Evaluating Neural Text Simplification in the Medical Domain

van den Bercken, Laurens; Sips, Robert-Jan; Lofi, Christoph

DOI
10.1145/3308558.3313630

Publication date
2019

Document Version
Final published version

Published in
WWW'19 The World Wide Web Conference (WWW)

Evaluating Neural Text Simplification in the Medical Domain

Laurens van den Bercken
Delft University of Technology
myTomorrows
laurens.vandenbercken@
mytomorrows.com

Robert-Jan Sips
myTomorrows
robert-jan.sips@mytomorrows.com

Christoph Lofi
Delft University of Technology
c.lofi@tudelft.nl

ABSTRACT

Health literacy, i.e. the ability to read and understand medical text, is a relevant component of public health. Unfortunately, many medical texts are hard to grasp by the general population as they are targeted at highly-skilled professionals and use complex language and domain-specific terms. Here, automatic text simplification making text commonly understandable would be very beneficial. However, research and development into medical text simplification is hindered by the lack of openly available training and test corpora which contain complex medical sentences and their aligned simplified versions. In this paper, we introduce such a dataset to aid medical text simplification research. The dataset is created by filtering aligned health sentences using expert knowledge from an existing aligned corpus and a novel simple, language independent monolingual text alignment method. Furthermore, we use the dataset to train a state-of-the-art neural machine translation model, and compare it to a model trained on a general simplification dataset using an automatic evaluation, and an extensive human-expert evaluation.

CCS CONCEPTS

• **Information systems** → **Data extraction and integration**; • **Applied computing** → **Consumer health**; *Health care information systems*; • **Computing methodologies** → Supervised learning.

KEYWORDS

Medical Text Simplification, Test and Training Data Generation, Monolingual Neural Machine Translation

ACM Reference Format:
Laurens van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating Neural Text Simplification in the Medical Domain. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313630

1 INTRODUCTION

The rapid increase of health information on the internet has resulted in more patients turning to the Internet as their first source of health information. In a recent structural review, Tan and Goonawardene found that patients consult the internet primarily to be actively involved in the decision making related to their health [29]. While finding more evidence of positive impact on decision making and the patient-physician relation from online health information searches, Tan and Goonawardene warn that information found on the internet "has the potential to misguide patients and make them excessively anxious". Moreover, they observe that discrepancies between a physician's advice and a patient's conclusions from online information can erode the patient-physician relationship, which could limit timely access to care.

In a 2018 report [31] by the World Health Organization, contextualize these warnings, by concluding that a majority of European citizens has insufficient health literacy (the ability to read and understand healthcare information, make appropriate health decisions and follow health instructions). The report stipulates that one of the major causes of this low health literacy is that health information is often inaccessible to the general public because the literacy demands of health information and the literacy skills of average adults are mismatched, the information is often poorly written, poorly designed and/or geared to an highly sophisticated audience. One way to improve health literacy is by simplifying (online) medical text, to match the literacy level and vocabulary of average adults.

To illustrate this, we provide an example from Wikipedia, and its simplified version from Simple Wikipedia below.

- Pituitary adenomas represent from 10% to 25% of all intracranial neoplasms and the estimated prevalence rate in the general population is approximately 17%.
- Pituitary adenomas represent from 10% to 25% of all brain tumors and is thought to happen in about 17% to 25% of most people.

Observe that the Wikipedia editors simplify text on two levels: (1) complicated medical terminology (*intracranial neoplasms*) is replaced by simpler terms and (2) complicated non-medical sentence structures (*estimated prevalence rate in the general population*) are simplified. There is an rapidly increasing body of online medical texts, such as electronic health records, clinical trials, medical research, drug labels and patient information leaflets. This rapid increase makes manual simplification unfeasible.

Therefore, in this paper, we focus on the following research question:

> **RQ**: To what extent can we use automated methods to simplify expert level health text to laymen level?

Current work in automated medical text simplification is mostly limited to simplifying medical terminology, either by the generation of explanations (explanation generation), or by replacing these terms with laymen terms or definitions (lexical simplification) [1, 5–8, 24, 25]. This ignores complex non-medical terms and

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313630

complicated sentence structures, which also hamper readability [33]. The state-of-the-art in automated text simplification, Neural Machine Translation [22, 28], shows promise to solve this second problem, but requires large parallel corpora for training, which are lacking in the medical domain. Recent work by Adduru et al. focused on the creation of such a medical text simplification corpus [2]. Unfortunately, the resulting set is not publicly available.

Original Contribution As there are no publicly available medical text simplification corpora, we create a new aligned corpus by semi-automatically filtering a set of aligned health-related sentences from an existing parallel text simplification corpus. In addition, we introduce a language independent monolingual text alignment method and use it for aligning additional health sentences from Wikipedia and Simple Wikipedia disease articles. The resulting data set is made publicly available for future research[1]. Furthermore, we propose a method using state-of-the-art Neural Machine Translation for medical text simplification, which can both simplify general text but also learns to translate medical concepts into simpler terms. We perform experiments with that method on the filtered health sentences. We report results of quantitative evaluations, and a qualitative one.

2 RELATED WORK

In this section we discuss relevant work on text simplification in the medical domain. We first discuss lexical simplification, text simplification which focuses on the replacement of complex terms. Secondly, we discuss syntactic simplification: simplification which focuses on the replacement of complicated sentence structures and conclude with combined approaches: text simplification which targets both the replacement of complicated terms and the replacement of complicated structures.

Lexical simplification Most work on medical lexical simplification is focused on the usage of large vocabularies, most prominently the Unified Medical Language System (UMLS) [4] to replace expert medical terms with consumer oriented synonyms. The UMLS is a meta-thesaurus, which contains unified entities from a large number of medical vocabularies (such as SnomedCT [9], MeSH [18] and CHV [37]). In summary, the state-of-the-art in lexical simplification is recognizing UMLS concepts from text and replacing them with a consumer-oriented synonym from the Consumer Health Vocabulary [37], an open access collection of consumer oriented synonyms for medical concepts. Despite significant efforts to automatically enrich and correct the Consumer Health Vocabulary from user-generated data [11, 13, 30, 36], evaluation of the effect of using these terms on the perceived simplicity of medical text by the lay population is lacking and recent work by Xie et al. articulates that medical concept replacement alone is not sufficient for medical text simplification [33].

Syntactic simplification There is little work investigating medical syntactic simplification in isolation. Leroy et al. investigate the (manual) splitting of complex noun phrases to improve readability of long sentences. However, they conclude that this approach does not necessarily improve readability and recommend that sentences should only be split when the split phrases "feel more natural" [17]. Furthermore, negations in medical texts were investigated and it was shown that easier text contains less morphological negations than difficult text [21]. An easy text contains for example "not clear" instead of "unclear", which could effectively be solved by a lexical simplification tool based on frequency analysis.

Combined approaches Monolingual machine translation, i.e. machine translation algorithms trained on a parallel corpus in the same language, have shown great promise in recent years. Such systems learn how to translate complex language into simple language, when trained of a parallel corpus of complex and simple sentences. In theory, such a translation combines lexical and syntactic simplification. Most prominent is the progress in Neural Machine Translation [22], which has demonstrated to achieve state-of-the-art performance on text simplification tasks for common language. Neural Machine Translation relies on the availability of a large parallel corpus for training and evaluation purposes. For common language, publicly available corpora are available from aligned Wikipedia and Simple Wikipedia sentences, the Parallel Wikipedia Simplification (PWKP) corpus [38] and a more recent corpus presented Hwang et al. [12] and from news articles, the Newsela corpus [34].

Algorithms trained on these datasets perform well on general language simplification, but have been shown to perform poorly on medical text simplification [2, 16]. For instance, an (statistical) algorithm trained on the PWKP dataset for simplifying cancer and other health text produced output that was "imperfect and required a second manual step to be consistent and accurate" [16].

To successfully employ Neural Machine Translation on health text, we would need a health specific parallel corpus. Unfortunately such a corpus is not available and, a first attempt by Adduru et al. to creating one [2] showed that this is not as easy as it seems. Adduru et al. used an array of methods to automatically align sentences from the medical subset of Wikipedia and Simple Wikipedia, as well as https://www.webmd.com and https://www.medicinenet.com. The result is a -proprietary- medical text simplification corpus of 2,493 sentence pairs. Adduru et al. present an automated evaluation of a Neural Machine Translation algorithm on these data, but do not include an human evaluation.

3 DATA

In this section we present two datasets we created for text simplification in the medical domain. The first dataset (EXPERT) is an expert-evaluated medical subset filtered from the aligned wikipedia corpus presented by Hwang et al. [12]. It is focusing on reliable high-quality sentence alignments such that it can be used as a test set for benchmarking. The second dataset (AUTOMATED) is a novel dataset created by automatic collection of aligned sentences from the medical subset of Wikipedia. Here, the focus lies on having a large dataset which can serve as training data, but we accept smaller losses in quality resulting from the automatic alignment. In table 1, we provide a summary of the presented datasets, compared to the medical corpus presented by Adduru et al. [2].

3.1 EXPERT dataset

Our EXPERT dataset is created using the aligned corpus presented in [12] as a baseline, which aligns sentences between Wikipedia and SimpleWikipedia. As the corpus does not focus on a particular

[1] available at https://research.mytomorrows.com/

	Fully aligned	Partially aligned
EXPERT	2,267	3,148
AUTOMATED	3,797	-
Adduru et al. [2]	2,493	-

Table 1: An overview of the datasets

domain, only few medical sentences are covered which motivates the creation of our EXPERT dataset. This initial corpus consists of manually and automatically generated *good* and *good partial* aligned sentence pairs, the former defined as "the semantics of the simple and standard sentence completely match, possibly with small omissions (e.g., pronouns, dates, or numbers)" and the latter as "a sentence completely covers the other sentence, but contains an additional clause or phrase that has information which is not contained within the other sentence". In the remainder of the paper we will refer to the *good* sentence pairs as *fully aligned* and to the *good partial* as *partially aligned* sentence pairs.

To generate the EXPERT dataset, we use a state-of-the-art medical named entity recognition tool, QuickUMLS [26] to sentences which may contain a medical topic from the fully aligned and partially aligned datasets. QuickUMLS is an approximate dictionary matching algorithm which matches terms from text with synonyms in the UMLS. We used QuickUMLS with the default setting for similarity threshold (0.7) and limited the semantic types to *Disease or Syndrome* and *Clinical Drug*). We consider a sentence pair a candidate medical sentence pair, when QuickUMLS recognizes at least one medical concept in either the complex or the simple medical sentence. After QuickUMLS processing, we provided the resulting candidate medical sentence pairs to a domain expert for additional validation, i.e. to confirm whether the sentence pair is indeed health-related. Using this approach, we created a filtered corpus of 2,267 *fully aligned* medical sentences and 3,148 *partially aligned* sentences.

3.2 AUTOMATED dataset

In addition to the labour-intensive manual filtering, we created a pipeline to automatically create an aligned dataset from Wikipedia and Simple Wikipedia, which allows efficient creation of a much larger dataset, at the cost of a slight loss in quality. In principle, such a pipeline has 3 distinct steps: (1) Collection of relevant articles and their related simplified version, (2) Splitting the articles in sentences and (3) Aligning the sentences into pairs.

3.2.1 Finding relevant articles. Recent work, by Kajiwara and Komachi [14] and Adduru et al. [2] focused on the creation of an aligned corpus from Wikipedia and Simple Wikipedia. The former presented a methodology to create a general corpus, the latter a medical corpus. Kajiwara and Komachi used a full dump of Wikipedia and Simple Wikipedia and aligned the articles with matching titles. Given the goal of creating a general-purpose corpus, they did not attempt to select articles based on topic. In their work, they identify a total of 126,725 Wikipedia articles with a matching Simple Wikipedia article in the English language. In contrast, Adduru et al. present an approach to collect a specific subset of medical Wikipedia articles. They manually selected a set of 164 articles, which they match to Simple Wikipedia articles with a matching title. Manual collection of such a dataset seems unnecessarily cumbersome. We propose an approach using DBPedia [3] and select all articles that fall in the *dbo: Disease* class. After title matching to Simple Wikipedia, this gives us a set of 1,098 aligned articles.

3.2.2 Splitting. Analogous to Kajiwara and Komachi, we extract the text from the Wikipedia and Simple Wikipedia articles, using the python Wikipedia API [2] and the NLTK 3.3 sentence tokenizer [3]. This gives an average number of words per sentence of 26.1 for the normal articles and 19.5 for the simple articles. The average numbers of sentences per article are 123.4 and 20.3 respectively. In comparison, Kajiwara and Komachi report an average number of words per sentence of 25.1 for the normal articles and 16.9 for the simple articles and an average number of sentences per article of 57.7 and 7.65, respectively. Medical articles (simple and normal), seems to be longer and more complex (they contain more and longer sentences).

3.2.3 Aligning. To align sentences from Wikipedia to Simple Wikipedia, we employ a two step approach: first we generate *candidate pairs*, by combining each sentence from the Wikipedia articles which each sentence of the related Simple Wikipedia article. This gives us a total of 3,660,064 candidate pairs from the 1,098 articles. Adduru et al. report 818,520 candidate pairs from 164 articles, demonstrating that their manually collected set contains longer articles than ours (3333.4 candidate pairs per article in our set versus 4991 candidate pairs per article in their set). Secondly, we select the most similar pairs from the candidate pairs. Various methods have been reported to perform this task, Kajiwara and Komachi employ pre-trained Word2Vec word embeddings to determine sentence similarity. Similarly, Hwang et al. present a method that relies on Wiktionary [12]. When aligning sentences where the distinctive (medical) terms are arguably very infrequent, such dependencies may not be wanted, as also noted by Adduru et al. who use a classifier to identify matching sentences.

BLEU alignment. We propose a simple metric, the BLEU score [23] to align the AUTOMATED dataset. The BLEU score is used commonly to evaluate Machine Translation algorithms, by comparing the similarity between the output of a translation algorithm to references sentence. In short, BLEU does this by counting overlapping word n-grams. For the sake of brevity, we refer to Papineni et al.[23] for details on the method. To the best of our knowledge, we're the first to employ BLEU for a sentence alignment task. To evaluate the quality of the BLEU alignment for the sentence alignment task, we compare BLEU alignment to the Maximum alignment reported by Kajiwara and Komachi[14], using the manual alignment set from Hwang et al. [12] for evaluation. This evaluation set contains 67,853 candidate sentence pairs, judged by human annotators. 277 were considered fully aligned, 281 partially aligned and 67,295 considered either not good enough partial alignments or bad alignments.

We test both methods in two sentence alignment scenarios: (1) Full alignment: Fully aligned sentences versus the rest and (2) Partial alignment: Fully and partially aligned sentences versus the rest.

Table 2 reports maximum F1-score and AUC for both methods in both scenarios. We observe that BLEU alignment performs on

[2] https://pypi.org/project/wikipedia/
[3] http://www.nltk.org

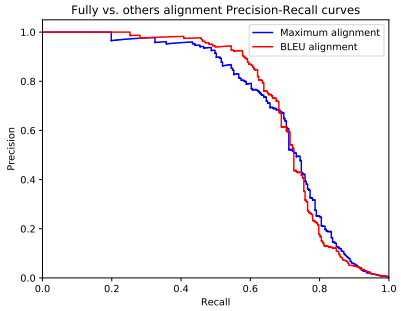

Figure 1: Precision-recall curves of Maximum alignment and BLEU alignment

Fully vs. rest	Max F1	AUC
BLEU alignment	0.717	0.714
Maximum alignment in [14]	0.717	0.730
Alignment used to align Wikipedia [12]	0.712	0.694
Fully and partially vs. rest		
BLEU alignment	0.534	0.484
Maximum alignment in [14]	0.638	0.618
Alignment used to align Wikipedia [12]	0.607	0.529

Table 2: Max F1 and AUC scores for identifying fully aligned and fully and partially aligned sentences

par with Maximum alignment for fully aligned sentences, but performance is worse on partially aligned sentences. In figure 1 we report the precision-recall curve for the fully aligned scenario.

We observe that BLUE alignment provides a useful method when performing sentence alignment on highly domain specific text. Despite the poorer performance on partial alignment, it does not depend on pretrained embeddings or external datasources to function. In addition, when aligning medical data, the vocabulary might contain a lot of words that are neither in the vocabulary of pretrained Word2Vec models nor in Wiktionary, which may deteriorate performance of approaches that use such sources. BLEU alignment only looks at overlapping n-grams, which makes it language independent.

In the AUTOMATED dataset, we only include fully aligned sentence pairs, given the poor performance BLEU alignment demonstrated on partial alignment. In presented dataset, we include sentence pairs with a BLEU score >0.29, which led to a maximum F1 score during evaluation. After filtering out sentences with MediaWiki mathematical formulas, the final set consists of 3,797 fully aligned medical sentences. In table 3 we present example aligned sentences from this set.

4 NEURAL TEXT SIMPLIFICATION

Most current research on text simplification in the medical domain focuses on simplifying medical concepts only. However, monolingual NMT has shown great potential in text simplification research but has not been applied to the medical domain yet. Therefore, we replicate the state-of-the-art NMT text simplification system of [22] as a baseline, and evaluate it on our expert-curated

Wikipedia	Simple Wikipedia	BLEU
Spinal tumors are neoplasms located in the spinal cord.	Spinal tumors is a form of tumor that grows in the spinal cord.	0.39
Aspirin is an appropriate immediate treatment for a suspected MI.	Aspirin is an early and important treatment for a heart attack.	0.33

Table 3: Example alignments using BLEU alignment

dataset. This system outperformed phrase-based [32] and syntax-based statistical machine translation [35] approaches, as well as an unsupervised lexical simplification approach [10]. Furthermore, we design a second NMT model which uses a combination of our AUTOMATED and other datasets, and replacing medical concepts with identifiers.

4.1 Training and Evaluation sets

For the setup of our experiments, we rely on the general dataset presented by Hwang et al.[12] and combine this with the EXPERT and AUTOMATED datasets described in the previous section. This gives us 4 datasets:

- **Fully aligned health sentences** f_{health} - Filtered and expert evaluated fully aligned health sentences from f_{wiki}: 2,267 sentences.
- **Partially aligned health sentences** p_{health} - Filtered and expert evaluated partially aligned health sentences from p_{wiki}: 3,148 sentences.
- **Fully aligned general domain sentences** $f_{\text{general}} = f_{\text{wiki}} - f_{\text{health}}$: 152.538 sentences.
- **Partially aligned general domain sentences** $p_{\text{general}} = p_{\text{wiki}} - p_{\text{health}}$: 126.785 sentences.

4.2 Baseline

We implemented the baseline system in OpenNMT[4], an open source framework for NMT. The architecture consists of two LSTM layers, states of size 500 and 500 hidden units and a 0.3 dropout probability. The vocabulary size is pruned to 50,000 in both the source and target language. Word embedding size is set to 500. We used pretrained Word2Vec embeddings from the Google News corpus [20] of size 300. The remaining part is learned during training of the NMT (while the pre-trained part remains fixed). Lastly, the decoder uses global attention with input feeding [19]. The system is trained for 15 epochs, using a SGD optimizer and an initial learning rate of 1.0. After epoch 9, the learning rate decay is 0.7, i.e. learning_rate = learning_rate * learning_rate_decay.

At translation time beam search is used to find the best prediction given the input. Beam search is an approximation of the best possible translation. At each step of the translation the k most likely words are generated given the input sentence. Here, k is called the beam size. Then, the most likely sequence (i.e. translation) is called hypothesis 1, the next hypothesis 2, etc. We will evaluate both hypothesis setting in the next section. The system that performed most changes and highest percentage of correct ones in [22] used a

[4]http://opennmt.net/

beam size of 12. This system is trained on general domain corpus, i.e. $f_{\text{general}} + p_{\text{general}}$.

4.3 Medical+CHV Replacement

Our second NMT system (MED-CHV) we evaluate in this paper follows a similar architecture as the baseline, but is trained on a combination of the general corpus and our health-related corpora (minus the corpus which is used as a test set): $f_{\text{general}} + p_{\text{general}} + f_{\text{health}} + p_{\text{health}} + f_{\text{BLEU-health}}$

In addition, we replace each medical concept encountered in the complex text with a Concept Unique Identifier (CUI) from UMLS. This approach reduces the (medical) vocabulary (and medical concept sparsity), since any textual variation of a concept is mapped (or normalized) to a single CUI, aggregating the references for each concept. For example, *atherosclerotic heart disease* and *coronary heart disease*, which are synonyms, are both replaced with C0010054. Furthermore, it enables us the replace CUIs with their CHV-preferred term if the CUI is not part of the source vocabulary of the NMT (i.e. rare medical concepts / CUIs). We used QuickUMLS [26] with a similarity threshold of 0.7, a value for which highest F1-scores were achieved in [26], to detect medical concepts and link them to a CUI. For the decoder we use pre-trained Word2Vec embeddings of size 200, trained on 10,876,004 English abstracts of biomedical articles from PubMed [15].

Note that we include the 50,000 most frequent words in the source and target vocabulary (so we have enough reference translations for each word in the vocabulary). This may cause that some CUIs are not in the source vocabulary and are therefore not translated. To overcome this, we replace CUIs that are out of vocabulary with their CHV-preferred term, if it exists, or copy the original source token. To do this, we make use of a phrase table, which can be pre-constructed before translation. Each entry in the phrase-table contains a CUI with its CHV-preferred term or its original source token. Instead of substituting out of vocabulary words with source words that have the highest attention weight, a possible translation in the phrase-table is looked up. This way the output does not contain any raw CUI.

5 EVALUATION

In this evaluation, we focus on the question of how well does NMT-based text simplification work in the medical domain. For our first experiment, we rely on an automated evaluation approach based on a reference test-set drawn from our EXPERT dataset. The second evaluation relies on human evaluators, and focuses on simplicity, understandability, and correctness of simplified sentences.

We randomly select 500 and 350 sentences as validation set and test set respectively from f_{health}. Automatic evaluation is done on the test set of size 350. Human evaluation is done in the first 70 sentences of the test set (since human evaluation is rather costly).

Automatic Evaluation: Text simplification is typically automatically evaluated using a traditional machine translation metric BLEU [23] and a text simplification specific metric SARI [35].

BLEU compares the output against references and produces a score between 0 and 1, with 1 representing a perfect translation (i.e. identical to one of the references). In our evaluation we use word n-grams up to 4. However, when used for simplification, it has to be handled with care as it is not uncommon that the source sentences (from Wikipedia) and the reference sentences (from Simple Wikipedia) are identical or very similar as Wikipedia editors just copied them over without or only with minor modifications. Therefore, a machine simplification which just keeps the source sentence as-is often has high BLEU scores, but is not simpler.

Hence, a specific text simplification metric was introduced in [35], called SARI, which compares **S**ystem output **A**gainst **R**eferences and against the **I**nput sentence. It focuses on lexical simplification, i.e. replacing complex words and phrases with simpler alternatives. "It explicitly measures the goodness of words that are *added*, *deleted* and *kept* by the systems" [35], by comparing the output with the source and the reference or multiple. SARI combines several aspects of adding and deleting words into a single numeric measure: the terms added by the simplification algorithm with respect to if they are also added in the reference simplification; and the terms removed by the simplification algorithm also with respect to if they are removed in the reference, and the terms which are kept stable between the reference and a simplification.

For this experiment, we evaluate the baseline system and the MED-CHV system, both with hypothesis 1 and 2 selection strategies (i.e., choose the most likely simplification and the second most likely one.) Furthermore, we consider an "Identity" simplification which just copies the source sentences without modifying.

Human Evaluation: As both metrics used in the automatic evaluation are insufficient to fully describe the capabilities of machine simplification, such evaluation need to be accompanied by a human evaluation. To this end, we obtain feedback on simplified sentences focusing on grammar, meaning preservation (both measured on a 1-5 Likert scale), and simplicity (on a scale of -2 to 2, with negative values representing that the text has become more complex). This follows the setup outlined in [22]. An evaluator is presented with a sentence pair (complex, simple) and asked to give the scores. We base our annotation guidelines on [27]. We slightly edited the guidelines, since their focus was on splitting (and deleting parts of) sentences, while our system mainly replaces words and deletes parts of sentences.

6 RESULTS AND DISCUSSION

In this section we report results of automatic and human evaluation.

Automatic Evaluation: In table 5, SARI, along with its three components, and BLEU scores are reported. The scores represent if the system is actually modifying the text, and how it relates to the test set reference sentences. "Identity" does not perform any text simplification, but simply uses the source sentence. This tells us how similar the source is to the reference. It serves as calibration scores for SARI and BLEU; e.g., not simplifying anything results in a BLEU score of 0.53 and a SARI score of 21.56. Both hypothesis 1 and 2 of the baseline (i.e. choosing the most likely or second likely simplification) are able to improve SARI scores. The main difference between them is that hypothesis 2 deletes with higher precision than hypothesis 1. Both hypotheses of the MED-CHV show comparable numbers for keeping and deleting terms, but a slightly higher number for adding terms. This may be because of the additional terms (medical concepts) the medical NMT is translating. BLEU scores of the identity and the baseline's

Source	Sentence
Wikipedia	Coronary artery disease (CAD) also known as atherosclerotic heart disease , coronary heart disease , or ischemic heart disease (IHD) , is the most common type of heart disease and cause of heart attacks .
Simple Wikipedia	Atherosclerosis is a form of heart disease .
Baseline, h-1	Coronary artery is the most common type of heart disease .
Baseline, h-2	Coronary artery is a type of disease .
Medical input	C1956346 (CAD) also known as C0010054 , C0010054 , or C0151744 (C0151744) , is the most common type of C0018799 and cause of C0027051 .
MED-CHV, h-1	$\{\text{coronary artery disease}\}_{\text{copied}}$ is the most common type of $\{\text{heart disease}\}_{\text{NMT}}$.
MED-CHV, h-2	$\{\text{coronary artery disease}\}_{\text{copied}}$ is the most common type of $\{\text{heart disease}\}_{\text{NMT}}$ and cause of $\{\text{heart attack}\}_{\text{NMT}}$.

Table 4: Example translations from different systems, medical concepts in MED-CHV are replaced with their CUI.

Approach	SARI	F_{add}	F_{keep}	P_{del}	BLEU
Identity	21.56	0.00	64.68	0.00	53.07
Baseline, h-1	28.14	1.91	60.37	22.15	54.78
Baseline, h-2	32.73	2.03	55.82	40.34	44.51
MED-CHV, h-1	32.27	2.24	57.10	37.47	47.48
MED-CHV, h-2	33.92	2.96	54.93	43.88	44.37

Table 5: Evaluations with automatic metrics

hypothesis 1 are highest. This may be due to that in hypothesis 1 the baseline is often producing the exact same sentence. The others are less conservative, i.e. perform more changes, which reduces BLEU. We showed that the NMT systems indeed improve SARI scores and therefore we expect that the output is simpler than the input. The medical NMT slightly increased SARI over the baseline (due to its F_{add} component). Therefore, we expect that simplicity scores will be at least similar to the baseline.

Manual Evaluation: Three laymen provided feedback on the first 70 sentences of the test set with respect to grammar, meaning preservation, and simplicity.

Table 6 shows that the baseline produces decent grammar and meaning preservation scores and indeed simplifies the text. However, MED-CHV scores show that grammar, meaning preservation and simplicity scores are all lower than the baseline. We assume that this is due to MED-CHV replacing out of vocabulary concepts with their CHV-preferred terms (which are expert curated simplified terms) instead of substituting them with source words that have the highest attention weight. While we assumed that using these expert term simplifications should perform well, also previous research concluded that "some CHV-preferred terms can be above the level of consumers' comprehension" [24].

Example translations are given in table 4. Note that the input of medical NMT is the Wikipedia sentence with medical concepts replaced with their CUI. Common medical concepts, such as heart disease (C0018799) and heart attack (C0027051), are part of the vocabulary and correctly translated by the NMT. Coronary artery disease (C1956346) is neither part of the vocabulary, nor a CHV-preferred term exists for it. Therefore, the source term is copied.

Approach	G	M	S
Simple Wikipedia	4.91	4.24	0.53
Baseline, h-1	4.85	4.30	0.22
Baseline, h-2	4.49	3.87	0.23
MED-CHV, h-1	4.23	3.82	-0.05
MED-CHV, h-2	4.19	3.76	-0.05

Table 6: Human evaluation scores. G:Grammar, M:Meaning preservation, S:Simplicity

7 CONCLUSION AND FUTURE WORK

Automated Medical Text Simplification can be a cornerstone technology to address insufficient health literacy. However, research into this domain is hampered by the lack of open training and test corpora. Therefore, in this paper we introduced such an open corpus which is based on the widely available Wikipedia-Simple Wikipedia text simplification corpus, and expanded with additional aligned sentences focusing on the medical domain. This corpus was created based on filtering with a medical expert from an existing aligned dataset, and by a novel simple, language independent monolingual text alignment method.

We used this corpus to evaluate two Neural Machine Translation models: one was trained on the aligned Wikipedia corpus (baseline), the other one was in addition trained on our corpus, but with medical terms replaced by their UMLS Concept Unique Identifiers. We assumed that the replacement would further boost performance. Both models were evaluated automatically and manually focusing only on the medical subset of the test data set we created. During automatic evaluation, it could be shown that the baseline performs fewer changes to sentences when simplifying. However, in the manual human-driven evaluation, it became clear that changing too many parts of the sentence can be detrimental, and that the baseline sentences were judged to be more understandable and simpler. We assume that this can be attributed to the act of replacing out of vocabulary medical concepts with their CHV-preferred terms. We therefore assume that training only with our extended dataset without additional replacements should yield superior performance. Due to the extreme costs of manually evaluating simplification results, this experiment will be covered in our future work. While this result was disappointing, it shows that automatic text simplification is a difficult task which demands future research.

In summary, we contributed a novel and open test and training dataset of aligned sentences focused on medical text simplifications, which easily allows such future research. Furthermore, we could show that even training a Neural Machine Translation model on a non-specialized corpus can still yield acceptable results in a complex domain like medical texts, clearly hinting at the potential of future endeavours.

REFERENCES

[1] Emil Abrahamsson, Timothy Forni, Maria Skeppstedt, and Maria Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. 57–65.

[2] Viraj Adduru, Sadid Hasan, Joey Liu, Yuan Ling, Vivek Datla, and Kathy Lee. 2018. Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification. In *The 3rd International Workshop on Knowledge Discovery in Healthcare Data*.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[4] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[5] Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, and Hong Yu. 2018. A Natural Language Processing System That Links Medical Terms in Electronic Health Record Notes to Lay Definitions: System Development Using Physician Reviews. *Journal of Medical Internet Research* 20, 1 (2018), e26.

[6] Jinying Chen, Abhyuday N Jagannatha, Samah J Fodeh, and Hong Yu. 2017. Ranking medical terms to support expansion of lay language resources for patient comprehension of electronic health record notes: adapted distant supervision approach. *JMIR medical informatics* 5, 4 (2017).

[7] Jinying Chen and Hong Yu. 2017. Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients. *Journal of biomedical informatics* 68 (2017), 121–131.

[8] Jinying Chen, Jiaping Zheng, and Hong Yu. 2016. Finding important terms for patients in their electronic health records: a learning-to-rank approach using expert annotations. *JMIR medical informatics* 4, 4 (2016).

[9] Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 121 (2006), 279.

[10] Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: do we need simplified corpora?. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2. 63–68.

[11] Zhe He, Zhiwei Chen, Sanghee Oh, Jinghui Hou, and Jiang Bian. 2017. Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach. *Journal of biomedical informatics* 69 (2017), 75–85.

[12] William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 211–217.

[13] Ling Jiang and Christopher C Yang. 2015. Expanding consumer health vocabularies by learning consumer health expressions from online health social media. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 314–320.

[14] Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 1147–1158.

[15] Aris Kosmopoulos, Ion Androutsopoulos, and Georgios Paliouras. 2015. Biomedical semantic indexing using dense word vectors in bioasq. *J BioMed Semant Suppl BioMedl Inf Retr* 3410 (2015), 959136040–1510456246.

[16] Poorna Kushalnagar, Scott Smith, Melinda Hopper, Claire Ryan, Micah Rinkevich, and Raja Kushalnagar. 2018. Making cancer health text on the Internet easier to read for deaf people who use American Sign Language. *Journal of Cancer Education* 33, 1 (2018), 134–140.

[17] Gondy Leroy, David Kauchak, and Alan Hogue. 2016. Effects on text simplification: Evaluation of splitting up noun phrases. *Journal of health communication* 21, sup1 (2016), 18–26.

[18] Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88, 3 (2000), 265.

[19] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[21] Partha Mukherjee, Gondy Leroy, David Kauchak, Srinidhi Rajanarayanan, Damian Y Romero Diaz, Nicole P Yuan, T Gail Pritchard, and Sonia Colina. 2017. NegAIT: A new parser for medical text simplification using morphological, sentential and double negation. *Journal of biomedical informatics* 69 (2017), 55–62.

[22] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 85–91.

[23] 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*. Association for Computational Linguistics, 311–318.

[24] Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. 2017. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *Journal of medical Internet research* 19, 12 (2017).

[25] Isabel Segura-Bedmar and Paloma Martínez. 2017. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of biomedical semantics* 8, 1 (2017), 45.

[26] Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*.

[27] Sanja Štajner and Goran Glavaš. 2017. Leveraging event-based semantics for automated text simplification. *Expert systems with applications* 82 (2017), 383–395.

[28] Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and Effective Text Simplification Using Semantic and Neural Methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 162–173.

[29] Sharon Swee-Lin Tan and Nadee Goonawardene. 2017. Internet health information seeking and the patient-physician relationship: a systematic review. *Journal of medical Internet research* 19, 1 (2017).

[30] VG Vinod Vydiswaran, Qiaozhu Mei, David A Hanauer, and Kai Zheng. 2014. Mining consumer health vocabulary from community-generated text. In *AMIA Annual Symposium Proceedings*, Vol. 2014. American Medical Informatics Association, 1150.

[31] World Health Organization (WHO and others. 2018. Health literacy. The solid facts. *Self* (2018).

[32] Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 1015–1024.

[33] Deborah X Xie, Ray Y Wang, and Sivakumar Chinnadurai. 2018. Readability of online patient education materials for velopharyngeal insufficiency. *International journal of pediatric otorhinolaryngology* 104 (2018), 113–119.

[34] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics* 3, 1 (2015), 283–297.

[35] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics* 4 (2016), 401–415.

[36] Ming Yang and Melody Kiang. 2015. Extracting Consumer Health Expressions of Drug Safety from Web Forum. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 2896–2905.

[37] Qing T Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association* 13, 1 (2006), 24–29.

[38] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 1353–1361.