two principal approaches to CLIR: document translation (DT) and query translation (QT). The existing comparison studies of the two approaches are outdated (e.g. Oard, 1998) and do not reflect the current advances in Machine Translation (MT). Even in very recent works, the authors have blindly assumed that DT is superior to QT (Khiroun et al., 2018), giving the argument that in DT, the text is translated in a larger context compared to the translation of short isolated queries in QT. The larger context should help in translation disambiguation and better lexical selection during translation, which should subsequently lead to better retrieval results.

This hypothesis needs to be revised, taking into consideration the significant improvement of machine translation quality in recent years, despite the strong practical disadvantages of DT over QT: DT is computationally expensive and hard to scale (every document needs to be translated into each supported language and then indexed) while QT is performed in query time and only a short text (the query) is translated into the document language.

In this work, state-of-the-art Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) systems are deployed for document translation and query translation to investigate their effect on retrieval quality in the cross-lingual setting. The experiments are conducted using the cross-lingual test collection produced within the CLEF eHealth tasks on patient-centered information retrieval in 2013–2015 extended with additional relevance assessments and manual query translations (Saleh and Pecina, 2019). Though this is a very specific domain and the results cannot be thoughtlessly generalized to other domains, the choice of this test collection was motivated by two facts: First, it provides resources for large-scale experimentation (1 million in-domain documents, 166 queries in 8 languages, thorough relevance assessment). Second, the medical domain in MT has been well studied (Jimeno Yepes et al., 2017; Dušek et al., 2014), and there are enough resources to develop well-performing MT systems for multiple languages.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6849–6860
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

2 Related work

In CLIR, documents and queries are written in different languages. The traditional term-matching retrieval methods require both documents and queries to be represented in the same language. In practice, either the queries need to be translated into the document language (QT), or the documents need to be translated into the query language (DT). Not many studies and experiments have been conducted in order to compare these two approaches.

Oard (1998) investigated the performance of DT, QT, and a hybrid system combining both. They found that the system translating English queries into German (the document language) outperformed the system translating the documents from German into English (the query language). They hypothesized that documents, which are typically longer than queries, provide more contextual and linguistic information that helps reduce translation ambiguity and thus improves translation quality. McCarley (1999) presented a hybrid DT/QT system, which averaged the retrieved document scores from DT and QT systems and thus outperformed both of them. Fujii and Ishikawa (2000) employed a two-step method where QT was first used to retrieve a limited number of documents that were translated into the query language and reranked by their DT retrieval scores.

Pirkola (1998) presented a new method for CLIR, which was referred to as *structured queries*. The idea was that a document containing one possible translation candidate of a query term is more relevant than a document that contains multiple translations of that term. This probabilistic structured queries approach was also applied to Cross-Language Speech Retrieval (Nair et al., 2020). Darwish and Oard (2003) also exploited alternative translations of query terms. Their experiments showed that combining multiple translations outperformed the selection of one best translation.

Nikoulina et al. (2012) investigated reranking SMT translation hypotheses towards better CLIR performance and showed that SMT systems are usually trained to give the best results in terms of translation accuracy, adequacy, and fluency. However, an improvement will be achieved when they are optimized towards retrieval quality. We followed this approach in our previous work and introduced a richer set of features and adopted the hypothesis reranker for multiple languages in the medical domain (Saleh and Pecina, 2016b,a).

Several recent papers employed methods based on Deep Learning. Litschko et al. (2018) presented an unsupervised CLIR approach employing shared cross-lingual word embedding model, which was trained using monolingual data only. They used those embeddings to translate query terms word by word into the document language. Rücklé et al. (2019) trained NMT model for CLIR using out-domain data and synthetic data (created by translating in-domain monolingual English into German) to retrieve answers to German questions from English collection in the technical domain (AskUbuntu and StackOverflow).

CLIR in the medical domain has been investigated within the series of CLEF ShARe/eHealth labs since 2013 which focused on improving access of laypeople (non-medical experts) to reliable medical information (Goeuriot et al., 2013, 2014; Palotti et al., 2015; Kelly et al., 2016; Palotti et al., 2017; Jimmy et al., 2018; Kelly et al., 2019).

In this paper, we compare the performance of both QT and DT using the traditional SMT and state-of-the-art NMT methods trained on the same data to make the comparison as fair as possible. We present a novel approach for NMT model selection that is optimized towards CLIR performance and investigate the effect of morphological pre- and post-processing on the performance on CLIR.

3 Data

Two types of data were used in our experiments: The data for training, tuning, and testing MT (Section 3.1) and the CLIR test collection (Section 3.2).

3.1 Machine Translation Resources

Parallel data is essential for training both SMT and NMT systems. We exploited the UFAL Medical Corpus[1] which was assembled during the course of several EU projects aiming at more reliable machine translation of medical texts and used for the purposes of WMT Biomedical Translation Task (Bojar et al., 2014). It mainly includes the EMEA corpus by Tiedemann (2009), UMLS metathesaurus (Humphreys et al., 1998), titles from Wikipedia articles in the medical categories mapped to other languages using Wikipedia Interlingual links, medical domain patent applications (Wäschle and Riezler, 2012; Pouliquen and Mazenc, 2011), and various web-crawled data.

[1] `http://ufal.mff.cuni.cz/ufal_medical_corpus`

Monolingual data is used to build a language model during the development of SMT systems. The language model helps select a candidate translation that is as coherent and fluent as possible in the target language (which is certainly important for document translation, but less important for query translation). Our procedure of data selection (both parallel and monolingual data) follows the work of Pecina et al. (2014), where two language models are trained on in-domain and general-domain data respectively, then each sentence from the corpus is scored by its cross-perplexity between the two models. Finally, the top 10 million scored sentences are chosen. In NMT training, the monolingual data is used to enlarge the parallel data training data by back-translation, where target language monolingual data is machine translated to the source language and added to parallel data for training. The monolingual data used in our experiments includes multiple resources such as the CLEF eHealth 2014 English document collection (Goeuriot et al., 2014), Genia corpus (Ohta et al., 2002), and medical Wikipedia articles in English.

MT development and test data: used for tuning and evaluating our MT systems consists of the Khresmoi Summary Translation Test Data[2] used by the DT models and Khresmoi Query Translation Test Data 2.0[3] used by the QT models. Both were developed within the Khresmoi project[4] and later extended within the KConnect[5] and HimL[6] projects. The summary test data includes sentences (1,000 for testing and 500 for development) from summaries of English medical articles manually translated from English to all relevant languages. The query test data includes English queries (1,000 for testing and 500 for tuning) sampled from a query log of a medical search engine and manually translated to the same set of languages.

3.2 CLIR Test Collection

For CLIR experiments, we use the CLIR test collection[7]. that we developed in our previous work (Saleh and Pecina, 2019). It is based on the data used within the CLEF eHealth lab IR tasks in 2013–2015 (Suominen et al., 2013; Goeuriot et al., 2014; Palotti et al., 2015). It contains about 1.1 million web pages that were crawled automatically from various trusted medical websites (Goeuriot et al., 2015). There are 166 queries in total (100 for training and 66 for testing) originally formulated in English (to mimic real patient queries) and then manually translated by medical experts into seven European languages (Czech, French, German, Spanish, Swedish, Polish, and Hungarian). The relevance judgments consist of the official relevance assessments provided by the task organizers and additional assessments, as described in (Saleh and Pecina, 2019).

We clean the document collection by removing HTML tags and other scripts in the documents. All the lemmatization experiments in our work are done using UDPipe (Straka and Straková, 2017), while for stemming, we use the Snowball algorithm (Moral et al., 2014).

4 Retrieval System

The document collection is indexed using Terrier (Ounis et al., 2005), an open-source tool for information retrieval experiments. For retrieval, we use Terrier's implementation of the language model with Bayesian smoothing and Dirichlet prior (Smucker and Allan, 2005) with the default value of the smoothing parameter.

5 Machine Translation Systems

In this section, we provide details on training the SMT and NMT systems used in the CLIR experiments. The SMT systems fully replicate the work by Dušek et al. (2014); we only provide the most important information. The NMT systems are described in full detail.

5.1 Statistical Machine Translation

The SMT systems are based on the phrase-based SMT paradigm implemented in Moses (Koehn et al., 2007). The system for the QT experiments was developed within the Khresmoi project (Dušek et al., 2014). The system was tuned to translate medical search queries (using the Khresmoi Query development set) and optimized towards PER (Position-independent word Error Rate, Tillmann et al., 1997) instead of the traditionally preferred BLEU (Papineni et al., 2002) as this was shown to be more effective for tuning SMT parameters for translating search queries (Pecina et al., 2014). The system is denoted as *QT-SMT-form*.

[2] `http://hdl.handle.net/11234/1-2122`
[3] `http://hdl.handle.net/11234/1-2121`
[4] `http://khresmoi.eu/`
[5] `http://www.kconnect.eu/`
[6] `http://www.himl.eu/`
[7] `http://hdl.handle.net/11234/1-2925`

For the DT experiments, we train two SMT systems: *DT-SMT-form*, which is a replication of the SMT system that translates standard sentences by Dušek et al. (2014), and our own system *DT-SMT-pre-lem* that translates English sentences into lemmatized sentences in the target language. This is done by lemmatizing the monolingual data and the target side of the parallel data prior to training. In both the systems, we use *fast_align* (Dyer et al., 2013) to train word alignment model on the lowercased word forms between English and the target language, then we replace the word forms in the target language with word lemmas. Moses (with its default settings) is used to train a phrase-table model using the tokenized and lowercased English word forms, and the tokenized and lemmatized data in the target language plus a 5-gram language model. Minimum Error Rate Training (MERT, Och, 2003) is used to tune the model parameter weights using the development data sets. We also experiment with another system (*DT-SMT-post-lem*), which produces lemmatized output but obtained as post-lemmatization of the output of the *DT-SMT-form* system, and a system (*DT-SMT-post-stem*) which produces stemmed output obtained by the Snowball stemmer applied again to the output of *DT-SMT-form*. This is to allow better comparison of the DT and QT approaches. Translating documents into a morphologically richer language enlarge the vocabulary (term diversity) and thus make retrieval more difficult. The three systems produce morphologically reduced translations of documents and thus make them comparable to the English ones (in terms of vocabulary size).

5.2 Neural Machine Translation

Neural Machine Translation (NMT) has become the state-of-the-art approach in MT and recently achieved superior results and lead to a significant improvement over the SMT systems (Jean et al., 2015). We implement two types of NMT systems: one for query translation (denoted as *QT-NMT-form*) and one for document translation (denoted as *DT-NMT-form*). Both produce standard (non-lemmatized) output.

The systems are based on the Marian (Junczys-Dowmunt et al., 2018) implementation of the Transformer (Vaswani et al., 2017) model with back-translation (Edunov et al., 2018). SMT has an advantage over NMT in employing monolingual data in its language model. This gap can be bridged by back-translation, a technique that exploits another MT model to translate monolingual data from the target language into the source language and adds this "synthetic" data to the original parallel data(Sennrich et al., 2016a). This approach also helps for domain adaption of NMT when the monolingual data is taken from a specific domain. We follow the back-translation approach in this work iteratively.

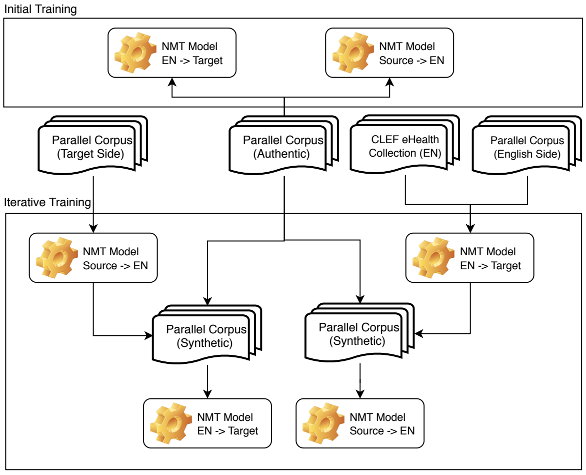
Initial Training
NMT Model EN -> Target
NMT Model Source -> EN
Parallel Corpus (Target Side)
Parallel Corpus (Authentic)
CLEF eHealth Collection (EN)
Parallel Corpus (English Side)
Iterative Training
NMT Model Source -> EN
NMT Model EN -> Target
Parallel Corpus (Synthetic)
Parallel Corpus (Synthetic)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

Figure 1: A schema of the iterative back-translation mechanism for NMT training.

5.2.1 Task-Oriented NMT Training

The NMT systems are trained using the same training data as the SMT systems. However, in NMT, all data sets (monolingual and parallel) are encoded into Byte-Pair Encoding (BPE), which helps reduce the out-of-vocabulary problem in NMT by encoding rare words as sequences of subword units (Sennrich et al., 2016b). We train the Transformer model using the same parameters as reported by Vaswani et al. (2017). Figure 1 shows the architecture of the proposed iterative back-translation NMT model, inspired by the work of Hoang et al. (2018): for each language pair, we first train initial models for both directions, English to target, and source to English. We use the authentic (non-synthetic) parallel data that is presented in Section 3.1 for training the initial models.

During training the Transformer models, multiple epochs (iterations through the entire training data) are needed. It is known that too many training epochs can cause over-fitting of the model, and a few iterations might cause under-fitting (Popel and Bojar, 2018). To avoid this, the early-stopping of the training is employed to terminate the process when the intermediate model satisfies some stopping criteria (training objective). We stop training when there are three consecutive checkpoints without any improvement in the translation performance of the validation data. Then, we use the initial model to translate monolingual text in the target language coming from two resources:

MT parallel training corpus: the target side of the parallel training data (Section 3.1) is translated into English using the *SRC→EN* NMT model to create the synthetic data for the models that are used in DT experiments. The English side of the parallel corpus is translated using the *EN→TGT* model for the QT experiment. This is done to investigate the effect of the source of the monolingual data on the CLIR performance. We randomly select 2 million sentences in each iteration.

CLIR test collection: we select randomly 2 million sentences from the test collection (Section 3.2) (after filtering sentences that are longer than 80 words), then we use *EN→TGT* model to translate them into the target language. This is done for models that are used for the query translation approach. The motivation of choosing the collection is to make the model adapted to translate the medical queries into English (the document language).

After translating this monolingual data, we create the synthetic data by adding the monolingual data and their translations to the authentic parallel data. Then we continue training of the models in both directions. We conduct back-translation three times, and in each iteration, we use the updated models from the previous one.

5.2.2 NMT Model Selection

We setup Marian to save the intermediate models (checkpoints) after every 5,000 iterations where each iteration is a batch sized of instances from the training data. This is done instead of saving each epoch to avoid loosing effective intermediate models in between. The model selection is based on evaluating each checkpoint by BLEU (Papineni et al., 2002) and PER (Tillmann et al., 1997) using the Khresmoi Summary development set (DT) and Khresmoi Query development set (QT).

Figure 2 shows the evaluation results of the intermediate models using the two MT metrics and how they correlate with P@10 (IR metric). P@10 is calculated by query translation of the Czech training queries into English using the corresponding NMT model, and then conducting retrieval as we describe in Section 4. Choosing the model that gives the best BLEU scores (iteration 400,000) does not correlate with the best value for P@10, nor the best score for PER (500,000). This is understandable because these metrics evaluate translation quality.

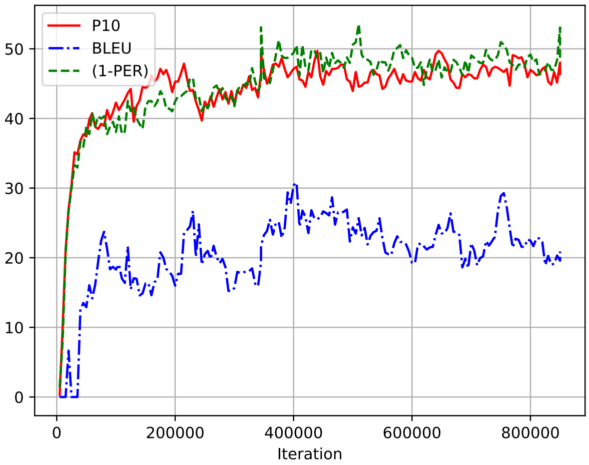

Figure 2: Performance comparison of the intermediate *QT-NMT-form* models at each checkpoint (after each 5,000 iterations) in terms of BLEU, 1-PER, and P@10 when employed in the Czech QT CLIR system.

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

6 Experiments and Results

6.1 MT Evaluation

In this section, we present *intrinsic* evaluation of the MT systems. We evaluate how well the systems translate sentence/queries given their reference translations in the test data. We present both BLEU and PER scores (all as percentages). The higher the BLEU score, the better the translation quality is. BLEU is based on measuring the similarity of n-grams counts between a translation hypothesis and its reference translation(s), and as such is sensitive to word order. PER, on the other hand, does not penalize word order between a translation hypothesis and its reference translation as BLEU does. Instead, it considers both as a “bag of words”. PER captures all words that appear in a translation hypothesis but do not exist in the reference. These words are known as PER errors; thus, the higher the PER value, the lower the translation quality.

	EN–CS		EN–FR		EN–DE		EN–HU		EN–ES		EN–SV		EN–PL	
MT System	BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER
DT-SMT-form	19.0	**51.1**	37.8	68.3	18.7	53.4	**10.5**	41.6	**25.7**	63.2	33.6	64.6	**11.5**	41.3
DT-NMT-form	**25.9**	56.5	**38.8**	**66.5**	**19.8**	**51.4**	8.2	**39.5**	23.2	**55.2**	**35.1**	**64.4**	10.2	**35.9**
DT-SMT-post-lem	30.9	65.6	43.5	74.7	23.6	60.4	13.2	48.6	35.4	72.3	40.9	69.9	16.1	50.5
DT-SMT-pre-lem	28.7	64.2	41.2	72.6	13.0	48.0	14.3	51.9	28.4	65.7	39.1	70.0	12.5	46.9

Table 1: Intrinsic evaluation of MT systems for **document translation** using the Khresmoi Summary Test set.

	CS–EN		FR–EN		DE–EN		HU–EN		ES–EN		SV–EN		PL–EN	
MT System	BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER	BLEU	PER
QT-SMT-form	**36.4**	70.2	**38.7**	75.9	**37.0**	65.2	**39.7**	67.3	**31.2**	73.7	39.2	**62.7**	**26.0**	58.6
QT-NMT-form	22.5	**48.9**	30.6	**65.4**	28.7	**58.1**	36.7	**63.2**	17.8	**45.5**	**40.9**	63.0	18.7	**47.9**

Table 2: Intrinsic evaluation of MT systems for **query translation** using the Khresmoi Query Test set.

The MT evaluation scores cannot be directly compared across language pairs, and for the **-form* and **-lem* systems (since the test sets differ), but they indicate to what extent the translated queries differ from the reference translations, which in term-matching IR is important. Also, the results of the two systems producing lemmas instead of the word forms are indicative only. They cannot be directly compared to those producing word forms.

Table 1 displays the (intrinsic) evaluation of the MT systems for document translation using the Khresmoi Summary test set (in terms of BLEU and PER). The results are not very consistent: For six out of the seven translation directions, *DT-NMT-form* outperforms *DT-SMT-form* in terms of PER. In terms of BLEU, *DT-NMT-form* wins for four language pairs.

The effect of lemmatization on the scores is not surprising. Naturally, lemmatization reduces the vocabulary size in the target language; thus, the BLEU scores are higher for the systems which employ lemmatization in either way. However, post-lemmatization is constantly better (with the exception of Hungarian, which is a very specific language, and its scores are generally much lower than for other languages). In terms of PER, the situation is different, and despite the fact that lemmatization reduces the target language, the systems without lemmatization often achieve better scores (except in German and Spanish).

Table 2 presents the (intrinsic) evaluation of the MT systems for QT using the Khresmoi Query test set. *QT-SMT-form* outperforms *QT-NMT-form* in terms of BLEU in all the languages except Swedish. However, in terms of PER (which is preferred), *QT-NMT-form* is always better. This can be partially explained because of the way we ensembled NMT models towards better CLIR performance. The bold font indicates which of the two **-form* systems is better (for each language pair and each measure).

6.2 CLIR experiments

Table 3 presents the results of the CLIR experiments altogether. Motivated by the organization of the CLEF eHealth CLIR tasks, we adopt P@10 (the percentage of relevant documents among the top ten retrieved ones) as the main evaluation measure. In all the experiments, all the top 10 ranked documents for each query are assessed for relevance. We also report MAP (Mean Average Precision) as a secondary evaluation measure. The **-SMT-form* systems are treated as baselines. The figures in bold denote results better than the baseline. Those, which are statistically significantly better are in bold and also in italics. The significance tests were performed using the paired Wilcoxon signed-rank test (Hull, 1993) with $\alpha = 0.05$, and no correction was applied.

First, we conduct monolingual experiments using the English queries and the English document collection to set a reference (oracle) system for our CLIR task, that is why all the results of monolingual systems are the same for all the languages. We report the following: *Mono-form* system uses the original English queries and the English collection (no morphological processing applied). *Mono-lem* and *Mono-stem* report the results after performing lemmatization and stemming of the document collection and the English queries, respectively. The purpose of these systems is to study the effect of the morphological processing of the English documents on retrieval performance.

	Czech		French		German		Hungarian		Spanish		Swedish		Polish	
MT System	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
Monolingual (Oracle)														
Mono-form	53.0	28.3	53.0	28.3	53.0	28.3	53.0	28.3	53.0	28.3	53.0	28.3	53.0	28.3
Mono-lem	52.1	27.5	52.1	27.5	52.1	27.5	52.1	27.5	52.1	27.5	52.1	27.5	52.1	27.5
Mono-stem	52.1	26.4	52.1	26.4	52.1	26.4	52.1	26.4	52.1	26.4	52.1	26.4	52.1	26.4
Query translation														
QT-SMT-form	47.2	22.6	48.0	23.6	44.2	21.7	45.9	22.9	46.9	23.2	40.0	20.2	42.1	20.1
QT-NMT-form	***57.2***	***26.0***	***51.5***	***24.1***	***50.3***	**22.5**	***50.7***	***24.0***	***49.0***	22.6	***50.1***	***23.8***	***47.2***	***22.3***
Document translation														
DT-SMT-form	39.0	17.4	42.1	21.5	40.4	**22.1**	40.0	17.2	45.6	**26.9**	38.3	17.0	40.7	20.4
DT-SMT-post-stem	36.9	16.7	44.5	22.7	39.2	***22.9***	35.4	17.0	46.3	***27.3***	33.9	16.7	35.3	18.7
DT-SMT-post-lem	39.3	18.3	41.9	21.7	37.7	**22.4**	37.1	17.0	42.7	25.0	33.0	16.0	37.1	**22.2**
DT-SMT-pre-lem	42.8	21.3	43.6	20.6	42.1	19.8	36.5	16.8	47.7	22.4	30.7	12.6	34.8	19.7
DT-NMT-form	42.1	15.6	46.0	19.8	36.6	14.0	26.0	10.5	43.9	17.5	33.9	11.6	38.9	12.3

Table 3: Extrinsic evaluation of the MT systems in the CLIR task. The CLIR experiments are evaluated using the Extended CLEF eHealth 2013–2015 test collection and compared with the results of monolingual retrieval (queries in English).

The QT experiments are done using the SMT and NMT systems, both translating into word forms (*QT-SMT-form* and *QT-NMT-form*). We want to stress here that the used MT systems for QT are different from the MT systems for DT, not only in the translation direction but also in the way that they were trained and tuned. Details are presented in Section 5.1 and Section 5.2.

In DT experiments, we exploit several configurations of the MT systems. *DT-SMT-form* translates the collection from English into the target language by the SMT system for document translation (no morphological processing applied). *DT-SMT-post-stem* refers to the results obtained by stemming the output of the the *DT-SMT-form* system. *DT-SMT-post-lem* lemmatizes the output of the *DT-SMT-form*, while *DT-SMT-pre-lem* lemmatizes the training data prior SMT training. (i.e., the translated documents in this system are already lemmatized). To compare the performance of DT when employing the NMT model, we report *DT-NMT-form*, which uses the presented NMT models to translate the collection into all the languages.

6.3 Result Analysis

In this work, we are mainly interested in comparing NMT vs. SMT employed in both the CLIR approaches (DT and QT), comparing the two approaches as such and analyzing the effect of morphological normalization in DT.

NMT versus SMT: For the QT approach, we can conclude that in terms of P@10, the NMT-based CLIR systems (using the *QT-NMT-form* MT systems) significantly outperform the SMT-based ones. Moreover, *QT-NMT-form* in Czech outperforms not only all other QT systems but also outperforms the monolingual system, which means that the NMT translations are on average better than the reference ones. This situation is illustrated in Table 4 which provides several examples of queries in which NMT not only provides translations which are better (in terms of P@10) than the ones provided by SMT but also better than the reference translations (for each translation, the P@10 score is in parentheses). This can be explained by the fact that the NMT models in our work are adapted to translate medical content by employing the collection itself in the back-translation process. This gives the model access to the collection vocabularies that are frequent in the retrieval collection, and in the relevant documents eventually.

To investigate this hypothesis. We train another *QT-NMT-form* system (for CS→EN only) using a different source of the back-translation data, namely the English side of the MT parallel text, which is also from the medical domain but different from the CLIR collection (the other settings of the system remain the same). The performance of this system decreased (as expected) from 57.2% to 54.2% (statistically significant). This shows that employing the document collection in back-

Query	SRC	REF	SMT	NMT
Query: 2013.38 (Czech)	*IM a dědičný*	*mi and hereditary (0.0)*	*mi and hereditary (0.0)*	*hereditary myocardial infarction (10.0)*
Query: 2014.19 (Swedish)	*L aneurysm i halspulsåder*	*l common carotid aneurysm (60.0)*	*l aneurysm in halspulsåder (0.0)*	*carotid artery aneurysm (100.0)*
Query: 2015.61 (French)	*hématomes sous les ongles*	*fingernail bruises (40.0)*	*bruising under the nail (10.0)*	*nail hematoma (60.0)*
Query: 2015.61 (Spanish)	*hematomas en la uña del dedo*	*fingernail bruises (40.0)*	*bruising in toe nail (20.0)*	*nail hematoma (60.0)*

Table 4: Comparison of query translations by two systems (*QT-SMT-form* and *QT-NMT-form*) and reference translations and their effect on retrieval quality. The figures in parentheses represent P@10 (in percentages) of retrieval when using the translation as a single query.

translation indeed helps produce translations that are more adapted to the collection domain.

NMT also helps deal with out-of-vocabulary (OOV) words (i.e., words do not appear in the training data), which is a common problem in SMT. For instance, the translations of Swedish queries produced by *QT-SMT-form* contain 40 untranslated terms. However, in *QT-NMT-form* translations, due to BPE, there are no OOVs at all (all words get translated, though the correct translation is not guaranteed). Very likely, this has a positive effect on the CLIR performance too.

QT versus DT: The most surprising observation in this work is the predominance of QT over DT in our experiments. In terms of P@10, for all the languages, *QT-SMT-form* provides significantly better translations than *DT-SMT-form*. For German and Spanish, the systems based on the translation of documents into morphologically normalized forms (lemmas, stems) perform on par with the systems based on *QT-SMT-form*, but for the other languages, the baseline *QT-SMT-form* is the best performing SMT option. The NMT models unsurprisingly boost translation quality for both QT and DT, but QT unexpectedly stays superior to DT, and the results get very close to the monolingual performance (and even higher for the Czech system, see above).

This can be explained by a simple hypothesis that a well-trained MT system based on the state-of-the-art techniques and sufficient amounts of training data is good enough to provide query translations of sufficient quality and does not require to see any larger context. The translation quality may not be perfect, but still sufficient for retrieval. For example, the Czech query *clef2015.test.33*, which is “*bílá infekce hltanu*“, is translated into English as “*white infection of pharynx*“. The reference translation for that query is “*white infection in pharynx*“. We can see that the CS→EN SMT system fails in translating prepositions (“*of*“ instead of “*in*“), but this does not affect the CLIR performance. However, we should keep in mind that our experiments are carried out in a very specific domain. This means that the queries are short, and often include symptoms and health conditions in which linguistics and contextual information may not play a significant role in solving the translation ambiguity
.

Morphological normalization: Producing document translations (lemmatized or stemmed) reduces collection vocabularies and improves term matching. However, in our experiments, none of the DT-SMT systems employing morphologically normalized translations of documents outperforms (in terms of P@10) the *QT-SMT-form* systems.

An example of a query where morphological normalization improved retrieval is the Czech query *clef2013.test.18*: “*aspirační pneumonie a dysfágie hltanu*“ (“*aspiration pneumonia and pharyngeal dysphagia*“ in English). The word “*hltanu*“, which means “*pharyngeal*“ is lemmatized in the training data of the SMT system and the Czech query into “*hltan*“, which means “*pharynx*“. When translating the English documents into Czech, “*pharynx*“ and “*pharyngeal*“ are translated back into “*hltan*“. This helps retrieve more relevant documents, increasing P@10 to 0.9 in *DT-SMT-pre-lem* from 0.7 in the monolingual systems (*Mono*, *Mono-lem* and *Mono-stem*), 0.6 in *QT-SMT-form* and 0.0 in *DT-SMT-form*. In comparison of pre-lemmatization and-post lemmatization, there is no clear winner. In the intrinsic MT evaluation, *DT-SMT-post-lem* outperforms *DT-SMT-pre-lem* for most languages.But in the extrinsic CLIR evaluation, *DT-SMT-pre-lem* is better for four languages and worse for Hungarian, Swedish, and Polish. *DT-SMT-pre-lem* in Spanish is the only DT system that outperforms the QT system. No clear conclusion can be done regarding the *DT-SMT-post-stem* models.

Finally, it is important to give insights about the cost-oriented comparison of the two approaches in terms of time complexity. The training time of our MT systems (both NMT and SMT) for both the approaches (QT and DT) is almost the same. The major difference was in the translation process. In the DT approach, translating the document collection using SMT took on average around three days using 200 CPU cores (each has 20 GB of RAM) for each language, which means it took us 21 days to translate 1.1 mil English documents into seven languages. While NMT translation was around ten times faster, using 20 GPUs only (GeForce RTX 2080Ti and Quadro P5000) with 10 GB of GPU RAM took around 20 days to translate the documents into the target languages. While for the QT approach, the translation process was pretty fast, where it took around 15 minutes to translate 66 queries from seven languages into English using SMT systems and around 3 minutes to do the same using NMT.

7 Conclusions

We presented a comparative study between query-translation (QT), and document translation (DT) approaches in the Cross-Lingual Information Retrieval (CLIR) task. To conduct this study, we investigated various MT systems and their configurations and performed a thorough large-scale evaluation based on the test collection produced within the CLEF eHealth tasks on patient-centered information retrieval during 2013–2015, and extended with additional relevance assessments.

We experimented with both statistical and neural MT paradigms. The SMT systems for QT were specifically trained and tuned to translate medical search queries. For DT, we trained two SMT systems: the first one was built to produce word forms, and the second one to produce word lemmas. We then used these two systems to translate the test collection into seven European languages. Furthermore, we performed lemmatization and stemming on the collection that was translated using the SMT system that produces word forms. The results showed that a well-tuned QT system outperforms DT, which is a positive result with an important impact on practical applications. So far, the QT approach has been preferred mainly for efficiency reasons (less space and computation needed). Our experiments suggest that this approach is even more effective (better retrieval results).

We also investigated the effect of using neural machine translation, which is now considered the state-of-the-art in many domains. This completely new paradigm in machine translation tends to improve the fluency of generated output (which is appreciated by humans), but often mismatches content and adequacy (which might hurt the performance in IR). In our experiments, NMT improved retrieval results in both QT and DT, but the QT approach is still superior, so the results are consistent with the findings from the SMT experiments.

However, we emphasize that the way we trained our MT systems is very domain-specific (medical domain), and we made use of a vast amount of medical data (monolingual and parallel). This makes our comparative study very task-oriented. When dealing with general domain test collection, some search terms might have a different meaning in different domains. For example, the word ”development” probably in most cases means in medicine the growth or spread of a disease (or a tumor), while in the general domain we can not say without a context, and in that case, the need for linguistics information in the queries will be more important to solve the translation ambiguity. This should be considered when comparing QT and DT approaches; thus, the reader should be careful when drawing the same conclusion of this work while working on a different domain.

Acknowledgments

This work was supported by the Czech Science Foundation (grant n. 19-26934X) and data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

We would like to thank everyone who helped us with this work, including Ondřej Bojar for his feedback on the design of the MT systems and Zain Baaity for this effort in the annotation of our system results.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. ACL.

Kareem Darwish and Douglas W. Oard. 2003. Probabilistic Structured Query Methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 338–344, New York, USA. ACM.

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

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 644–648, Atlanta, GA, USA. ACL.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. ACL.

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

Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen, and Guido Zuccon. 2013. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information Retrieval to Address Patients' Questions when Reading Clinical Reports. *CLEF 2013 Online Working Notes*, 8138:1–16.

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

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

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

David Hull. 1993. 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*, pages 329–338, Pittsburgh, USA.

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

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

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 Biomedical Translation Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark. ACL.

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

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

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

Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, and João Palotti. 2019. Overview of the CLEF eHealth Evaluation Lab 2019. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 322–339, Cham. Springer.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, and et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions*, pages 177–180, Stroudsburg, PA, USA.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised Cross-lingual Information Retrieval Using Monolingual Data Only. In *The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '18, pages 1253–1256, New York, NY, USA. ACM.

J. Scott McCarley. 1999. Should We Translate the Documents or the Queries in Cross-language Information Retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 208–214, College Park, Maryland. ACL.

Cristian Moral, Angélica de Antonio, Ricardo Imbert, and Jaime Ramírez. 2014. A Survey of Stemming Algorithms in Information Retrieval. *Information Research: An International Electronic Journal*, 19(1):n1.

Suraj Nair, Anton Ragni, Ondrej Klejch, Petra Galuščáková, and Douglas Oard. 2020. Experiments with Cross-Language Speech Retrieval for Lower-Resource Languages. In *Information Retrieval Technology*, pages 145–157, Cham. Springer.

Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of Statistical Machine Translation Model for Cross-lingual Information Retrieval in a Service Context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 109–119, Stroudsburg, PA, USA. ACL.

Douglas Oard. 1998. A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *Machine Translation and the Information Soup*, volume 1529, pages 472–483. Springer, Berlin, Germany.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. ACL.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. The GENIA Corpus: An Annotated Research Abstract Corpus in Molecular Biology Domain. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 82–86, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. 2005. Terrier Information Retrieval Platform. In *Advances in Information Retrieval*, pages 517–519, Berlin, Heidelberg. Springer.

Joao Palotti, Guido Zuccon, Pavel Pecina Jimmy, Mihai Lupu, Lorraine Goeuriot, Liadh Kelly, and Allan Hanbury. 2017. CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, pages 1–10, Dublin, Ireland. CEUR-WS.

João RM Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanbury, Gareth JF Jones, Mihai Lu pu, and Pavel Pecina. 2015. CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving information about medical symptoms. In *CLEF (Working Notes)*, pages 1–22, Berlin, Germany. Spriner.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, USA. ACL.

Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlavářová, Gareth J.F. Jones, and et al. 2014. Adaptation of Machine Translation for Multilingual Information Retrieval in the Medical Domain. *Artificial Intelligence in Medicine*, 61(3):165–185.

Ari Pirkola. 1998. The Effects of Query Structure and Dictionary Setups in Dictionary-based Cross-language Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 55–63, New York, USA. ACM.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Bruno Pouliquen and Christophe Mazenc. 2011. COPPA, CLIR and TAPTA: Three Tools to Assist in Overcoming the Patent Barrier at WIPO. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 24–30, Xiamen, China. Asia-Pacific Association for Machine Translation.

Andreas Rücklé, Krishnkant Swarnkar, and Iryna Gurevych. 2019. Improved Cross-Lingual Question Retrieval for Community Question Answering. In *WWW Conference*, WWW '19, pages 3179–3186, New York , USA. ACM.

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

Shadi Saleh and Pavel Pecina. 2016b. *Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval*, pages 54–66. Springer, Évora, Portugal.

Shadi Saleh and Pavel Pecina. 2019. An Extended CLEF eHealth Test Collection for Cross-lingual Information Retrieval in the Medical Domain. In *Advances in Information Retrieval - 41th European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings*, Lecture Notes in Computer Science. Springer.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, Berlin, Germany. ACL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany. ACL.

Mark D. Smucker and James Allan. 2005. An Investigation of Dirichlet Prior Smoothing's Performance Advantage. Technical report, University of Massachusetts.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. ACL.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, and et al. 2013. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. Springer, Berlin, Germany.

Jörg Tiedemann. 2009. News from OPUS: A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248, Borovets, Bulgaria. John Benjamins.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated DP Based Search for Statistical Translation. In *In European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aid an N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnet t, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Katharina Wäschle and Stefan Riezler. 2012. Analyzing Parallelism and Domain Similarities in the MAREC patent corpus. In *Multidisciplinary Information Retrieval*, volume 7356 of *Lecture Notes in Computer Science*, pages 12–27. Springer.