Evaluating Neural Text Simplification in the Medical Domain

Laurens van den Bercken

Master Thesis
Computer Science
Software Technology

Web Information Systems

TU Delft

myTomorrows

Evaluating Neural Text Simplification in the Medical Domain

by

Laurens van den Bercken

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on February 28, 2019.

Student number:	4587545	
Project duration:	May 23, 2018 – February 28, 2019	
Thesis committee:	Dr. C. Lofi,	TU Delft, supervisor
	R. Sips MSc,	myTomorrows, company supervisor
	Prof. G.J.P.M. Houben,	TU Delft, chair
	Dr. J.S. Rellermeyer,	TU Delft

An electronic version of this thesis is available at `https://repository.tudelft.nl/`.

TU Delft myTomorrows

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 health professionals and use complex language and domain-specific terms. Here, automatic text simplification making text commonly understandable would be very beneficial. In this thesis we evaluate the state-of-the-art in automatic text simplification in the medical domain. We train a Neural Machine Translation (NMT) system on aligned complex and simple sentences from Wikipedia and Simple Wikipedia. As there are no publicly available aligned medical text simplification corpora, we create one semi-automatically with the help of a domain expert and one fully automatically using a novel monolingual alignment method introduced in this thesis. We analyse the effect of in-domain data when training an NMT system. Furthermore, we describe two strategies for medical term simplification in combination with NMT: 1) An extra pre-processing step that boosts medical term simplification 2) A post-processing dictionary approach using the Open-Access and Collaborative Consumer Health Vocabulary (CHV). We analyse the effect of both strategies separately. We let humans evaluate the output on grammar, meaning preservation (from the complex sentence) and simplicity (compared to the complex sentence).

Results show that an NMT trained on general aligned complex and simple sentences is able to simplify medical sentences at the level of Simple Wikipedia. An NMT trained on medical sentences (in addition to general sentences) in combination with the boosting strategy for medical term simplification is able to translate more medical concepts, but the output is not simpler than the NMT trained on general sentences only. Interestingly, NMT in combination with the CHV did not boost simplicity, but had the opposite effect.

Preface

This thesis is written to obtain the degree of Master of Science at the Delft University of Technology. We conducted the research at myTomorrows[1] located in Amsterdam. myTomorrows is a company that "provide[s] patients with unmet medical needs, and their doctors, information about treatment options worldwide and facilitate access to medicines in development".

Firstly, I would like to thank my Delft University of Technology supervisor Christoph and myTomorrows supervisor Robert-Jan for their guidance and sincere involvement throughout the project. Together, we succeeded in submitting a scientific paper about (the first part of) this work to The Web Conference 2019[2]. On January 21, 2019 we were informed that our submission was accepted as a short paper. The paper can be found in appendix A.

I would like to thank Gerhard Mulder, Medical Data Research Coordinator at myTomorrows, who provided in-domain knowledge needed for filtering out medical sentences from aligned Wikipedia and Simple Wikipedia. Also, thanks to myTomorrows colleagues Annette Valentijn, Friso Reitsma and Koen Witteman for manually evaluating system outputs. Thanks to all other colleagues from myTomorrows who where always in for a chat.

Lastly, I would like to thank my parents for their great support throughout my whole education period and their financial support. Also thanks to my brother and friends for keeping me motivated and helping me get through this. Special thanks go to my girlfriend for patiently standing by my side and for always believing in me.

Without any of you this thesis would certainly have not existed. Thank you to you all. I hope you enjoy your reading.

Laurens van den Bercken
February 20, 2019

[1] https://mytomorrows.com/
[2] https://www2019.thewebconf.org/

Contents

List of Figures

List of Tables

1 Introduction

As health care processes are getting more transparent and patients are getting more involved in their health care, e.g. through patient portals, it becomes increasingly important that health care information is understandable by patients to prevent misinterpretation. A report [78] by the World Health Organization (WHO) concluded that a majority of European citizens has insufficient health literacy (the ability to read and understand health care information, make appropriate health decisions and follow health instructions). One of the major causes of this low health literacy is that health care information contains complex language and specific medical terminology and is often geared to health care professionals. Low health literacy has been associated with poorer health [63] and higher mortality amongst (older) adults [11].

One way to improve health literacy is simplifying medical text, to eliminate complex language and translate specific medical terminology to laymen terms. To illustrate this, we provide an example from Wikipedia and its (manually) simplified version from Simple Wikipedia in table 1.1. Observe that indeed complex language (*estimated prevalence rate in the general population*) and medical terminology (*intracranial neoplasms*) are simplified. However, manually simplifying the continuous stream of medical content is not feasible and therefore an automatic text simplification approach is needed. In this thesis 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?

Wikipedia	**Simple Wikipedia**
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.

Table 1.1: Two health sentences from (English) Wikipedia and Simple (English) Wikipedia

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) [2, 12–15, 62, 65]. This ignores complex non-medical terms and complicated sentence structures, which also hamper readability [80]. The state-of-the-art in automated text simplification, Neural Machine Translation (NMT) [53, 74], 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 [5]. Unfortunately, the resulting set is not publicly available.

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. The resulting corpus consists of a few thousand aligned sentences, which is

generally not enough for training neural models. That is why we introduce a simple novel language independent monolingual text alignment method. The method aligns sentences from one source with sentences (in the same language) from another source. The result is a parallel corpus. We show that it can compete with more complex alignment methods that rely on additional sources like Word2Vec models and Wiktionary, while the introduced method does not. We use it to align additional sentences from Wikipedia disease articles with sentences from the corresponding Simple Wikipedia articles. The NMT can learn simplification of domain-specific terminology more reliably as it gets more examples to learn from. The resulting parallel datasets are made publicly available for future research[1].

We train a state-of-the-art text simplification system (NMT) on aligned complex and simple sentences. Since domain specialization (to the medical domain) is shown to improve translation quality in regular NMT [67], we also analyse the effect of domain-specific training data (in the context of text simplification). We compare an NMT system trained on general aligned sentences only to an NMT system trained on additional aligned health sentences. In addition, we introduce two strategies for translating medical terms to laymen terms. The first is an extra pre-processing step before training the NMT that boosts medical term simplification. We reduce the medical vocabulary by replacing each medical concept encountered in the complex text with a Concept Unique Identifier (CUI) from the Unified Medical Language System (UMLS). Any textual variation of a concept is mapped (or normalized) to a single CUI. Instead of a number of variations of one medical concept, only the CUI is part of the vocabulary, which makes medical concepts less sparse. This can boost medical term translation. The second is a post-processing dictionary approach using the Open-Access and Collaborative Consumer Health Vocabulary (CHV), similar to previous work in medical text simplification [62]. The CHV connects laymen terms to terms used by professionals. After translation we replace medical concepts that are left untranslated with laymen terms from the CHV. We compare both strategies separately with vanilla NMT to measure the effect of both isolated.

We constructed a general set and a medical set of aligned complex and simple sentences. Evaluation is done in the medical domain (a test set of medical sentences). We discuss automatic evaluation (with Simple Wikipedia as a reference) as well as human evaluation. We let humans rate generated simplifications on grammar, meaning preservation from the original sentence and simplicity compared to the original sentence.

Our contributions are summarized as follows:

1. We introduce a novel language independent monolingual text alignment method
2. We publish a medical text simplification corpus[1]
3. We show the effect of domain-specific training data on medical text simplification when training a state-of-the-art automatic text simplification system (NMT)
4. We introduce a method to reduce the medical vocabulary to boost medical concept translation
5. We use the CHV for replacing medical terms with laymen terms in combination with NMT
6. We discuss automatic and human evaluation of NMT in the medical domain and compare the output with manual simplifications from Simple Wikipedia

First, we give an overview of text simplification and medical text simplification research in chapters 2 and 3 respectively. Next, in chapter 4, we describe how we created a new medical text simplification corpus. In chapter 5 we give an explanation of the NMT system we use and the two strategies for medical concept simplification. Chapter 6 explains our first (exploratory) in-house evaluation. Next, in chapter 7 we present and discuss results of that evaluation. Chapter 8 describes changes we did based on the first evaluation. Chapter 9 describes the very similar external evaluation. Chapter 10 then presents and discusses the results of our second evaluation. Lastly, we end with conclusions in chapter 11.

[1] http://research.mytomorrows.com

2

Background: Text Simplification

In this chapter we give an overview of text simplification research. This chapter, together with the next chapter about text simplification in the medical domain, serves as the foundation of our research. We identify the state-of-the-art in both chapters and select and combine useful pieces from both based on the specific challenges in the medical domain.

2.1. Introduction

Text simplification is the process of making text easier to understand, e.g. by reducing vocabulary or grammatical complexity, while preserving its meaning (at some level which is important for the target user). Text simplification approaches can be divided into three main categories:

- Lexical simplification
- Syntactic simplification
- Monolingual Machine translation

Lexical simplification is the replacement of difficult terms or phrases with easier synonyms. First, a set of complex words is identified. Then, for each complex word, a set of candidate terms is generated, regardless of the complex word sense. This stage is followed by word sense disambiguation, to eliminate candidates that do not fit in the context. Lastly, these candidates are ranked on simplicity and the top ranked term replaces the complex term. A general example is given in figure 2.1.

Syntactic simplification is the process of identifying sentences with complex grammatical structure and replacing these with simpler ones. For example long sentences that can be split up into multiple simple sentences and optionally reordered, paraphrased or even dropped. Four tasks are defined in syntactic simplification: splitting, reordering, paraphrasing and deleting. Examples of common syntactic structures are given in table 2.1. A typical example approach is given in figure 2.2. Note that syntactic simplification / transformation is often done in combination with lexical simplification.

Construct	**Example sentence**
Coordination	I worked at home *and he went to the office.*
Subordination	*Before I go*, I need to do the dishes.
Adjectival clause	I go to the university, *which is in Delft.*
Participial phrase	I, *traveling by car*, was fastest.
Appositive phrase	Arthur, *the mechanic*, murders people.
Punctuation	Arthur (*fifty one years old*) is not a real mechanic.
Passive phrase	*Bob was killed by Eve.*

Table 2.1: Syntactic constructs with examples

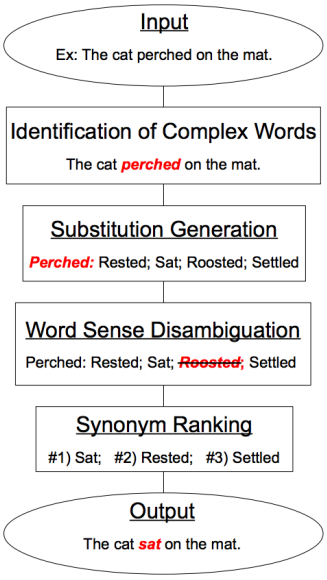


Figure 2.1: Lexical simplification pipeline by example, from [69]

Lastly, monolingual machine translation is divided into Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). On a very basic level SMT models the translation of a foreign text to e.g. English with the probability:

$$\arg\max_e p(e|f) = \arg\max_e p(f|e)p(e)$$

according to Bayes' theorem, with $p(e)$ the language model and $p(f|e)$ being the translation model. The language model is the probability of a sentence e occurring in a language, which can be learned from a general domain corpus in that language. This is used for generating grammatical correct sentences. The translation model is then learned from a parallel corpus by frequency counts. A decoder then searches for the sentence e that maximizes $p(f|e)p(e)$. NMT uses a neural network to learn this statistical model. An encoder-decoder model is often used, which encodes the input sentence to an internal representation and decodes this to the translation in the target language, while maximizing the probability that it is indeed a correct translation. Again, on a very basic level both SMT and NMT can be applied to *translate* from English to Simple English.

Automatic text simplification is usually evaluated using humans that score the output on grammar (fluency), meaning preservation (adequacy), and simplicity on a 1–5 Likert scale. In addition, text simplification based on machine translation is often automatically evaluated using a traditional machine translation metric BLEU [58] and a text simplification specific metric SARI [81].

We review the state-of-the-art in lexical simplification, syntactic simplification and machine translation for text simplification in sections 2.2, 2.3 and 2.4 respectively. We end with a discussion in section 2.5.

2.2. Lexical Simplification

As depicted in figure 2.1, lexical simplification consists of four tasks. This section gives a very brief overview of each of the tasks, based on a recent survey on lexical simplification [57].

2.2.1. Complex Word Identification

Complex word identification (CWI) approaches fall into five categories:

- Simplify everything
  Each token in the text is a candidate for substitution.

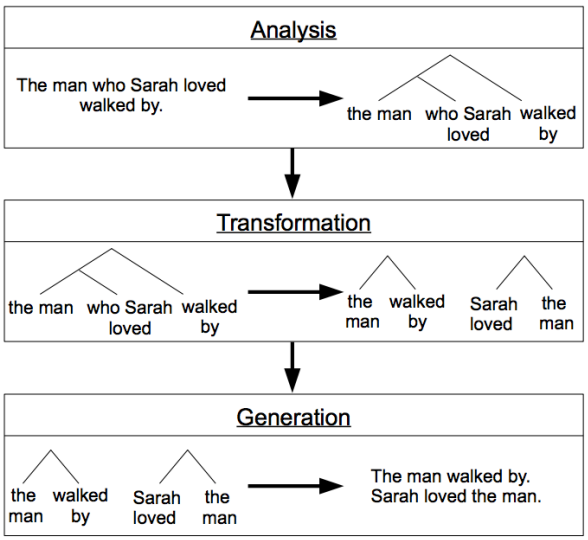


Figure 2.2: Syntactic simplification pipeline by example, from [69]

- Threshold-based
  A threshold-based system searches for a threshold given a simplification metric and a training set. Word frequency is a popular metric.
- Lexicon-based
  Domain-specific text simplification systems can use a lexicon. When a word exists in a lexicon, it should be simplified. For example, in the medical domain the Unified Medical Language System (UMLS) is often used.
- Implicit CWI
  Each token in the text is a candidate for substitution, but only substitutions that replace a word with a simpler word are applied.
- Machine learning based
  A machine learning model is trained on an annotated training set.

In the shared task of complex word identification in 2016 [54] a new data set is constructed combining the CW [68], LexMTurk [28] and Simple Wikipedia [32] corpora. Most submitted systems use machine learning. Interestingly, Decision Trees and Ensemble methods outperform Neural Networks in the task, but this may be due to the limited amount of training data. It turned out that word frequency remains the most effective simplification metric.

In 2018 the data set from [83] is used, which includes annotated English, German and Spanish sentences. In addition a French data set was collected. It was (again) concluded that “traditional feature engineering-based approaches (mostly based on [word] length and frequency features) perform better than neural network and word embedding-based approaches”.

2.2.2. Substitution Generation

The generation of substitution candidates can be done using

- thesauri, e.g. WordNet;
- parallel corpora, e.g. Simple Wikipedia and English Wikipedia or the Newsela corpus;
- or word embeddings, e.g. GloVe [59] or Word2Vec [50];

Most recent work uses (context-aware) word embeddings [55, 56]. A word embedding model is trained on a corpus with words concatenated by their POS tag. Additionally, the model can be retrofitted over synonym relations from WordNet [55]. Then, a set of most similar words are selected as candidate substitutions. Similarity is measured with cosine similarity. Furthermore, the Newsela corpus was used after paragraph, sentence and word alignment [55].

2.2.3. Sense Disambiguation

This stage selects from the set of candidates the set of words which fit in the context of the complex word being simplified, which should maximize meaning preservation, but also grammatical correctness. This stage can be divided in five categories:

- Select all candidates
  Simply select all candidates generated in the previous stage and pass them on, i.e. no disambiguation.
- Explicit sense labelling
  Model sense labelling as a classification task with labels from e.g. WordNet.
- Implicit sense labelling
  Group words that appear in similar context based on a latent variable language model.
- Part-of-speech tag filtering
  Select only candidates that have the same POS tag.
- Semantic similarity filtering
  Select only candidates that have a high semantic similarity with the complex word.

Recent work on lexical simplification uses semantic similarity filtering [24, 55, 56]. Not only the cosine similarity between the complex word and the target word is used, but also the similarity between the context around the complex word and the candidate [24]. Alternatively, a machine learning approach is used with context-aware word embedding features and generated training data based on the hypothesis: “a given target complex word is the only word suitable to replace itself” [55, 56].

The problem can also be defined as an optimization problem with WordNet as a sense inventory [4]. Note that the sense disambiguation is done on the complex word and subsequently substitution generation is done with WordNet. Since this optimization problem is an NP-hard problem, two solutions are proposed to find near-optimal solutions: simulated annealing [33] and D-Bees [3].

2.2.4. Substitution Ranking

The last stage is ranking the remaining candidates based on simplicity. Approaches fall in three categories:

- Frequency based
  This is based on the assumption that more frequent words are easier. Frequencies come from corpora, such as movie subtitles, English Wikipedia and Simple Wikipedia, Google 1T Corpus, Microsoft N-gram Services and search engine results.
- Simplicity measures
  Using some defined metric of simplicity, e.g. combining frequency and word length.
- Machine learning based
  A machine learning based ranker is trained.

Each category has recent notable contributions. A frequency based approach using subtitles for children and families was presented in [56]. A ranking based on simplicity measures, such as word embedding similarities, frequencies, and informativeness, was presented in [24]. Lastly, a neural model for substitution ranking was presented in [55]. The neural model reports the best results, but requires training data. The unsupervised ranking approach does not perform much worse though and outperforms frequency based ranking.

2.3. Syntactic Simplification

Syntactic simplification simplifies whole sentences as opposed to only complex words or phrases. Four (separate) tasks in syntactic simplification are splitting, reordering, paraphrasing and deleting (parts of) complex sentences. We found that most approaches focus on splitting and deleting, while reordering and paraphrasing is done implicitly (as a consequence of splitting and deleting) and is mostly limited to lexical simplification, respectively.

2.3.1. Splitting

Complex sentences can be split into multiple simple sentences, e.g. the complex sentence “The man, carrying numerous books, entered the room.” can be split into two sentences “The man entered the room.” and “He was carrying numerous books.”. Text is often parsed using the Stanford dependency parser [34]. Hand-written rules are then applied to split sentences containing complex grammatical structures [19, 20, 41, 45, 64, 70], as given in table 2.1. In addition, this is combined with automatically learned lexical simplification rules [45, 70] from aligned sentences or a traditional lexical simplification component [19, 20], as given in figure 2.1.

More recent approaches use an additional step when a complex pattern is detected. A detected complex pattern is classified by a decision tree to predict a split [41]. The decision tree is trained on Simple English Wikipedia triplets. Each triplet consists of the original sentence and the split version consisting of two sentences (similar to the example given earlier). Similarly, a complexity checker and confidence model are used for identifying sentences to simplify and for deciding whether the output is good enough to show to the end-user respectively [64].

Lastly, Boxer [16] is used to parse a complex sentence to a Discourse Representation Structure (DRS). Then each pair of event variables (i.e. node that describes an event) is a split candidate [52]. Split probabilities are learned from aligned sentences from English and Simple Wikipedia. Similarly, [72] generates a sentence for each event mention. Additionally, pronominal entity mentions are resolved [72], using Stanford’s co-reference resolution tool [40]. This approach is combined with the unsupervised lexical simplification approach introduced in [24].

2.3.2. Deleting

Irrelevant information can be deleted from sentences, which could make a sentence simpler. For example, the sentence “A couple of years ago I published a tiny scientific article about text simplification for the first time.” can be compressed to “A couple of years ago I published a scientific article about text simplification.”, deleting the adjective *tiny* and the phrase “*for the first time*”. Note that we do not consider sentence compression as a task on its own, but sentence deletion in the context of sentence simplification. All work described in this subsection is accompanied with other simplification operations.

Discourse is used to delete irrelevant information [52, 72]. First, sentences are parsed to a graph of discourse structure. Then, events are extracted. Prepositional phrases, adverbs, adjectives and orphan words (words that do not belong to either sentence after a split) are candidate deletions [52]. Deletion probabilities are learned from Wikipedia. Alternatively, sentences that do not contain an event mention are simply deleted [72].

A system that adds deletion to [70], compresses sentences by generating candidates that miss parts of the sentence and that largely comply with English syntax. Then for each sentence the best candidate is selected in a way that benefits the whole text (by linear programming) [46].

Machine learning approaches try to learn what to delete. An approach “using conditional random fields over top-down traversals of dependency graphs that jointly predicts possible compressions and paraphrases” is introduced here [9]. First, a dependency tree is constructed. Then, a top-down approach is used: The CRF predicts for each sub-tree whether it can be deleted, paraphrased or left. For paraphrasing the Paraphrase Database (PPDB) [22] is used. For deleting the Google compression data set is used. The problem can also be seen as a sequence labeling task [6]. Each word is labeled as `DELETE`, `ORIGINAL` or `REPLACE` (and others, but only these are used). The labeling is done using a “bidirectional recurrent neural network, with an initial embedding layer of size 300 and two hidden LSTM (Long-Short Term Memory) layers of size 100” [6]. Training data is automatically generated from aligned sentences. Words that were classified as `DELETE` are simply deleted and words with the `REPLACE` label are replaced using a lexical simplification approach [55].

2.4. Monolingual Machine Translation

Machine translation can be used to *translate* from English to Simple English. Such systems learn from parallel corpora, e.g. aligned Wikipedia and Simple Wikipedia sentences. While the previous sections explicitly deal with lexical and / or (parts of) syntactic simplification, machine translation models both implicitly based on training data.

Machine translation approaches fall into two main categories, namely Statistical Machine Translation (SMT) and Neural Machine Translation (NMT), which both are supervised approaches. Due to the

lack of large parallel corpora for non-popular language pairs, a third category was introduced, namely Unsupervised Machine Translation (both SMT and NMT).

2.4.1. Statistical Machine Translation

Early work that used SMT for text simplification observed that it mainly reorders and performs substitutions and therefore proposed a hybrid system that "combines a model encoding probabilities for splitting and deletion with a monolingual [phrase-based statistical] machine translation module which handles reordering and substitution" [52]. After splitting and deleting, the phrase-based SMT substitutes words and reorders parts of sentences. The SMT is implemented using the Moses toolkit[1] and trained on aligned Simple and English Wikipedia sentences.

Later, the impact of quality of the data set is investigated by training several phrase-based SMT models on data sets with different sizes and sentence similarities [73]. It was concluded that quality has a greater impact than quantity, i.e. models trained on carefully selected sentence pairs with moderate similarity based on S-BLEU perform best, regardless of the training set size, which ranges from 2,000 to 10,000.

Xu et al. [81] propose to use their text simplification specific metrics when tuning a syntax-based SMT system using the synchronous context-free grammar in the PPDB [22], which can be used directly by a SMT decoder. The SMT is then tuned using manual simplification from workers from Amazon Mechanical Turk, who provided multiple reference sentences, and the two new tuning metrics. The system was implemented in the open source syntactic machine translation decoder Joshua[2] [61] and tested on a test set with 8 references from workers from Amazon Mechanical Turk. This test set is re-used in later works, described in the next section.

2.4.2. Neural Machine Translation

More recently, machine translation moved to NMT. This work claims to be the first to apply NMT to the problem of text simplification [53] and use the OpenNMT framework [35] to build the network. An NMT architecture is used with "two LSTM layers, hidden states of size 500 and 500 hidden units, and a 0.3 dropout probability". Additionally, global attention with input feeding is used [44]. Furthermore, two Word2Vec models are trained. One trained on Google news and the original English text and one trained on Google news and the simplified English text. During translation beam search is applied to find the best translation (i.e. sequence of words). 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. BLEU with NIST [10] smoothing and SARI [81] are then used to select the best beam size and hypothesis using the validation set.

Later, this approach is combined with a semantic splitting algorithm [74]. Training data is first parsed to the UCCA (Universal Cognitive Conceptual Annotation) [1] scheme using the TUPA parser [27]. Two types of sentence structures are then split:

1. Parallel scenes, e.g. "He came back home and played piano" to "He came back home" and "He played piano" [74]

2. Elaborate scenes, e.g. "He observed the planet which has 14 known satellites" to "He observed the planet" and "Planet has 14 known satellites" [74]

After applying these split rules in the training data, the same NMT as [53] is trained on it. Human evaluation showed that this approach produces simpler and simpler structural output than [53]. Grammar and meaning preservation scores, as well as automatic metrics BLEU and SARI, are lower though.

2.4.3. Unsupervised Machine Translation

Unsupervised machine translation does not need parallel data [38]. It learns from separate monolingual corpora. Figure 2.3 shows the trade-off in training data size when training machine translation systems (for translating from English to French). BLEU ranges from 0 to 100 and higher is better. To the best of

[1] `http://www.statmt.org/moses/`
[2] `http://joshua-decoder.org/`

our knowledge this has not yet been tested for the text simplification task (i.e. translating from English to Simple English).

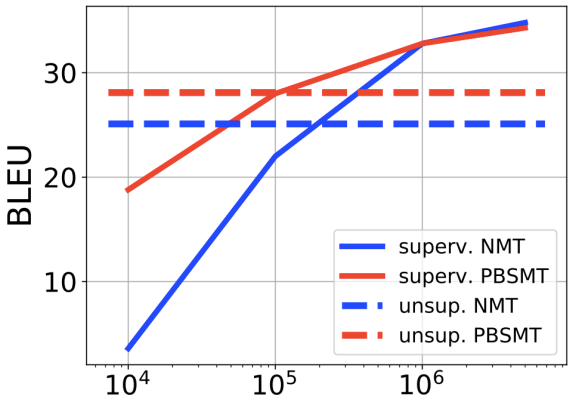


Figure 2.3: Comparison between supervised and unsupervised machine translation on WMT'14 En-Fr, from [38]

2.5. Discussion

Lexical simplification is the identification of complex words and phrases and replacing them with simpler synonyms. In this chapter we discussed state-of-the-art approaches in lexical simplification. Frequency and word length features are often used for identifying complex words. Context-aware word embedding models are used for generating substitutions and word sense disambiguation. Ranking candidates with a neural network shows best results, but requires training data. An (unsupervised) approach using frequency on a subtitle corpus was also shown to be effective. Note that these approaches perform single word simplification. While lexical simplification can especially be beneficial for people with a limited vocabulary, e.g. second language learners, it does not alter the grammatical structure of the text.

Syntactic simplification aims to split, delete, paraphrase and reorder complex sentences, based on the identification of complex sentence constructs, like in table 2.1. However, approaches that eliminate such structures by identifying complex sentences and splitting them into simpler ones rely on costly and carefully constructed hand-written rules. Alternatively, discourse is used to split sentences, i.e. not using hand-written rules and therefore not explicitly tackle complex structures, are showed to produce more readable and simpler texts, while achieving similar grammar and meaning preservation scores when compared to hand-written rules approaches [72]. Additionally, they are able to delete irrelevant information. Syntactic simplification can also be seen as a sequence labeling task, but this is not suitable for splitting sentences [6].

Machine translation models both lexical and syntactic simplification implicitly based on training data, which is available on English and Simple English Wikipedia and the Newsela corpus. Machine translation can either be done with traditional Statistical Machine Translation or Neural Machine Translation, both supervised and unsupervised. Most recent work in text simplification using machine translation uses NMT, for which it is shown that it outperforms SMT approaches. Unsupervised machine translation has not been explored in the text simplification task.

In the next chapter we discuss text simplification specifically in the medical domain. We list the state-of-the-art in the same three categories. Lastly, we end with a discussion how to combine useful pieces from both chapters based on the specific challenges in the medical domain.

3

Background: Medical Text Simplification

In this chapter we describe the state-of-the-art in medical text simplification. First, we discuss challenges in the domain and how it is different from general text simplification. The chapter is further outlined in sections describing the different approaches to text simplification in the medical domain. In the discussion we make the comparison between medical and non-medical (previous chapter) text simplification research. We conclude the chapter with how we can combine useful pieces from (general) text simplification and medical text simplification based on the specific challenges in the medical domain.

3.1. Introduction

There are many types of medical texts, such as clinical text found in Electronic Health Records (EHRs), clinical trial results, medical research, drug labels and patient information leaflets. Such texts can contain specific terminology, i.e. medical concepts. Simpler synonyms, i.e. terms that are generally understood by the general public, cannot be found in typical resources such as WordNet and existing parallel corpora, such as Wikipedia and Simple Wikipedia and the Newsela corpus. Moreover, medical text in general, in addition to containing medical terminology not known to laymen, is acknowledged to be too complex for laymen [80]. This was already illustrated in table 1.1.

This chapter follows the same structure as previous chapter. For each category of text simplification we include a section about recent literature. We end the chapter with a discussion.

3.2. Lexical Simplification

A lot of research in medical text simplification focused on (medical) lexical simplification [2, 12–14, 62, 65], i.e. focusing on replacing professional medical terms by laymen terms. In this section we give an overview of research that focused on (parts of) lexical simplification within the medical domain.

3.2.1. Complex Word Identification

In medical text complex words are usually medical concepts found in professional medical vocabularies and terminologies. One approach simply checks for each word whether a simpler synonym exists in Medical Subject Headings (MeSH) [2], which is a controlled vocabulary for indexing medical articles. More recent approaches [12, 13, 62] use MetaMap [7] to detect medical concepts and map them to the Unified Medical Language System (UMLS). UMLS is a metathesaurus, which contains unified entities from different source vocabularies and terminologies.

All medical concepts identified by MetaMap can simply be considered as complex words [62]. Alternatively, these concepts can be considered as candidates. These candidates are then ranked by e.g. a Support Vector Machine [13] or a random walk algorithm [12]. Transfer learning was also applied to detect important terms for lay language annotation [14]: Feature Space Augmentation (FSA) [17] and Supervised Distant Supervision, based on [77]. FSA combines shared features from two domains and the domain-specific features, while SDS minimizes error from two domains. The source-domain is distantly supervised and the target-domain is the manually labeled data by experts. A log-linear model

is then trained and ranking is based on output probabilities. Manually annotated data can be used for training [13]. Others are distantly supervised by the Open-Access and Collaborative Consumer Health Vocabulary (CHV), which is a vocabulary for health consumers, which contains laymen synonyms for medical concepts, mined from medical forums and Q&A systems [12, 14]. Candidate terms that occur in the CHV are assumed to be important for the patient.

Segura-Bedmar and Martínez developed a Named Entity Recognition system [66], based on a dictionary-based approach, to detect medical concepts to be simplified. The dictionary is built from the ATC system, CIMA, MedDRA and medical websites, such as MedlinePlus and other Spanish medical websites.

3.2.2. Substitution Generation

Word2Vec models, trained on Wikipedia, can be used to generate substitutions for medical concepts [65]. Semantically similar terms are selected as candidates, based on similarity between the candidate and the original word and the candidate and the context around the original word. But, this brings the limitation that candidates consist of one word only. To overcome this phrase embeddings are used in addition to word embeddings, i.e. the average of all word embeddings of (noun) phrases that contain at least one MedDRA word, keeping certain patterns of Part-of-Speech only. Similarity is calculated using cosine similarity.

Alternatively, controlled vocabularies can be used. As was mentioned before, [2] simply replaces all words that have a simpler synonym in MeSH with that synonym. When making use of MetaMap one could leverage the mapping to the UMLS, where the CHV is part of. Each identified medical concept is replaced with its CHV-preferred term, which, intuitively, should make text easier to comprehend for health consumers [62]. The section below explains works dedicated to constructing and enriching such vocabularies.

Consumer Health Vocabulary

A number of works has focused on constructing or enriching the CHV, which can be used for complex word identification and substitution generation. Wikipedia is used to mine consumer and professional health terms using a pattern-based approach [76]. Articles on Wikipedia often contain alternate forms, abbreviations and synonyms. For example the article about Xerostomia begins with: "**Xerostomia**, also known as **dry mouth** and **dry mouth syndrome**, is dryness in the mouth...". They identified twelve common linking phrases: also called, also known as, also referred to as, commonly called, commonly known as, commonly referred to as, sometimes called, sometimes known as, sometimes referred to as, also termed, previously known as and colloquially known as. First, only health / medical articles are filtered. Then, from the title and the leading text paragraph, all bold faced, italicized and hyperlinked phrases are identified as candidates. Next, terms should be labeled as consumer or professional. This is done based on the common assumption that a term frequently used by laypersons and medical professionals are likely consumer-preferred terms and professional terms respectively. MedHelp was chosen as a consumer text corpus, while abstracts of articles published in scientific journals were chosen as the professional text corpus. By manually evaluating 100 pairs, it was observed "that 89% of the pairs are between synonymous or equivalent concepts".

Other works focused on extracting new consumer health expressions only, i.e. not on extracting new pairs. A co-occurrence analysis based approach is used to expand the CHV by learning from a popular online health community – MedHelp [30]. It is based on the assumption that "two (or more) words that tend to occur in similar linguistic context (i.e. to have similar co-occurrence patterns) tend to resemble each other in meaning" [39]. CHV terms are used as seed terms. Then, candidate terms are extracted using co-occurrence analysis and ranked based on the strength of the association between seed and candidate. MedHelp was also used to extract consumer health expressions in [82]. First, given a post, similar comments are selected using the Kullback-Leibler divergence of two topic vectors computed with LDA. The top k comments are merged together with the original post. On this key-phrase extraction is applied, using frequent word sequences [84]. Such a sequence appears in at least σ documents and is initially a sequence of two that ends with a noun. Then, each iteration one word is added, until the condition does not hold anymore. Only nouns and adjectives are kept. The posts and comments are tagged using Stanford Part-of-Speech tagger [75]. Sequences are given a phrase score and the top m are considered consumer health expressions.

Lastly, *Yahoo! Answers* is used, specifically the question and the best answer, to extract health

consumer expressions [26]. First, the text is split into sentences and tokens, including Part-of-Speech tags using OpenNLP[1]. Tokens are normalized using the Lexical Variant Generation tool [48]. Stop words are removed. N-grams up to 5 are extracted. N-grams that occur less than 5 in the data set are also removed. Using fuzzy matching, n-grams are matched to the CHV, which are considered as seed terms. Note that each CHV term is also normalized. Non-CHV terms are considered as candidate terms. Seed terms are then clustered using K-means with Euclidean distance. Both word and context features are used. The distance from a candidate term to the nearest cluster is the score of that term. In addition, clusters are given a score based on their average TF of the terms belonging to that cluster. Candidate terms are first sorted on cluster score, then on distance to nearest cluster.

3.2.3. Sense Disambiguation

Recall that this stage selects from the set of substitution candidates the set of substitutions which fit in the context of the medical concept being simplified, which should maximize meaning preservation, but also grammatical correctness. Just a few works actually focus on the whole lexical simplification task [2, 62, 65], but only one has some sort of word sense disambiguation [65]. Note that a dictionary approach using some vocabulary is not context aware. In [65] Word2Vec models are used to generate candidate terms. Instead of just taking the synonym candidate which is semantically similar the most, the context around the original word is also considered. The average of the cosine distance of all context words and the synonym candidate is computed as this similarity [24] as well.

3.2.4. Substitution Ranking

For assessing if a synonym is actually easier, frequency analysis is often used. If a synonym is more frequently occurring in a general corpus, then it is considered as easier [2]. [65] uses the degree of informativeness, defined in [18], also using frequency analysis. Only if the candidate is less informative than the original word it is replaced. The MetaMap approach simply selects the CHV-preferred term [62].

3.2.5. Explanation Generation

Explanation generation can be seen as a variant of lexical simplification. It is the automatic generation of explanations of complex words. Hence, this is different from lexical simplification, since complex words are not simply replaced by simpler synonyms, but instead a whole sentence containing a(n) explanation / definition is showed.

One study examined the effect of automated health explanations on older adults (rather than focusing on automatically generating these explanations) [47]. To identify medical concepts a controlled vocabulary was created from the Plain Language Health Thesaurus and Medline Plus [21]. For each term that appears in the vocabulary an explanation is given. Results showed that participants were more satisfied and found the text more helpful. However, the ability to recognize over-the-counter medication risks did not increase.

A later approach that focused on automatically generating explanations, i.e. lay definitions, for medical terms, introduced NoteAid[2] [15]. The proposed method is again two-fold: first identify medical concepts, then link these concepts to laymen definitions. CoDeMed was developed to contain these lay definitions. It was constructed using human efforts, but also by mining medical synonyms from Wikipedia. They “used the interwiki links in pages on the Wikipedia Health tree to extract candidate synonyms for medical terms”. These candidates are then ranked using cosine similarity between word embeddings of the target term and its candidate term and pseudo-relevance feedback. For identifying medical concepts MetaMap is used. These terms are then linked to terms in CoDeMed by dictionary lookup by the second module called MedLink. An example output of the system is depicted in figure 3.1.

3.3. Syntactic Simplification

There is little work on automatic syntactic simplification in the medical domain. However, there is an approach that focused on manually splitting noun phrases and investigating the effect of it [43], after an initial validation which found that for easy texts noun phrase complexity is lower [42]. From the

[1] https://opennlp.apache.org/
[2] http://www.clinicalnotesaid.org/emrreadability/notesaid.uwm

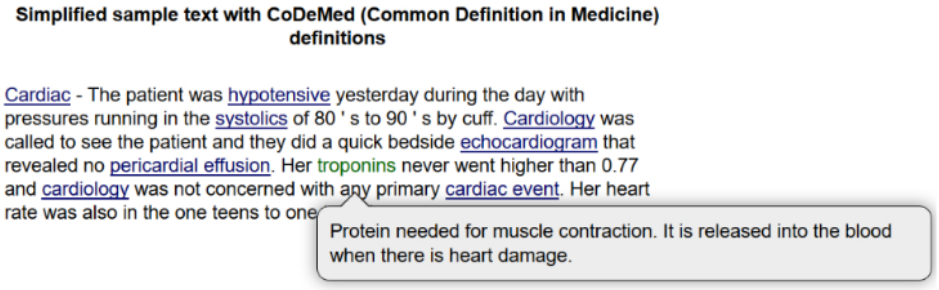


Figure 3.1: Example output of generated explanations, from [15]

Disease category on English Wikipedia all articles (8,247) were collected. Then sentences "containing six or more nouns and at least one noun phrase containing two words (2-gram), three words (3-gram) or four words (4-gram)." are selected, using the Berkeley Parser [60]. Sentences that contain medical concepts or proper nouns were discarded and only sentences where the noun phrase is in the Google Web Corpus are kept. Furthermore, only sentences are kept of which its constituents after splitting the noun phrase occur more frequently in the Google Web Corpus (based on previous work that showed familiar terms are easier to understand). An example is given: "the frequency of "motor nerve conduction velocities" is 791 (in the Google Web Corpus) and after splitting into "conduction velocities of motor nerves" the frequency increases to 10,498 (conduction velocities) and 12,804 (motor nerves) with an average of 11,650". The approach was evaluated using Amazon Mechanical Turk, to measure perceived and actual text difficulty. For measuring perceived difficulty workers were asked to rate sentences on a 5-point Likert scale. For measuring actual difficulty an adjusted Cloze test was used: four nouns were deleted from the sentence; workers were then asked to fill in the blanks through choosing 1 out of 4 options. It was concluded that splitting noun phrases did not make the sentences easier, although they were perceived as easier. It was also concluded that term comprehension has a great impact on overall comprehension through the inclusion of pseudowords (representing difficult words) in the sentences. In fact, "when pseudowords were present, the perceived difficulty of the sentences remained the same regardless of whether the noun phrase was split or not", indicating the importance of lexical simplification in the medical domain.

Other work looked at negation and the impact of it on text readability and the ability to predict text difficulty based on it. A negation parser for medical text simplification is introduced [51]. Three types of negation are detected: sentential (e.g. no, none, not), morphological (e.g. irrelevant, impossible, limitless) and double negation (e.g. not impossible). Blogs and Simple Wikipedia represent easy medical texts and PubMed, Cochrane, clinical trials and English Wikipedia represent difficult medical texts. Results show that morphological negations occur less frequently in easy texts than in difficult texts. This is possibly due to that e.g. "not clear" is more familiar to a reader than "unclear", since "not" and "clear" occur more frequently than "unclear", and is therefore considered as easier. At the same time that could be the reason that in easy text sentential negation occurs more frequently than in difficult text.

3.4. Monolingual Machine Translation

Limited work has been done on machine translation for medical text simplification. One approach uses Moses, an SMT system, to automatically simplify medical text on slide shows about breast cancer for deaf people. The system was trained on Simple and English Wikipedia. However, "the resulting simplified English paragraphs were imperfect and required a second manual step to be consistent and accurate" [37]. But, after the two-step simplification method, it was concluded that deaf people performed significantly better on a multiple choice quiz about the material.

Another approach focused on generating paraphrases with NMT and compare it to SMT [25]. Training data comes from the PPDB and SNOMED CT. The latter is a medical terminology, which contains description terms for medical concepts. It was concluded that NMT performs on par with SMT, but SMT tends to copy the input, while NMT introduces more novelty, meaning that it potentially produces more useful output. It should however be noted that this approach focused on paraphrasing parts of sentences and medical concepts, e.g. simplifying "*contagious diseases*" to "*communicable diseases*",

essentially performing lexical simplification only. Moreover, it was not evaluated on whole sentences, but on PPDB paraphrases and medical concepts and description terms from SNOMED CT.

Lastly, a first attempt towards creating a medical text simplification data set is done in [5]. Sentences from complex medical sentences and simple medical sentences are aligned using the mean of several sentence similarity scores and the output of a "stacked bidirectional long short term memory (BiLSTM) layers in a Siamese architecture" trained on Quora question pairs, and Paralex question pairs. Then, an optimal threshold is searched. Sentences from 164 health articles from Wikipedia and Simple Wikipedia are aligned, resulting in 1,491 sentence pairs. In addition, sentences from `https://www.webmd.com` and `https://www.medicinenet.com` are aligned. Sentences from medicinenet are considered complex and sentences from webmd are considered simple. A Google search with the titles of the 164 articles is done and sentences in the results from both websites are extracted and subsequently aligned, resulting in 1,002 additional sentence pairs. The result is a (not publicly available) medical text simplification parallel data set of 2,493 sentence pairs. It is used to train an NMT system. The system was evaluated using traditional machine translation metrics only and not by human evaluators.

3.5. Discussion

Most work on medical text simplification focused on lexical simplification. More specifically, the focus is on the replacement of medical concepts by laymen synonyms or expressions. For finding these synonyms and expressions the CHV is used. However, it is found that this is incomplete and still too difficult for laymen [62]. A word embedding approach was promising, since this has worked in the general domain with great success [24], but, evaluation showed that it performs far worse than a basic dictionary approach [65]. However, an "error analysis shows that some of the system answers might be valid and simple synonyms, even though they are not the same as proposed by the gold-standard corpus". Therefore, a second evaluation using multiple synonyms for the gold-standard or perhaps using humans is needed.

The generation of explanations of medical concepts was also explored, which is considered to be a variant of lexical simplification. It was shown that readers were indeed more satisfied with explanations of medical concepts, nonetheless not better at recognizing medication risks. A semi-automatic system was built using human efforts and Wikipedia, but was only evaluated by physicians. One clear advantage that explanation generation has over other simplification approaches is that it does not alter the original text and hence it remains in its original state. Therefore, errors made by such a system do not affect the readability of the text, while errors made by other approaches does. On the other hand, the already complex and potentially long medical texts become even longer.

Since the CHV could play a vital role in medical lexical simplification, several work focused on enriching or constructing CHVs from user-generated content. While most work only extract laymen terms and expressions, aiming to make CHVs more comprehensive for laymen, we found only one work that adds new pairs of medical concepts and its laymen synonyms, aiming to make CHVs more complete. Interestingly, the impact of such approaches on lexical simplification using the CHV was not evaluated in these works. While lexical simplification is widely acknowledged to be very important for comprehension of medical text (due to the vocabulary gap between health professionals and consumers), it is also acknowledged that lexical simplification alone is not sufficient for medical text simplification [80].

Yet few works investigated medical text simplification other than (only) simplifying professional medical concepts. The (manual) splitting of complex noun phrases did not make medical text easier. Furthermore, negations in medical texts were investigated, but nothing more than that. It was shown that easier text contain less morphological negations than difficult text. An easy text contains for example "not clear" instead of "unclear", which is what a general lexical simplification approach based on frequency would also produce. A preliminary work extracted a (fairly small not publicly available) medical text simplification dataset from web-based knowledge sources and trained an NMT system on it, but do not include human evaluation.

Given that

1. current work in (automatic) medical text simplification mainly focuses on simplifying medical terminology only;

2. medical text can also benefit from simplifying complex non-medical terms and complicated sentence structures;

3. limited work is done on using state-of-the-art text simplification (NMT) in the medical domain

we focus in this thesis on evaluating the state-of-the-art in text simplification (NMT) in the medical domain. It can learn to perform lexical simplification and syntactic simplification implicitly, based on training data. We investigate the impact of two directions that intuitively should improve text simplification capabilities of NMT in the medical domain:

1. Data: We want to learn the impact of domain-specific training data on text simplification, since it improves translation quality in regular NMT [67]. As there is no publicly available medical text simplification corpus, we create one and make it publicly available for future research[3]. We compare NMT trained on general sentences with NMT trained on general and medical sentences.
2. Simplification of medical terms: We introduce an extra pre-processing step that boosts an NMT to learn medical term simplification. We also evaluate NMT in combination with a dictionary approach using the CHV for replacing medical terms with laymen terms.

All approaches are evaluated using automatic metrics and human evaluators. The latter, a common practice in text simplification research, is not yet done in the medical domain for NMT. First, we need to create a medical text simplification corpus, which is where the next chapter continues.

[3] http://research.mytomorrows.com

4

Medical Text Simplification Corpus

In this chapter we describe how we created two new aligned medical text simplification datasets for in-domain training, as there are no publicly available medical text simplification corpora. A medical text simplification corpus may contain simplifications of medical terminology to laymen terms which an NMT can learn. The first dataset (expert) is an expert-evaluated medical subset filtered from the aligned Wikipedia corpus presented by Hwang et al. [29]. The second dataset ($\mathrm{automatic}$) is a novel dataset created by automatically aligning sentences from disease articles on Wikipedia and Simple Wikipedia using a novel alignment method.

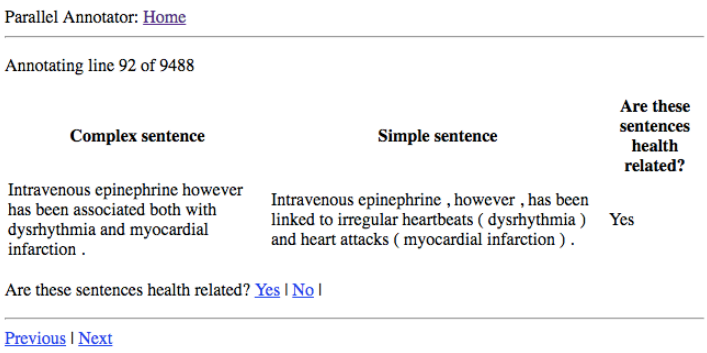
4.1. Expert-evaluated

Our expert dataset is created using the aligned corpus presented in [29], which aligns sentences between Wikipedia and Simple Wikipedia. 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 this section we will refer to the *good* sentence pairs as *fully aligned* and to the *good partial* as *partially aligned* sentence pairs.

Our expert dataset is a subset of the corpus presented in [29]. We filter out the set of aligned medical sentences. We use a state-of-the-art medical named entity recognition (and linking) tool, QuickUMLS [71], to determine whether a given sentence pair (complex - simple) is health related or not. QuickUMLS is an approximate dictionary matching algorithm which matches terms from text with terms 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 (or both). After QuickUMLS processing, we provide 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. We designed an extremely simple web page, illustrated in figure 4.1, for annotation. Each *fully aligned* and *partially aligned* sentence pair in [29] is run through this approach, resulting in a filtered corpus of 2,267 *fully aligned* medical sentences and 3,148 *partially aligned* sentences.

4.2. Automatically Aligned

Since a few thousand aligned sentences is generally not enough for training neural models, we describe how we created a second medical text simplification dataset fully automatically in this section. We propose a simple novel language independent monolingual text alignment method and demonstrate it by aligning additional complex and simple health sentences from Wikipedia and Simple Wikipedia disease articles. First, we describe the novel text alignment method. Next, we describe how we used it to align additional health sentences. This results in our second dataset, $\mathrm{automatic}$.


Parallel Annotator: Home

Annotating line 92 of 9488

Complex sentence	Simple sentence	Are these sentences health related?
Intravenous epinephrine however has been associated both with dysrhythmia and myocardial infarction .	Intravenous epinephrine , however , has been linked to irregular heartbeats (dysrhythmia) and heart attacks (myocardial infarction) .	Yes

Are these sentences health related? Yes | No |

Previous | Next

Figure 4.1: An extremely simple annotation tool

4.2.1. A Novel Monolingual Text Alignment Method

Given a sentence pair (complex - simple), we need some metric to measure how similar to each other they are: a sentence similarity score. If the two sentences are highly similar, it is highly likely that they are a valid alignment. We use the sentence BLEU score [58] as a sentence similarity score. The BLEU score is used in machine translation to measure how good an output translation is by comparing it to a reference or multiple. More specifically, it is measured by counting overlapping word n-grams. We do not include the equation here, since it needs a lot of background insights. For more details we refer to [58]. The score is between 0 and 1, with 1 being a perfect translation, i.e. the two sentences of a pair are identical to each other (or in other words are a perfect alignment). In our sentence alignment we use *character* n-grams up to 12 and uniform weight distribution, i.e. $1/12$, since these settings gave best results in the evaluation described below. In addition we use the NIST [10] smoothing function. To the best of our knowledge, BLEU score has not been used before in the alignment task.

We compare BLEU alignment with our own implementation of maximum alignment, which was considered to be the best alignment method among several in [31]. Given a sentence x and y a sentence similarity $\text{sim}_{\text{max}}(x, y)$ is calculated by aligning each word in x with the most similar word in y and vice versa. Word similarity $\phi(x_i, y_j)$ is calculated using pre-trained Google news Word2Vec embeddings [49]. Word similarity threshold is set to 0.5, similar to [31]. Words in the input sentences that are not in the vocabulary of the Word2Vec model are filtered out.

To evaluate the quality of the BLEU alignment for the sentence alignment task, we compare BLEU alignment to our implementation of Maximum alignment and results reported by Kajiwara and Komachi [31]. We use the manual alignment set from Hwang et al. [29] as evaluation set. 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 binary classification scenarios:

1. Fully aligned sentences versus the rest
2. Fully and partially aligned sentences versus the rest

We search for the best threshold value t. A sentence pair with sentence similarity $s > t$ is considered an alignment, else it is not. Precision p, recall r and F_1 score are then defined as follows:

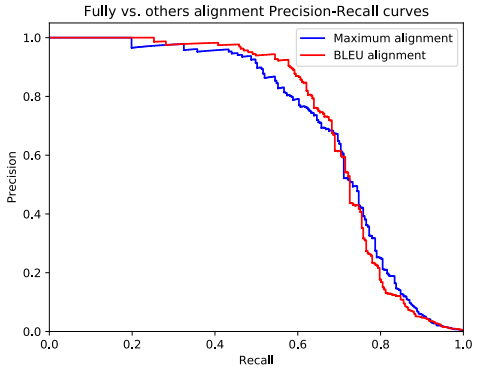
$$p = \frac{|\{\text{correct alignments}\}| \cap |\{\text{predicted alignments}\}|}{|\{\text{predicted alignments}\}|}$$

$$r = \frac{|\{\text{correct alignments}\}| \cap |\{\text{predicted alignments}\}|}{|\{\text{correct alignments}\}|}$$

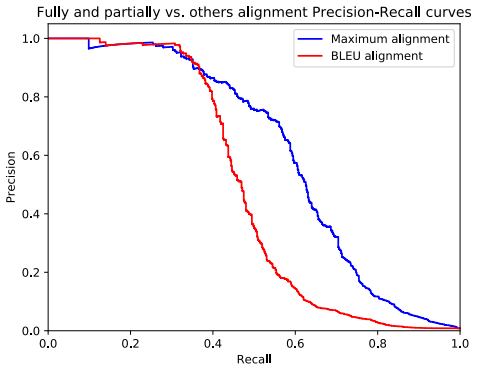
$$F_1 = 2 \cdot \frac{p \cdot r}{p + r} \text{ (harmonic mean of } p \text{ and } r)$$

Figure 4.2 shows precision-recall curves of the two scenarios of both methods (our implementation). Table 4.1 summarizes the figure, reporting maximum F_1 score and Area Under the Curve (AUC) (also known as average precision). Maximum F_1 score is reached at a certain threshold t_m for each alignment

method m. It combines precision and recall scores. The AUC score is an overall measure over the whole range of precision and recall. Note that the maximum of precision, recall, F_1 score and AUC is 1.



(a) Precision-recall curve of fully aligned vs. rest: Maximum alignment (AUC=0.704) and BLEU alignment (AUC=0.714)



(b) Precision-recall curve of fully and partially aligned vs. rest: Maximum alignment (AUC=0.611) and BLEU alignment (AUC=0.484)

Figure 4.2: Precision-recall curves of Maximum alignment and BLEU alignment

For fully aligning sentences BLEU score performs on par (see figure 4.2a) with more complex sentence alignment methods based on word embeddings [31] and Wiktionary [29]. Despite the poorer performance on partial alignment (see figure 4.2b), 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 not in the vocabulary of such external datasources, which may deteriorate performance of approaches that use them. BLEU only looks at overlapping n-grams, which makes it domain and even language independent. Next, we describe how we used BLEU to align additional complex and simple health sentences.

Fully vs. rest	**Max F_1**	**AUC**
BLEU alignment	0.717	0.714
Maximum alignment	0.687	0.704
Maximum alignment in [31]	0.717	0.730
Alignment used to align Wikipedia [29]	0.712	0.694
Fully and partially vs. rest		
BLEU alignment	0.534	0.484
Maximum alignment	0.624	0.611
Maximum alignment in [31]	0.638	0.618
Alignment used to align Wikipedia [29]	0.607	0.529

Table 4.1: Max F_1 and AUC scores for identifying fully aligned and fully and partially aligned sentences

4.2.2. Additional Alignments

Because more in-domain data seems to have a positive effect on translation quality when training an NMT [67], we demonstrate our proposed method for sentence alignment by aligning additional disease articles from Wikipedia and Simple Wikipedia. Recent work, by Kajiwara and Komachi [31] and Adduru et al. [5] 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. Therefore, we propose an approach using DBPedia [8] and select all English articles that fall in the *dbo: Disease* class. After title matching to Simple Wikipedia, this gives us a set of 1,098 aligned articles. Analogous to Kajiwara and Komachi, we extract the text from the Wikipedia and Simple Wikipedia articles, using the python Wikipedia API [1] and tokenize into sentences using NLTK 3.3 [2]. This gave an average number of words per sentence of 26.1 for the English articles and 19.5 for the simple articles. The average numbers of sentences per article were 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 numbers of sentences per article were 57.7 and 7.65, respectively. Medical articles (normal and simple) seem to be longer and more complex (in terms of sentence length).

To align sentences from Wikipedia to Simple Wikipedia, we employ a two step approach: first we setup *candidate pairs*, by combining each sentence from the English Wikipedia articles which each sentence of the corresponding Simple Wikipedia article. This gives us a total of 3,660,064 candidate pairs from the 1,098 articles. Secondly, we select the most similar pairs from the candidate pairs. Kajiwara and Komachi employ pre-trained Word2Vec word embeddings to determine sentence similarity. Similarly, Hwang et al. present a method that relies on Wiktionary [29]. 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. In our alignment method we use a simple metric, the sentence BLEU score, in the configuration described above, to align sentences.

We only include fully aligned sentence pairs, where BLEU alignment shows good performance in the evaluation. In the generation of the dataset, we include sentence pairs with a BLEU score above a treshold of 0.29, which provided the maximum F_1 score during evaluation on the general domain set. After filtering out sentences with MediaWiki mathematical formulas, we include 3,797 fully aligned medical sentences. Table 4.2 shows example pairs with different BLEU scores.

Wikipedia	**Simple Wikipedia**	**BLEU**
Aspirin is an appropriate immediate treatment for a suspected MI.	Aspirin is an early and important treatment for a heart attack.	0.33
Many cases of croup have been prevented by immunization for influenza and diphtheria.	Immunization (vaccines) for influenza and diphtheria can prevent croup.	0.43
Hypertension (HTN or HT), also known as high blood pressure (HBP), is a long-term medical condition in which the blood pressure in the arteries is persistently elevated.	Hypertension or high blood pressure, is a chronic medical condition in which the blood pressure in the arteries is higher than it should be.	0.52
In the case of psychotic patients the premier cause of auditory hallucinations is schizophrenia.	The main cause of auditory hallucinations in psychotic patients is schizophrenia.	0.63
All three types can be diagnosed by seeing the parasites under the microscope.	All three types can be diagnosed by identifying the parasites under a microscope.	0.72
Uncomplicated infections can be diagnosed and treated based on symptoms alone.	Simple infections can be diagnosed and treated based on symptoms alone.	0.83
There are no effective antiviral drugs for the common cold even though some preliminary research has shown benefits.	There are no effective antiviral drugs for the common cold, even though some preliminary research has shown benefit.	0.93

Table 4.2: Example alignments using BLEU alignment

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

4.3. Overview

In sum, we introduced two new medical text simplification datasets: expert and $\mathrm{automatic}$. Our expert dataset is derived from [29]. If we exclude expert from [29], we end up with non-health related aligned sentences, which we consider as the $\mathrm{general}$ set of aligned sentences. The $\mathrm{automatic}$ dataset contains our additional alignments of medical sentences from Wikipedia and Simple Wikipedia disease articles. An overview is given in table 4.3.

Servan et al. showed that regular NMT already benefits from 500 domain-specific aligned sentences [67]. Training on 5000 domain-specific aligned sentences showed an improvement of 10.5% in BLEU points, a traditional metric for machine translation. However, note that domain specialization was not investigated before for NMT in the context of text simplification.

Fully aligned	N
$\mathrm{general}$	152,538
expert	2,267
$\mathrm{automatic}$	3,797
Partially aligned	
$\mathrm{general}$	126,785
expert	3,148

Table 4.3: An overview of the datasets

5

Neural Medical 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 evaluated properly in the medical domain yet. Therefore, we replicate the state-of-the-art NMT text simplification system of [53] and evaluate it on our expert-curated dataset. This system outperformed phrase-based [79] and syntax-based statistical machine translation [81] approaches, as well as an unsupervised lexical simplification approach [24].

We train a system on the $\mathrm{general}$ set only and call it $\mathrm{nmt\text{-}general}$. To investigate the effect of the health related aligned sentences, we train a second system $\mathrm{nmt\text{-}medical}$ on $\mathrm{general}$, expert and $\mathrm{automatic}$. Note that because we train on both fully and partially aligned sentences, we expect that the NMT learns to delete irrelevant parts of sentences and keep the core meaning only. Furthermore, because we have a relatively small amount of aligned health sentences and medical concepts are therefore sparse, we propose two strategies for translating medical concepts. They can be used (independently) in combination with NMT.

1. We introduce a method that groups semantically similar medical concepts so medical concepts will become less sparse

2. We use the Open-Access and Collaborative Consumer Health Vocabulary (CHV) for replacing professional medical terms with laymen terms

Recall that the CHV is a vocabulary for health consumers, which contains laymen synonyms for medical concepts, mined from medical forums and Q&A systems [12, 14]. First, we describe the NMT system itself. Next, we describe the two strategies for simplifying medical concepts.

5.1. Neural Text Simplification

We implemented the NMT system in OpenNMT[1], 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. An NMT system learns embeddings of source and target words (and may use pre-trained embeddings as a starting point). A source and target vocabulary size v is chosen for which embeddings are learned and retrieved. The most frequent v unique words are in the vocabularies. Note that an NMT system can only learn translations from words that are in the source vocabulary to words that are in the target vocabulary. The vocabulary size is pruned to 50,000 in both the source and target language. Word embedding size is set to 500. $\mathrm{nmt\text{-}baseline}$ uses pre-trained Word2Vec embeddings from the Google News corpus [49] of size 300, while $\mathrm{nmt\text{-}medical}$ uses pre-trained Word2Vec embeddings of size 200, trained on 10,876,004 English abstracts of biomedical articles from PubMed [36]. The remaining part (of the total size of 500) is learned during training of the NMT (while the pre-trained part remains fixed). Lastly, the decoder uses global attention with input feeding [44]. The system is trained for 20 epochs, using a SGD optimizer and an initial learning rate of 1.0. Then we select the best model based on

[1] http://opennmt.net/

perplexity on the validation set. After epoch 9, the learning rate decay is 0.7, i.e. `learning_rate = learning_rate * learning_rate_decay`.

Simplification is learned on sentence level on a word basis. The input is a complex sentence and the output is a simple sentence. The input is tokenized into tokens (or words). Since the output is also a sequence (i.e. the order of words matters) we use 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. For each system we will evaluate both hypotheses 1 and 2 and a beam size of 12, since the system that performed most changes and highest percentage of correct ones in [53] used a beam size of 12 as well. Moreover, we found that hypothesis 1 was often too conservative, i.e. copying the input too much, while hypothesis 2 performed more changes.

5.2. Translating Medical Concepts

In this section we describe how we combined two strategies for medical concept translation with NMT. The relatively small amount of aligned health sentences makes medical concepts sparse and it is therefore hard for the NMT to learn simplifications for them. The two strategies are used on top of nmt-medical. We call the resulting system nmt-medical-cui-chv, which will become clear in the next sections.

5.2.1. Grouping Semantically Similar Medical Concepts

As an extra pre-processing step before training (and translating with) nmt-medical we replace each medical concept encountered in the complex text in expert and automatic with a Concept Unique Identifier (CUI) from the Unified Medical Language System (UMLS). We group medical concepts that have the same CUI to a single token. Recall that the UMLS is a metathesaurus, which contains unified entities from different source vocabularies and terminologies. We used QuickUMLS [71] with a similarity threshold of 0.7, a value for which highest F1-scores were achieved in [71], to detect medical concepts and link them to a CUI. QuickUMLS measures string similarity between words from the input and instances from the UMLS. An n-gram with a similarity higher than the threshold is considered a medical concept. For each detected concept we take the best match's CUI according to QuickUMLS and replace the concept with this CUI. An example is given below (Wikipedia sentence followed by the same sentence with medical concepts replaced by their CUI):

- 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.
- C1956346 (CAD) also known as C0010054, C0010054, or C0151744 (C0151744), is the most common type of C0018799 and cause of C0027051.

This approach reduces the (medical) vocabulary (and medical concept sparsity), since any textual variation of a concept is mapped, normalized or grouped to a single CUI. This way all variations are reduced to the same single token and also the references for each concept are now aggregated. For example, *atherosclerotic heart disease* and *coronary heart disease* are both replaced with C0010054. Note that a valid laymen translation can be (for both) *heart disease*.

5.2.2. Adding a Dictionary Approach for Out of Vocabulary Medical Concepts

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 (OOV) with their CHV-preferred term, if it exists, or copy the original source token. Remember that QuickUMLS assigns a CUI from the UMLS and that the CHV is part of the UMLS. Each CHV term has a CUI. This way we can get the corresponding CHV-preferred term based on the CUI. An example is given below:

He suffered from a $\{\text{myocardial infarction}\}_{\text{C0027051}}$.

The CUI C0027051 is assigned to *myocardial infarction* by QuickUMLS (because *myocardial infarction* is in the UMLS). Then, in the MRCONSO table in the UMLS we can get the corresponding CHV term. The MRCONSO table contains terms with their source vocabulary linked to their CUI. A simplified view of this table is given in table 5.1.

CUI	**VOCAB**	**STR**
C0027051	...	...
C0027051	MSH	myocardial infarction
C0027051	...	...
C0027051	CHV	heart attack
C0027051	...	...

Table 5.1: A simplified view of the MRCONSO table in the UMLS.

To translate OOV CUIs, 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. In the example above the pair C0027051 and *heart attack* is added to this phrase-table. 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. If C0027051 is OOV then it is replaced by *heart attack* from the phrase-table. This way the output does not contain any raw CUI.

6

Exploratory Evaluation

Text simplification is typically automatically evaluated and evaluated by humans. Automatic evaluation is based on a reference test-set drawn from our expert dataset. Note that Wikipedia sentences are the input sentences and Simple Wikipedia sentences are the references. We automatically evaluate each proposed system with hypothesis 1 and 2 selection strategies (the most likely simplification and the second most likely one), as well as the identity function (just copying the input). Results of automatic evaluation are used to get a first impression of the performance of the systems. Humans evaluated simplified sentences on grammar correctness, meaning preservation from the complex sentence and simplicity compared with the complex sentence. Again, we evaluate each proposed system with hypothesis 1 and 2 selection strategies, but in this case, also Simple Wikipedia, since we do not need a reference for human evaluation. We do need Wikipedia as input, to measure how much meaning is preserved and how much simpler the output is. An overview of systems we evaluate is given in table 6.1. We randomly select 350 sentences as test set and 500 as validation set (to select to best model) from the expert dataset. Automatic evaluation is done on the test set. Human evaluation is done in the first 70 sentences of the test set (since human evaluation is rather costly).

System	Automatic evaluation	Human evaluation
Identity (Wikipedia)	✓	✗
Simple Wikipedia	✗	✓
$\mathrm{nmt\text{-}general}$	✓	✓
$\mathrm{nmt\text{-}medical\text{-}cui\text{-}chv}$	✓	✓

Table 6.1: An overview of systems we evaluate

Next, we describe how automatic and human evaluation are executed, with their corresponding metrics.

6.1. Automatic Evaluation

Text simplification is typically automatically evaluated using a traditional machine translation metric BLEU [58] and a text simplification specific metric SARI [81]. 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 [81], 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” [81], by comparing the output with the

Grammar	
5	The output is meaningful and there are no grammatical mistakes
4	The output contains one or two minor errors, but the meaning can be easily understood
3	The output contains several errors, but it is still possible to understand the meaning
2	The meaning of the output is hard to understand due to many grammar errors
1	The meaning of the output is impossible to understand due to being ungrammatical
Meaning preservation	
5	The output has the same meaning as the original sentence (it is allowed that it misses some irrelevant parts of the original sentence)
4	The core meaning of the output is the same as the original sentence, but with subtle differences
3	The output contains part of the relevant information of the original sentence, but misses another part of relevant information of the original sentence
2	The output has very different meaning than the original sentence
1	It is impossible to compare
Simplicity	
2	The output is syntactically and lexically simpler than the original sentence
1	The output is easier to understand due to minor changes, e.g. word replacement
0	The output is as difficult as the original sentence
-1	The output is harder to understand than the original sentence due to incorrect minor changes
-2	The output is impossible to understand

Table 6.2: Guidelines for in-house annotation

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.

In sum, BLEU score is an indication how close a translation is to the reference, while SARI measures the correctness of changes (adding and deleting) and unchanged phrases (what phrases were kept unchanged correctly) when compared to the original input and the reference. The results will give an indication of how well systems are able to simplify health sentences on Wikipedia with Simple Wikipedia as a reference.

6.2. Human Evaluation

As both metrics used in the automatic evaluation are insufficient to fully describe the capabilities of machine simplification, such evaluation needs 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 [53]. An evaluator is presented with a sentence pair (complex, simple) and asked to give the scores.

We asked three in-house non-medical people from myTomorrows to score these sentences on a simple web page, illustrated in figure 6.1. We based our annotation guidelines on [72]. 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. The guidelines are outlined in table 6.2. Note that an exact copy of the input gets a score of 5, 5 and 0 for grammar, meaning preservation and simplicity respectively.

6.3. Expectations

The nmt-general system should be able to simplify a complex health sentence in general. That is, delete irrelevant content and keep the core meaning of the sentence only (which is learned from partial alignments). This should be reflected by SARI's delete component. The nmt-medical-cui-chv system

should be able to add more (domain-specific) terms that are also in the reference, since it can translate more domain-specific terms due to the additional in-domain training data. Moreover, the translation of medical concepts to laymen terms is boosted by the semantic grouping of medical concepts and the CHV. Depending on how much content is deleted BLEU score should increase compared the just copying the input.

Human evaluation firstly answers the question how much simpler Simple Wikipedia is than Wikipedia, which is helpful, since we used it as a reference in automatic evaluation. It also answers how much meaning the automatically aligned Simple Wikipedia sentences preserve, compared to the Wikipedia sentences. Next, we compare how close the proposed systems come to Simple Wikipedia. The nmt-general system should be able to simplify a complex health sentence in general (with a certain loss of meaning preservation). The nmt-medical-cui-chv should improve simplicity by, in addition to general simplifications, translating medical terms to laymen terms.

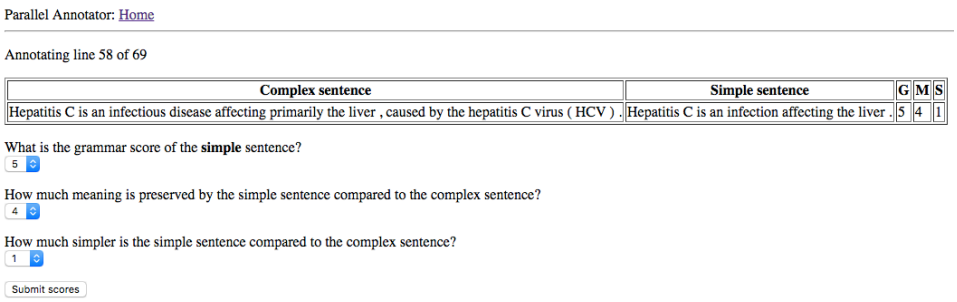


Figure 6.1: The annotation web page for in-house myTomorrows laymen

7

Results and Discussion I

This chapter gives an overview of and discusses automatic and human evaluation of the proposed systems:

- nmt-general
- nmt-medical-cui-chv

To briefly sum up, nmt-general is an NMT system trained on general Wikipedia and aligned Simple Wikipedia sentences. nmt-medical-cui-chv is in addition trained on domain-specific data. Also, before training (and translating), semantically similar medical concepts are grouped. For medical concepts that are not in the vocabulary (and are therefore not translated) we use the Open-Access and Collaborative Consumer Health Vocabulary (CHV). This is a vocabulary for health consumers, which contains laymen synonyms for medical concepts, mined from medical forums and Q&A systems [12, 14]. Note that we include Simple Wikipedia in our human evaluation.

7.1. Automatic Evaluation

In table 7.1, 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 nmt-general (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 phrases with higher precision than hypothesis 1. Both hypotheses of nmt-medical-cui-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) nmt-medical-chv is translating. BLEU scores of the identity and hypothesis 1 nmt-general are highest. This may be due to that in hypothesis 1 nmt-general 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. nmt-medical-cui-chv 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.

7.2. Human Evaluation

Three laymen provided feedback on the first 70 sentences of the test set with respect to grammar, meaning preservation, and simplicity, as is described in chapter 6.

Table 7.2 shows that nmt-general produces decent grammar and meaning preservation scores and indeed simplifies the text. However, nmt-medical-cui-chv scores show that grammar, meaning preservation and simplicity scores are all lower than nmt-general. We assume that this is due to nmt-medical-cui-chv replacing out of vocabulary concepts with their CHV-preferred terms (which are expert curated

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

Table 7.1: Evaluations with automatic metrics

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” [62].

Approach	G	M	S
Simple Wikipedia	4.91	4.24	0.53
nmt-general, h-1	4.85	4.30	0.22
nmt-general, h-2	4.49	3.87	0.23
nmt-medical-cui-chv, h-1	4.23	3.82	-0.05
nmt-medical-cui-chv, h-2	4.19	3.76	-0.05

Table 7.2: Human evaluation scores. G:Grammar, M:Meaning preservation, S:Simplicity

7.3. Example Translations

Two (random) example translations are given in table 7.3 and table 7.4. The first example shows that all systems are able to simplify the input and produce a grammatical correct sentence. The second example is more interesting, since it shows two things:

1. It shows that the automatic alignment of Wikipedia and Simple Wikipedia is not always perfect, from which meaning preservation can clearly suffer.
2. It shows that inserting a CHV-preferred term (*cell count* instead of *Cells*) decreases grammatical correctness and does not make the sentence simpler, while nmt-general does not suffer from this.

7.4. Discussion

While the results were unexpected, it is unable to identify what caused the deterioration of performance of nmt-medical-cui-chv compared to nmt-general. This is due to the fact that we did a number of things at the same time:

1. Adding domain-specific data
2. Replacing medical concepts with CUIs
3. Using the CHV to replace medical terms with laymen terms

Again, each component above should intuitively boost performance, but this was not the case. To measure the effect of each step we need to evaluate each step separately. We want to measure the effect of domain-specific training data. We also want to know the effect of grouping semantically similar medical concepts and using the CHV individually. Therefore, we add a second evaluation of the following systems:

1. nmt-medical - NMT trained on general, expert and automatic, i.e. we only add domain-specific data with respect to nmt-general

Source	Sentence	G	M	S
Wikipedia	On the second stage, as abnormal new blood vessels (neovascularisation) form at the back of the eye as a part of proliferative diabetic retinopathy (PDR), they can burst and bleed (vitreous hemorrhage) and blur vision, because the new blood vessels are weak.	-	-	-
Simple Wikipedia	As new blood vessels form at the back of the eye as a part of proliferative diabetic retinopathy (PDR), they can bleed (ocular hemorrhage) and blur vision.	4.67	3.00	1.00
nmt-general, h-1	The new blood vessels (neovascularisation) form at the back of the eye as a part of proliferative diabetic retinopathy (PDR).	5.00	2.67	1.00
nmt-general, h-2	On the second stage, the new blood vessels (neovascularisation) form at the back of the eye as a part of proliferative diabetic retinopathy (PDR).	4.67	3.33	0.67
nmt-medical-cui-chv, h-1	They can burst and bleed.	3.67	2.00	0.33
nmt-medical-cui-chv, h-2	In the second stage, the blood vessels are weak.	5.00	2.33	0.33

Table 7.3: Example translations from different systems with their scores. G:Grammar, M:Meaning preservation, S:Simplicity

2. nmt-medical-cui - The same as nmt-medical, but with grouping of semantically similar medical concepts in the complex text during training (and translation)
3. nmt-medical-chv - The same as nmt-medical, but with untranslated medical terms replaced by CHV-preferred terms

This way we can better measure the effect of the separate steps we took specifically for the medical domain. However, this does not explain the deterioration of performance. Since it is very reasonable to assume that domain-specific training data does not deteriorate performance we rule this out.

It is also reasonable that, assuming we grouped medical concepts correctly and replaced medical concepts with the *correct* CUI, CUI replacement should not really affect the NMT. Because for the NMT it does not matter whether it sees *coronary artery disease* or C1956346 or x, except that it becomes one token instead of three. In addition, recall that CUI replacement should group the same medical concepts in different terms, boosting the learning of medical concept translation to Simple Wikipedia terms. However, we might have assigned the wrong CUI to a medical concept, which impacts CUI translation learning and certainly CHV-preferred term replacement.

Source	Sentence	G	M	S
Wikipedia	Cells found in a ganglion are called ganglion cells, though this term is also sometimes used to refer specifically to retinal ganglion cells.	-	-	-
Simple Wikipedia	In another usage, ganglion cells are found in the retina of the vertebrate eye.	3.33	2.33	0.33
nmt-general, h-1	These cells are called ganglion cells.	4.67	3.00	1.67
nmt-general, h-2	This is called ganglion blood cells.	3.00	2.33	0.00
nmt-medical-cui-chv, h-1	cell count found in a ganglion are called ganglion cells, though this term is also sometimes used to refer specifically to retinal ganglion.	3.00	3.33	-1.00
nmt-medical-cui-chv, h-2	cell count found in a ganglion are called ganglion cells.	2.67	2.67	-0.67

Table 7.4: Example translations from different systems with their scores. G:Grammar, M:Meaning preservation, S:Simplicity.

Each medical concept detected by QuickUMLS with a similarity higher than a threshold of 0.7 is replaced by its CUI from the UMLS. However, it may be the case that a term has multiple matches in the UMLS and hence it could have multiple CUIs. A medical concept can have a different meaning in context a than in context b. Take for example the medical abbreviation *HT*, which can mean *hypertension* or *hyperthyroidism*. The UMLS might contain two instances of *HT* (for *hypertension* and *hyperthyroidism*) each with its own CUI. We simply took the best match (with highest similarity). In the case of a tie (i.e. matches with the same similarity, *HT* in this toy example) we randomly take one, which obviously is not always correct. To select the correct one, one should include a word sense disambiguation stage. A wrong CUI replacement in itself does not really effect the NMT, because it learns an embedding for a term (whatever it may be). However, when dealing with multiple occurrences of a CUI, a wrong CUI in a group of correct CUIs could indeed influence the embedding learning. Moreover, if selected a wrong CUI we replace it with the wrong laymen term from the CHV. We assume that this is the reason that the performance of $\mathrm{nmt\text{-}medical\text{-}cui\text{-}chv}$ deteriorated. In the next chapter we therefore revise CUI replacement.

8

Grouping Semantically Similar Medical Concepts Revisited

In the previous chapter we argued that we might have assigned the wrong Concept Unique Identifier (CUI) from the Unified Medical Language System (UMLS) to medical concepts. A medical concept can have multiple meanings, depending on the context around it. Continuing with the example from previous chapter, *HT* can mean *hypertension* or *hyperthyroidism*. Each of them may have an instance *HT* in the UMLS, with a different CUI. QuickUMLS returns all matches with a similarity higher than a certain threshold. Previously, we simply took the best match. In the case of a tie (multiple explanations with the same similarity) we randomly took one. In this chapter we describe a medical concept disambiguation method.

8.1. Disambiguation

QuickUMLS simply uses a string similarity between n-grams from the input sentence and UMLS instances and returns the UMLS instances that have a similarity higher than a certain threshold. If QuickUMLS returns exactly one instance, there is no need for disambiguation. If it returns multiple, we make use of two aligned sentences. For example, a sentence a and its simplification b. Sentence a might only mention *HT*, while in the simplified sentence b *HT* is translated to *hypertension*. Then we know that *HT* in sentence a is *hypertension* and not *hyperthyroidism*. The toy example below illustrates this:

1. The patient was suffering from *HT* during his surgery.
2. The patient showed signs of *hypertension* during surgery.

Using the second aligned sentence, we now know that *HT* refers to *hypertension*. Note that we do not necessarily need to use an aligned sentence, we can also use the sentence itself. Consider the following (real world) example from Wikipedia:

- Hypertension (HTN or HT), also known as high blood pressure (HBP), is a long-term medical condition in which the blood pressure in the arteries is persistently elevated.

Because *hypertension* is mentioned in the same sentence *HT* probably refers to *hypertension*.

In sum, given a sentence a, in which we want to replace medical concepts with CUIs, take a second aligned sentence b (remember that it is possible that b is equal to a). For each concept $c_{a,i}$ in a get the set of explanations $e_{a,i}$ from QuickUMLS. For each concept $c_{b,j}$ in b get the set of explanations $e_{b,j}$ from QuickUMLS. Then, for each concept $c_{a,i}$, if the intersection between $e_{a,i}$ and $e_{b,j}$ for any j is exactly one, the intersection of size 1 is added to the list of candidates l. Note that it can be the case that an intersection of size 1 occurs for multiple values of j, e.g. *HT* may have an intersection of explanations of size 1 with *hypertension* **and** *HTN*. If there is exactly one unique explanation e in l then we assign

explanation e to concept $c_{a,i}$. Finally, concept $c_{a,i}$ in a is replaced with the CUI of explanation e. The procedure is outlined in algorithm 1.

Input: a, b
Output: a
concepts_a = quickumls(a);
concepts_b = quickumls(b);
foreach $c_{a,i} \in$ *concepts_a* **do**
  candidates = [];
  $e_{a,i}$ = get_explanations($c_{a,i}$);
  foreach $c_{b,j} \in$ *concepts_b* **do**
    $e_{b,j}$ = get_explanations($c_{b,j}$);
    intersection = $e_{a,i} \cap e_{b,j}$;
    if |*intersection*| == *1* **then**
      candidates.append(intersection);
    end
  end
  if |*unique(candidates)*| == *1* **then**
    a = replace(a, $c_{a,i}$, candidates.pop());
  end
end

Algorithm 1: Concept disambiguation in sentence a, using a second (aligned) sentence b

8.1.1. Training

Recall from 5.2.1 that before training we replace all medical concepts with a CUI in the complex sentences. Different than before, we also replace medical concepts found in the general set of sentences. We only replace a medical concept when we know we selected the correct CUI, as described above. We use Simple Wikipedia as a set of second sentences for disambiguation. We make use of the fact that our training data consists of aligned (complex and simple) sentences. The resulting system is called nmt-medical-cui.

8.1.2. Translation

Before translation, we also need to replace medical concepts with CUIs (since we trained nmt-medical-cui on CUIs), but we do not have the Simple Wikipedia aligned sentence. We can use automatically generated simplifications from the already trained NMT systems (nmt-general and nmt-medical). We can also use the complex sentence itself, i.e. the identity. To avoid that we introduce CUIs that are OOV, we simply only replace medical concepts with CUIs that are in vocabulary. Table 8.1 shows how many CUIs we were able to disambiguate using different sources (and how many of them were in vocabulary). During testing we chose to use the identity to disambiguate CUIs.

Source	Disambiguated CUIs	In vocabulary
Identity	708	621
nmt-general, h-1	703	619
nmt-general, h-2	701	617
nmt-medical, h-1	707	621
nmt-medical, h-2	705	618

Table 8.1: Number of disambiguated CUIs that are in vocabulary using different sources.

8.2. Post-Translation Dictionary Approach

Instead of constructing a phrase-table that we use during translation for out of vocabulary medical concepts (see 5.2.2), we replace medical concepts after translation. We replace medical concepts that are left after translation. We replace medical concepts in the output of nmt-medical with laymen terms from the Open-Access and Collaborative Consumer Health Vocabulary (CHV) using its CUI. Similar

to nmt-medical-cui, we use the original complex sentence (the identity) for disambiguation. Table 8.2 shows how many CHV terms we replaced in the output of nmt-medical. The resulting system is called nmt-medical-chv.

Source	Number of CHV-preferred terms inserted
nmt-medical, h-1	487
nmt-medical, h-2	421

Table 8.2: Number of CHV-preferred terms we inserted after translation with nmt-medical.

9

Amazon Mechanical Turk Evaluation

In the discussion of our exploratory evaluation we mentioned we needed a second evaluation. We evaluate the same nmt-general baseline, trained on general. We add an evaluation of nmt-medical. This systems only differs in what training data it is trained on. It is in addition trained on medical sentences. Next we evaluate nmt-medical-cui and nmt-medical-chv. These two systems aim at translating medical concepts to laymen terms. We evaluate these separately to measure the performance of both separately. In sum, we evaluate the following systems:

- (Automatically aligned) Simple Wikipedia
- nmt-general - State-of-the-art Neural Text Simplification trained on the general dataset
- nmt-medical - State-of-the-art Neural Text Simplification trained on the general, expert and automatic datasets
- nmt-medical-cui - nmt-medical, but with semantically similar medical concepts grouped and replaced by their Concept Unique Identifier (CUI) before training and translating
- nmt-medical-chv - nmt-medical, but with medical concepts in the output replaced by Open-Access and Collaborative Consumer Health Vocabulary (CHV) terms

We again let humans evaluate the output of all systems. We ask three Amazon Mechanical Turk workers to rate grammar of the complex sentence gc (1-5), grammar of the simple sentence gs (1-5) of the output, meaning preservation m (1-5) of the output from the input and how much simpler s (-2-2) the output is than the input on a Likert scale, similar to what we did before. A pair p is defined as the input (complex sentence) and the output of a system. Each sentence pair p_i is evaluated by three unique workers on Amazon Mechanical Turk. For each sentence pair p_i we take the average of the scores given by the three workers for each metric, resulting in gc_i, gs_i, m_i and s_i. Because we evaluate all systems on the same set of sentences, we can do a Paired Sample t Test on gc, gs, m and s to find out whether there is a significant difference between two systems. We randomly selected 70 sentences from the test set. Note that we evaluate hypothesis 1 and 2 (the most likely simplification and the second most likely one) for each system (except for Simple Wikipedia). This results in $(4 * 2 + 1) * 70 = 630$ sentences / tasks on Amazon Mechanical Turk. The sentences are published in one batch and the sentences are served in a random order. This way workers cannot know the origin of the sentence.

We made sure we made the task as simple as possible. Each metric has its own question with its own predefined answers (representing the Likert scale). The set of questions and answers (given a sentence pair) is shown in figure 9.1.

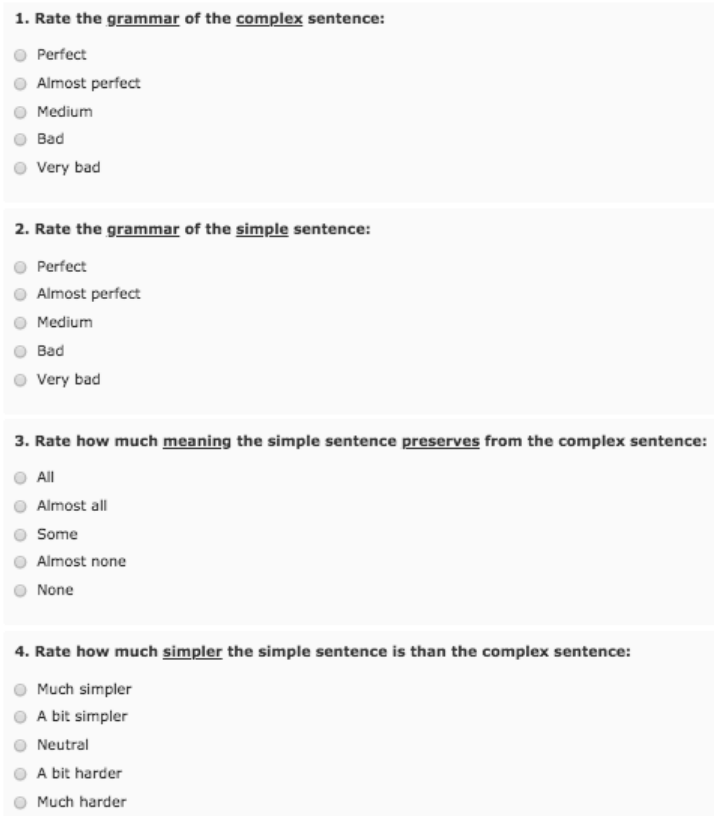


Figure 9.1: Set of questions and answers on Amazon Mechanical Turk

10

Results and Discussion II

10.1. Experimental Setup

In order to test to what extent we can use automated methods to simplify expert level health text to laymen level (main research question, see chapter 1) we evaluated four methods (chapter 9):

1. nmt-general - State-of-the-art Neural Text Simplification trained on the general dataset
2. nmt-medical - State-of-the-art Neural Text Simplification trained on the general, expert and automatic datasets
3. nmt-medical-cui - nmt-medical, but with semantically similar medical concepts grouped and replaced by their Concept Unique Identifier (CUI) before training and translating
4. nmt-medical-chv - nmt-medical, but with medical concepts in the output replaced by Open-Access and Collaborative Consumer Health Vocabulary (CHV) terms

In addition we evaluated Simple Wikipedia, to measure how close automatic simplifications are to manual simplifications. In this section we describe a setup for answering a number of sub-questions, which ultimately enables us to answer the main research question.

1. *What is the effect of domain-specific training data on medical text simplification when using the State-of-the-art in Neural Text Simplification?*
   We compare performance of nmt-general to performance of nmt-medical. These systems only differ in what training data is used.
2. *What is the effect of grouping semantically similar medical concepts to boost medical concept simplification on medical text simplification when using the State-of-the-art in Neural Text Simplification?*
   We compare performance of nmt-medical to nmt-medical-cui and nmt-general to nmt-medical-cui. Before training and translating with nmt-medical-cui, medical concepts are semantically grouped, which boosts medical concept simplification.
3. *What is the effect of using a post-translation dictionary approach for medical concept simplification using the Open-Access and Collaborative Consumer Health Vocabulary (CHV) on medical text simplification when using the State-of-the-art in Neural Text Simplification?*
   We compare performance of nmt-medical to nmt-medical-chv and nmt-general to nmt-medical-chv. The only difference between nmt-medical and nmt-medical-chv is that nmt-medical-chv does an extra step after translation. After translation medical concepts that are still in the output are replaced by laymen synonyms from the CHV.
4. *How close can our automated methods for medical text simplification get to manually simplified medical sentences from Simple Wikipedia?*
   We compare each system's performance against performance of Simple Wikipedia.

As explained in the previous chapter, we measure performance by asking Amazon Mechanical Turk workers to rate complex and simple sentence pairs on:

- the grammar of the input and output (1-5 Likert scale)
- how much meaning the output preserves from the input (1-5 Likert scale)
- how much simpler the output is than the input (-2-2 Likert scale)

In addition, we report three metrics that help interpret ratings from Amazon Mechanical Turk workers:

- The average length in words of output sentences
- The total number of translated medical concepts
  A medical concept is considered translated if it has an overlap of exactly one CUI with a medical concept in the source sentence and their terms are different.
- The total number of exact copied sentences from the input

We created a test set of 350 sentences, which is a random subset of fully aligned sentence pairs from the general dataset. We randomly selected 70 sentences from the test set for evaluation on Amazon Mechanical Turk. We run each system on the same 70 sentences, so that we can use a Paired Sample t Test to measure whether differences between two systems are significant. For each system we consider the two most likely outputs (h-1 and h-2). We compare h-1 of system a to h-1 of system b and h-2 of system a to h-2 of system b.

10.2. Results and Discussion

10.2.1. Effect of Domain-Specific Training Data

In this subsection we compare performance of nmt-general to performance of nmt-medical to answer the sub-question: *What is the effect of domain-specific training data on medical text simplification when using the State-of-the-art in Neural Text Simplification?*

We expect that training on domain-specific data, in addition to general data, will increase meaning preservation and simplicity scores and will not effect grammar. The NMT trained on general data can mostly delete parts of sentences to make a sentence simpler, since it does not know how to simplify most (domain-specific) terms. The NMT trained on medical data in addition to general data has more medical terms in its vocabularies (see table 10.1) and is therefore able to translate them (rather than deleting), arguably boosting meaning preservation and simplicity. This translates into the following hypotheses:

- Grammar:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d \neq 0$ (two-tailed)
- Meaning preservation:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d > 0$ (one-tailed, meaning preservation scores of nmt-medical is higher)
- Simplicity:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d > 0$ (one-tailed, simplicity scores of nmt-medical is higher)

Results are given in table 10.2. We indeed cannot reject the null hypothesis for grammar for both h-1 and h-2. Meaning preservation of nmt-medical in h-1 is significantly higher than that of nmt-general, but not in h-2. Simplicity of nmt-medical is actually lower than that of nmt-general (the opposite effect

Approach	Source vocabulary	Target vocabulary
nmt-general	4462	4265
nmt-medical	4794	4658

Table 10.1: This table shows how many medical terms are in the source and target vocabularies of nmt-general and nmt-medical.

of what we expected). Therefore, for simplicity we did an additional two-tailed t Test. The difference in h-1 is significant and in h-2 not.

The difference in meaning preservation in h-1 can be explained by the fact that nmt-medical produces 256 out of 350 exact copies of the input versus 215 out of 350 exact copies of nmt-general. In h-2 there is no difference in meaning preservation, because in h-2 both systems are mostly deleting parts of sentences. The number of exact copies could also explain the unexpected difference in simplicity scores, i.e. copying the input does not make the sentence simpler. Moreover, although nmt-medical has more medical terms in its vocabularies (see table 10.1), it does not translate more medical terms. This may be due to the fact that a (valid) translation of a medical concept can be the medical concept itself. In that case it is not counted as a translation. It could be that Simple Wikipedia contains the same medical concepts as Wikipedia, i.e. there is no translation, which the nmt-medical has learned.

In sum, training on domain-specific content increases meaning preservation, but decreases simplicity in h-1. This is probably because nmt-medical produces a lot of exact copies. In h-2, both systems are mostly deleting and hence there is no difference in performance observed.

h-1	G	M	S	Words	Translated medical concepts	Exact copies
nmt-general	4.41	3.94	0.55	20.86	30	215
nmt-medical	4.55	4.49	0.28	23.10	34	256
μ_d	0.14	0.55**	-0.27*			
h-2						
nmt-general	4.31	3.81	0.61	18.61	34	45
nmt-medical	4.35	3.81	0.58	19.22	31	56
μ_d	0.04	0.00	-0.03			

Table 10.2: *$p = 0.0079$ (two-tailed), **$p < 0.0001$. Results of nmt-general vs. nmt-medical. G:Grammar of simple sentence, M:Meaning preservation, S:Simplicity.

10.2.2. Effect of Semantically Grouping Medical Concepts

In this subsection we compare performance of nmt-medical to nmt-medical-cui and nmt-general to nmt-medical-cui to answer the sub-question: *What is the effect of grouping semantically similar medical concepts to boost medical concept simplification on medical text simplification when using the State-of-the-art in Neural Text Simplification?*

nmt-medical vs. nmt-medical-cui

Compared to nmt-medical, nmt-medical-cui translates more medical concepts, since we grouped semantically similar medical concepts, which makes them less sparse. This may have a positive effect on simplicity. Therefore we have the hypotheses:

- Grammar:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d \neq 0$ (two-tailed)
- Meaning preservation:
  - $H_0 : \mu_d = 0$

  - $H_1 : \mu_d \neq 0$ (two-tailed)

- Simplicity:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d > 0$ (one-tailed, simplicity scores of nmt-medical-cui is higher)

Results of nmt-medical vs. nmt-medical-cui are given in table 10.3. There indeed is no difference in grammar for both h-1 and h-2 and in meaning preservation in h-2. In h-1 there is a significant difference in meaning preservation. However, this may be due to that in h-1 nmt-medical is copying 256 out of 350 test sentences vs. 177 out of 350 by nmt-medical-cui. Simplicity scores of nmt-medical-cui are not significantly higher than that of nmt-medical, although nmt-medical-cui translates more medical concepts. However, we did not assess the correctness and quality of these translations. It could be that these translations are only slightly different (e.g. from singular to plural, or a different spelling).

In sum, translation of medical concepts is indeed boosted in nmt-medical-cui. However, it does not change the performance of medical text simplification. In h-1 meaning preservation is significantly lower, but probably due to the extremely high number of exact copies of nmt-medical.

h-1	**G**	**M**	**S**	**Words**	**Translated medical concepts**	**Exact copies**
nmt-medical	4.55	4.49	0.28	23.10	34	256
nmt-medical-cui	4.45	4.07	0.41	21.86	92	177
μ_d	-0.10	-0.42*	0.13			
h-2						
nmt-medical	4.35	3.81	0.58	19.22	31	56
nmt-medical-cui	4.31	3.88	0.51	20.29	115	45
μ_d	-0.04	0.07	-0.07			

Table 10.3: *$p = 0.0006$. Results of nmt-medical vs. nmt-medical-cui. G:Grammar of simple sentence, M:Meaning preservation, S:Simplicity.

nmt-general vs. nmt-medical-cui

Compared to nmt-general, meaning preservation (due to the ability to translate domain-specific terms rather than deleting) and simplicity should be higher (due to translating medical terms to laymen terms). This translates into the following hypotheses:

- Grammar:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d \neq 0$ (two-tailed)
- Meaning preservation:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d > 0$ (one-tailed, meaning preservation scores of nmt-medical-cui is higher)
- Simplicity:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d > 0$ (one-tailed, simplicity scores of nmt-medical-cui is higher)

Results of nmt-general vs. nmt-medical-cui are given in table 10.4. There are no significant differences between any of the metrics in h-1 and h-2, while we expected that meaning preservation and simplicity of nmt-medical-cui were higher. However, the way of simplifying of nmt-general and nmt-medical-cui seems different. While nmt-general produces shorter sentences, nmt-medical-cui translates more medical terms. This may indicate that deleting parts of sentences equally simplifies a medical sentence as translating medical terms. Either way, we lose as much meaning in the process.

h-1	G	M	S	Words	Translated medical concepts	Exact copies
nmt-general	4.41	3.94	0.55	20.86	30	215
nmt-medical-cui	4.45	4.07	0.41	21.86	92	177
μ_d	0.04	0.13	-0.14			
h-2						
nmt-general	4.31	3.81	0.61	18.61	34	45
nmt-medical-cui	4.31	3.88	0.51	20.29	115	45
μ_d	0.00	0.07	-0.10			

Table 10.4: Results of nmt-general vs. nmt-medical-cui. G:Grammar of simple sentence, M:Meaning preservation, S:Simplicity.

10.2.3. Effect of Post-Translation Dictionary Approach

In this subsection we compare performance of nmt-medical to nmt-medical-chv and nmt-general to nmt-medical-chv to answer the sub-question: *What is the effect of using a post-translation dictionary approach for medical concept simplification using the Open-Access and Collaborative Consumer Health Vocabulary (CHV) on medical text simplification when using the State-of-the-art in Neural Text Simplification?*

nmt-medical vs. nmt-medical-chv

Recall that we replace medical concepts that are still in the output after translation with nmt-medical with laymen terms from the CHV and call the resulting system nmt-medical-chv.

Compared to nmt-medical only simplicity should be boosted, since the only thing we do is replacing medical concepts with laymen synonyms.

- Grammar:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d \neq 0$ (two-tailed)
- Meaning preservation:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d \neq 0$ (two-tailed)
- Simplicity:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d > 0$ (one-tailed, simplicity scores of nmt-medical-chv is higher)

Results of nmt-medical vs. nmt-medical-chv are given in table 10.5. Grammar in both h-1 and h-2 is significantly lower. Meaning preservation is significantly lower in h-1. Simplicity is also lower, while we expected it would be higher. We therefore ran an extra two-tailed t Test. The difference in h-1 is significant.

nmt-general vs. nmt-medical-chv

Compared to nmt-general meaning preservation (due to the ability to translate domain-specific terms rather than deleting) and simplicity (due to replacing medical concepts with laymen synonyms) should be boosted. This translates into the following hypotheses:

- Grammar:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d \neq 0$ (two-tailed)
- Meaning preservation:

h-1	G	M	S	Words	Translated medical concepts	Exact copies
nmt-medical	4.55	4.49	0.28	23.10	34	256
nmt-medical-chv	4.23	4.05	0.07	23.46	307	103
μ_d	-0.32***	-0.44****	-0.21*			
h-2						
nmt-medical	4.35	3.81	0.58	19.22	31	56
nmt-medical-chv	4.05	3.71	0.40	19.54	263	16
μ_d	-0.30**	-0.10	-0.18			

Table 10.5: *$p = 0.0055$ (two-tailed), **$p = 0.0067$, ***$p = 0.0008$, ****$p < 0.0001$. Results of nmt-medical vs. nmt-medical-chv. G:Grammar of simple sentence, M:Meaning preservation, S:Simplicity.

  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d > 0$ (one-tailed, meaning preservation scores of nmt-medical-chv is higher)
- Simplicity:
  - $H_0 : \mu_d = 0$
  - $H_1 : \mu_d > 0$ (one-tailed, simplicity scores of nmt-medical-chv is higher)

Results of nmt-general vs. nmt-medical-chv are given in table 10.6. Grammar is significantly lower in h-2. Meaning preservation is not different. In both h-1 and h-2 simplicity is significantly lower, while we expected it to be higher. We tested significance with a two-tailed t test.

Using the CHV after translation does not improve simplicity. In fact, simplicity scores go down (in both comparisons). Moreover, grammar scores of the output also go down (also in both comparisons). It may be because we are replacing non-medical terms with terms from the Open-Access and Collaborative Consumer Health Vocabulary (CHV) in some cases. Remember that we used QuickUMLS to recognize medical concepts. We set the string similarity threshold to 0.7, which may in some cases introduce false positives. This may hurt grammar and simplicity. In addition, it can also be that CHV terms are still too difficult for laymen [62].

h-1	G	M	S	Words	Translated medical concepts	Exact copies
nmt-general	4.41	3.94	0.55	20.86	30	215
nmt-medical-chv	4.23	4.05	0.07	23.46	307	103
μ_d	-0.18	0.11	-0.48***			
h-2						
nmt-general	4.31	3.81	0.61	18.61	34	45
nmt-medical-chv	4.05	3.71	0.40	19.54	263	16
μ_d	-0.26**	-0.10	-0.21*			

Table 10.6: *$p < 0.0472$ (two-tailed), **$p = 0.0182$, ***$p < 0.0001$ (two-tailed). Results of nmt-general vs. nmt-medical-chv. G:Grammar of simple sentence, M:Meaning preservation, S:Simplicity.

10.2.4. Comparison with Simple Wikipedia

We compare each system with (by human written) simplifications from Simple Wikipedia to answer the sub-question: *How close can our automated methods for medical text simplification get to manually simplified medical sentences from Simple Wikipedia?*

Keep in mind that sentences from Wikipedia and Simple Wikipedia are automatically aligned. The test set is a random subset of our expert curated dataset expert. It consists of fully aligned sentence pairs (Wikipedia and Simple Wikipedia) only. Note that sentence pairs in expert were selected based

on medical relevance and not on quality of alignment. Automatic alignment is not perfect ($p = 0.798$, $r = 0.599$, $F_1 = 0.685$) [29], which may have an effect on meaning preservation and arguably on simplicity.

We want to find out whether our systems can simplify medical text at the level of Simple Wikipedia. For each metric for Simple Wikipedia vs. all our systems (h-1 and h-2) we test the following hypotheses:

- $H_0 : \mu_d = 0$
- $H_1 : \mu_d \neq 0$ (two-tailed)

Results of Simple Wikipedia and the differences of all our systems compared with Simple Wikipedia are reported in table 10.7. We identify three systems that are in all three metrics not significantly different from Simple Wikipedia:

- nmt-general, h-1
- nmt-medical-cui, h-1
- nmt-medical, h-2

Earlier, we already saw that nmt-general is mostly deleting parts of sentences to simplify a sentence, while nmt-medical-cui is translating medical concepts (rather than mostly deleting). Also, systems in h-2 tend to delete more than in h-1. These three systems reach the level of Simple Wikipedia according to our metrics (but keep in mind that Simple Wikipedia is automatically aligned). Interestingly, the system that uses the CHV for replacing medical concepts with laymen synonyms does not reach this level and has the lowest scores of all in all metrics (except in h-1 for meaning preservation).

	G	M	S	Words	Translated medical concepts	Exact copies
Wikipedia				26.11		
Simple Wikipedia	4.54	3.90	0.44	21.25	53	58
h-1						
μ_d nmt-general	-0.13	0.04	0.11	20.86	30	215
μ_d nmt-medical	0.01	0.59****	-0.16	23.10	34	256
μ_d nmt-medical-cui	-0.09	0.17	-0.03	21.86	92	177
μ_d nmt-medical-chv	-0.31**	0.15	-0.37***	23.46	307	103
h-2						
μ_d nmt-general	-0.23*	-0.09	0.17	18.61	34	45
μ_d nmt-medical	-0.19	-0.09	0.14	19.22	31	56
μ_d nmt-medical-cui	-0.23*	-0.02	0.07	20.29	115	45
μ_d nmt-medical-chv	-0.49****	-0.19	-0.04	19.54	263	16

Table 10.7: $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$. Results of Simple Wikipedia compared with all systems. G:Grammar of simple sentence, M:Meaning preservation, S:Simplicity.

10.3. Example Translations

Below in table 10.8 and 10.9 we present two examples. Each example consists of the input (Wikipedia), Simple Wikipedia and the output of all systems (h-1 and h-2). The first example shows that the systems are able to produce simpler output by preserving key information in the sentence only. The workers indeed find that these sentences are simpler (a score of 1.33) and meaning preservation <= 4. There is however no difference between the systems' output. For example, the CHV term for *typhoid fever* is *typhoid fever*. Workers score a copy of the complex as not simpler (0 and -0.33).

In the second example nmt-general and nmt-medical more or less copy the input. nmt-medical-cui, h-1 simplifies *common cold* to just *cold*. nmt-medical-chv replaces *asymptomatic* (false positive) with *asymptomatic HIV infection*, incorrectly, which has an impact on simplicity, according to the workers.

Approach	Sentence	G_c	G_s	M	S
Wikipedia	However, at the age of 18, he predeceased his father when he died of typhoid fever.	-	-	-	-
Simple Wikipedia	However, he died before his father because of typhoid fever.	4.33	5.00	3.67	0.00
nmt-general, h-1	However, at the age of 18, he predeceased his father when he died of typhoid fever.	3.67	4.67	3.67	-0.33
nmt-general, h-2	He died of typhoid fever.	4.33	5.00	3.33	1.33
nmt-medical, h-1	However, at the age of 18, he predeceased his father when he died of typhoid fever.	5.00	5.00	5.00	0.00
nmt-medical, h-2	He died of typhoid fever.	4.00	5.00	4.00	1.33
nmt-medical-cui, h-1	He died of typhoid fever.	4.67	5.00	4.00	1.33
nmt-medical-cui, h-2	However, at the age of 18, he predeceased his father when he died of typhoid fever.	4.00	5.00	4.67	-0.33
nmt-medical-chv, h-1	However, at the age of 18, he predeceased his father when he died of typhoid fever.	4.67	4.67	5.00	0.00
nmt-medical-chv, h-2	He died of typhoid fever.	3.33	3.33	3.00	1.33

Table 10.8: Example translations from different systems with their scores. G_c:Grammar of the complex sentence, G_s:Grammar of the simple sentence, **M**:Meaning preservation, **S**:Simplicity.

10.4. Threats to Validity

There are a number of threats to validity. Starting with the data, we assume that Wikipedia contains complex language and Simple Wikipedia contains simple language. Especially in the medical domain this might be debatable, given the results of nmt-medical. The questions remain whether medical Wikipedia articles contain expert level health text and whether the corresponding Simple Wikipedia articles contain the laymen translation.

Next, we used Amazon Mechanical Turk workers to evaluate the output of our systems. There is always a risk that workers did not care for the task, but completed the task as fast as possible. Also, workers might misinterpreted the task. However, we quantitatively (table 10.7) and qualitatively (table 10.8 and 10.9) checked the scores and they seem to be valid. For example, the grammar of human written text (Simple Wikipedia) approaches 5 (maximum score) and Simple Wikipedia is indeed considered simpler. The scores in the examples are also more or less what one would expect. We did not include a statistical analysis to filter out particular workers or sentences that seem suspicious.

Lastly, results indicate that outputs are indeed simpler. But, we did not include an analysis of the effect of it. The goal is that medical texts become more *accessible* for laymen. The sentences are simpler than the original input, but are people now better at understanding the *medical* content? This is a question that our results cannot answer.

Approach	Sentence	G_c	G_s	M	S
Wikipedia	A number of the viruses that cause the common cold may also result in asymptomatic infections.	-	-	-	-
Simple Wikipedia	A number of the viruses that cause the common cold may also result in no symptoms.	5.00	5.00	4.67	0.67
nmt-general, h-1	A number of the viruses that cause the common cold may also result in asymptomatic infections.	5.00	5.00	5.00	0.00
nmt-general, h-2	A number of the virus that cause the common cold may also result in asymptomatic infections.	4.67	5.00	4.00	0.33
nmt-medical, h-1	A number of the viruses that cause the common cold may also result in asymptomatic infections.	5.00	5.00	5.00	0.00
nmt-medical, h-2	A number of viruses that cause the common cold may also result in asymptomatic infections.	4.67	4.67	4.33	0.33
nmt-medical-cui, h-1	A number of the viruses that cause the cold may also result in asymptomatic infections.	5.00	5.00	4.33	0.00
nmt-medical-cui, h-2	A number of the viruses that cause the common cold may also result in asymptomatic infections.	4.00	4.67	4.00	0.67
nmt-medical-chv, h-1	A number of the viruses that cause the common cold may also result in asymptomatic HIV infection infections.	4.33	4.33	4.33	-1.00
nmt-medical-chv, h-2	A number of viruses that cause the common cold may also result in asymptomatic HIV infection infections.	4.33	4.00	3.33	0.00

Table 10.9: Example translations from different systems with their scores. $\mathbf{G}_c$:Grammar of the complex sentence, $\mathbf{G}_s$:Grammar of the simple sentence, **M**:Meaning preservation, **S**:Simplicity.

11 Conclusions

In this thesis we focused on the following research question:

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

First, we identified the state-of-the-art in general text simplification. We showed that Neural Machine Translation (NMT) shows most promise in performing both lexical and syntactic simplification. It learns from parallel corpora, such as aligned Wikipedia and Simple Wikipedia and the Newsela corpus. Based on training data, it can (implicitly) learn to perform lexical simplification (replacing complex words with simpler words) and syntactic simplification (reorder and delete parts of sentences).

In the medical domain, research focused mainly on lexical simplification (replacing medical concepts with laymen synonyms and generating laymen explanations of medical concepts). While lexical simplification is widely acknowledged to be very important for comprehension of medical text (due to the vocabulary gap between health professionals and consumers), it is also acknowledged that lexical simplification alone is not sufficient for medical text simplification [80]. State-of-the-art in general text simplification, NMT, might actually be a suitable technique to do both. This is still relatively unexplored in the medical domain. A preliminary work [5] extracted a (fairly small not publicly available) medical text simplification dataset from web-based knowledge sources and trained an NMT system on it, but do not include human evaluation, a common practice in text simplification research. In this thesis we used NMT to learn text simplification from aligned Wikipedia and Simple Wikipedia, published our datasets for future research and included a proper human evaluation.

We evaluated NMT in two directions: data and medical terms. To investigate the effect of adding domain-specific training data, we created a medical set of aligned sentences for in-domain training and testing. We semi-automatically filtered health sentences from an existing Wikipedia - Simple Wikipedia aligned corpus, using the knowledge of an in-house domain expert. Additionally, we used a novel method for aligning additional complex and simple health sentences from Wikipedia and Simple Wikipedia disease articles. For medical term simplification, we introduced two strategies. One is boosting the NMT itself to learn translations of medical concepts to medical concepts on Simple Wikipedia (arguably laymen terms). The second is a dictionary approach using the Open-Access and Collaborative Consumer Health Vocabulary (CHV) that is used after translation. Untranslated medical concepts in the output are then replaced by terms from the CHV. We measured the effect of adding domain-specific training data when training an NMT and of the two medical term simplification strategies in combination with NMT.

Workers on Amazon Mechanical Turk evaluated the original complex sentence on grammar and the output on grammar, meaning preservation (from the complex sentence) and simplicity (compared to the complex sentence). Adding domain-specific data did not make the output text simpler, but caused the NMT to produce more exact copies. In fact, NMT trained on general data produced simpler output, due to producing the shortest sentences, but with the logical loss of meaning. The boosting of learning translations of medical concepts indeed boosted the number of medical concept translations. Output sentences are longer than NMT trained on general data, but of the same simplicity (and meaning

preservation). It could indicate that translating medical concepts is as beneficial as deleting parts of sentences and keeping the core meaning only. Replacing untranslated medical concepts with CHV-preferred terms also did not boost simplicity. It actually makes grammar, meaning preservation and simplicity worse. It may be because we were not strict enough in matching terms from the text with terms from the UMLS or that CHV-preferred terms are still too difficult for laymen.

In sum, to answer the research question, we can use the state-of-the-art in general text simplification (NMT) to simplify expert level health text. It can learn lexical and syntactic simplification implicitly from training data. An NMT system trained on simplifications from general Wikipedia and Simple Wikipedia articles produced the simplest output, but mainly due to making the sentence shorter. Adding domain-specific data, in combination with a boosting strategy for learning medical concept translation, indeed causes the NMT system to translate more medical concepts, but scores are similar. It does delete less content, but meaning preservation was not different though. An NMT system trained on general data only is already able to simplify expert level health text to the level of Simple Wikipedia. The questions remain whether the original input (Wikipedia) is actually at expert level and whether Simple Wikipedia (and the output of our systems) is at laymen level.

11.1. Recommendations

We concluded that an NMT system trained on general Wikipedia and Simple Wikipedia articles already produces simplifications similar to that of Simple Wikipedia. However, it mainly does so by deleting parts of sentences (and keeping the core meaning). It does not translate medical concepts. The medical NMT system with boosting strategy does translate medical concepts and also produces simplifications similar to that of Simple Wikipedia. The question remains whether deleting parts of sentences or translating medical concepts is more important. It seems that Simple Wikipedia does not translate a lot of medical concepts either, but does not delete as much as the general system does.

Depending on how much in-domain data one can gather, we recommend to use an NMT system trained on general Wikipedia articles only. Apparently, there is not much translation of professional medical concepts on Wikipedia to laymen synonyms on Simple Wikipedia. If for example, additional in-domain data can be gathered (ideally from which we know it is understandable by laypersons), e.g. layperson summaries of clinical trial results [23], we should always compare a system trained on this in-domain data to one trained on general data.

11.2. Future Work

Future work can continue in several directions. Obviously, future work may test the suitability of Wikipedia and Simple Wikipedia for (medical) text simplification. Research that answers questions like:

- How are non-medical Wikipedia articles simplified compared to medical Wikipedia articles?
- Is there a translation of medical concepts to laymen terms on Simple Wikipedia?

Alternatively, one could align other in-domain data, train a system on it and compare it to our existing systems.

Another direction is to continue with using an NMT for learning medical concept translations. With QuickUMLS we can annotate known medical concepts and link them to the UMLS. We can use an NMT to learn translations of those concepts, in e.g. Simple English, that do not yet exist in the UMLS. In other words, we can learn new laymen terms for known medical concepts. Conveniently, these laymen terms are then also immediately linked to a CUI in the UMLS.

At last there remains one open question: Are people after automatic simplification better at understanding the *medical* content? This was out of scope for this thesis. Future work may include an additional evaluation that answers this question.

Bibliography

[1] Omri Abend and Ari Rappoport. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 228–238, 2013.

[2] Emil Abrahamsson, Timothy Forni, Maria Skeppstedt, and Maria Kvist. 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)*, pages 57–65, 2014.

[3] Sallam Abualhaija and Karl-Heinz Zimmermann. D-bees: A novel method inspired by bee colony optimization for solving word sense disambiguation. *Swarm and Evolutionary Computation*, 27: 188–195, 2016.

[4] Sallam Abualhaija, Tristan Miller, Judith Eckle-Kohler, Iryna Gurevych, and Karl-Heinz Zimmermann. Metaheuristic approaches to lexical substitution and simplification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 870–880, 2017.

[5] Viraj Adduru, Sadid Hasan, Joey Liu, Yuan Ling, Vivek Datla, and Kathy Lee. 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*, 2018.

[6] Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 295–305, 2017.

[7] Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

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

[9] Joachim Bingel and Anders Søgaard. Text simplification as tree labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 337–343, 2016.

[10] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[11] Sophie Bostock and Andrew Steptoe. Association between low functional health literacy and mortality in older adults: longitudinal cohort study. *Bmj*, 344:e1602, 2012.

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

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

[14] Jinying Chen, Abhyuday N Jagannatha, Samah J Fodeh, and Hong Yu. 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.

[15] Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, and Hong Yu. 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):e26, 2018.

[16] James R Curran, Stephen Clark, and Johan Bos. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 33–36. Association for Computational Linguistics, 2007.

[17] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[18] Siobhan Devlin and Gary Unthank. Helping aphasic people process online information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 225–226. ACM, 2006.

[19] Daniel Ferrés, Montserrat Marimon, and Horacio Saggion. A web-based text simplification system for english. *Procesamiento del Lenguaje Natural*, (55), 2015.

[20] Daniel Ferrés, Montserrat Marimon, Horacio Saggion, et al. Yats: Yet another text simplifier. In *International Conference on Applications of Natural Language to Information Systems*, pages 335–342. Springer, 2016.

[21] Centers for Disease Control, Prevention, et al. *Plain Language Thesaurus for Health Communication*. Centers for Disease Control and Prevention, 2009.

[22] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, 2013.

[23] Claire L Gillow. Layperson summaries of clinical trial results: Useful resources in the vacuum of regulatory guidance. *Medical Writing*, 24(4):205–209, 2015.

[24] Goran Glavaš and Sanja Štajner. 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)*, volume 2, pages 63–68, 2015.

[25] Sadid A Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek Datla, Aaditya Prakash, and Oladimeji Farri. Neural clinical paraphrase generation with attention. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 42–53, 2016.

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

[27] Daniel Hershcovich, Omri Abend, and Ari Rappoport. A transition-based directed acyclic graph parser for ucca. *arXiv preprint arXiv:1704.00552*, 2017.

[28] Colby Horn, Cathryn Manduca, and David Kauchak. Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 458–463, 2014.

[29] William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 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*, pages 211–217, 2015.

[30] Ling Jiang and Christopher C Yang. 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*, pages 314–320. Springer, 2015.

[31] Tomoyuki Kajiwara and Mamoru Komachi. 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*, pages 1147–1158, 2016.

[32] David Kauchak. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, volume 1, pages 1537–1546, 2013.

[33] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[34] Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

[35] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, 2017.

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

[37] Poorna Kushalnagar, Scott Smith, Melinda Hopper, Claire Ryan, Micah Rinkevich, and Raja Kushalnagar. Making cancer health text on the internet easier to read for deaf people who use american sign language. *Journal of Cancer Education*, 33(1):134–140, 2018.

[38] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*, 2018.

[39] Franco Lancia. Word co-occurrence and theory of meaning. *Retrieved August*, 18:2007, 2005.

[40] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task*, pages 28–34. Association for Computational Linguistics, 2011.

[41] John Lee and J Buddhika K Pathirage Don. Splitting complex english sentences. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 50–55, 2017.

[42] Gondy Leroy and James E Endicott. Combining nlp with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754. ACM, 2012.

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

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

[45] Angrosh Annayappan Mandya and Advaith Siddharthan. Text simplification using synchronous dependency grammars: Generalising automatically harvested rules. In *8th International Natural Language Generation Conference*, 2014.

[46] Angrosh Annayappan Mandya, Tadashi Nomoto, and Advaith Siddharthan. Lexico-syntactic text simplification and compression with typed dependencies. In *25th International Conference on Computational Linguistics*, 2014.

[47] Aqueasha Martin-Hammond and Juan E Gilbert. Examining the effect of automated health explanations on older adults' attitudes toward medication information. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 186–193. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2016.

[48] Alexa T McCray, Suresh Srinivasan, and Allen C Browne. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 235. American Medical Informatics Association, 1994.

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

[50] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

[52] Shashi Narayan and Claire Gardent. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 435–445, 2014.

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

[54] Gustavo Paetzold and Lucia Specia. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, 2016.

[55] Gustavo Paetzold and Lucia Specia. Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 34–40, 2017.

[56] Gustavo H Paetzold and Lucia Specia. Unsupervised lexical simplification for non-native speakers. In *AAAI*, pages 3761–3767, 2016.

[57] Gustavo H Paetzold and Lucia Specia. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593, 2017.

[58] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[59] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[60] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics, 2006.

[61] Matt Post, Juri Ganitkevitch, Luke Orland, Jonathan Weese, Yuan Cao, and Chris Callison-Burch. Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 206–212, 2013.

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

[63] Gillian Rowlands, Joanne Protheroe, John Winkley, Marty Richardson, Paul T Seed, and Rima Rudd. A mismatch between population health literacy and the complexity of health information: an observational study. *Br J Gen Pract*, 65(635):e379–e386, 2015.

[64] Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. Musst: A multilingual syntactic simplification tool. *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28, 2017.

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

[66] Isabel Segura-Bedmar, Paloma Martínez, Ricardo Revert, and Julián Moreno-Schneider. Exploring spanish health social media for detecting drug effects. In *BMC medical informatics and decision making*, volume 15, page S6. BioMed Central, 2015.

[67] Christophe Servan, Josep Crego, and Jean Senellart. Domain specialization: a post-training domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06141*, 2016.

[68] Matthew Shardlow. The cw corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, 2013.

[69] Matthew Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70, 2014.

[70] Advaith Siddharthan and Angrosh Annayappan Mandya. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. Association for Computational Linguistics, 2014.

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

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

[73] Sanja Štajner, Hannah Béchara, and Horacio Saggion. A deeper exploration of the standard pb-smt approach to text simplification and its evaluation. 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)*, volume 2, pages 823–828, 2015.

[74] Elior Sulem, Omri Abend, and Ari Rappoport. 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)*, volume 1, pages 162–173, 2018.

[75] Kristina Toutanova and Christopher D Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics, 2000.

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

[77] Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Brian Zhu, and Iain J Marshall. Extracting pico sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, 17(132):1–25, 2016.

[78] World Health Organization (WHO et al. Health literacy. the solid facts. *Self*, 2018.

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

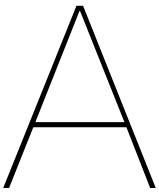
[80] Deborah X Xie, Ray Y Wang, and Sivakumar Chinnadurai. Readability of online patient education materials for velopharyngeal insufficiency. *International journal of pediatric otorhinolaryngology*, 104:113–119, 2018.

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

[82] Ming Yang and Melody Kiang. Extracting consumer health expressions of drug safety from web forum. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 2896–2905. IEEE, 2015.

[83] Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. Cwig3g2-complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 401–407, 2017.

[84] Jiaming Zhan, Han Tong Loh, and Ying Liu. Gather customer concerns from online product reviews–a text summarization approach. *Expert Systems with Applications*, 36(2):2107–2115, 2009.



WWW ’19 Accepted Short Paper

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

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
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 first semi-automatically filter 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 the 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.

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 domain, only few medical sentences are covered which motivates the creation of our EXPERT dataset. This initial corpus consists

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

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 the dataset to be much larger in size with slight losses in quality. In principle, such a pipeline has 2 distinct steps: (1) Collection of relevant articles and their related simplified version, (2) Splitting the articles in sentences and aligning them into pairs.

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.

Splitting and aligning Analogous to Kajiwara and Komachi, we extract the text from the Wikipedia and Simple Wikipedia articles, using the python Wikipedia API [2] and tokenize into sentences using NLTK 3.3 [3]. This gave 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 were 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 numbers of sentences per article were 57.7 and 7.65, respectively. Medical articles (simple and normal), seems to be longer and more complex (in terms of sentence length).

To align sentences from Wikipedia, to Simple Wikipedia, we employ a two step approach: as the first we setup *candidate pairs*, by combining each sentence from the normal articles which each sentence of the related simple 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. In order to do this, 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 clasifier to identify matching sentences. In our alignment task, we propose a simple metric, the BLEU score [23] to select matching sentences. 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 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] as evaluation set. 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 1 reports maximum F1-score and AUC for both methods in both scenarios. We observe that BLEU alignment performs on 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 not in the vocabulary of pre-trained Word2Vec models or not in Wiktionary, which may deteriorate

[2] https://pypi.org/project/wikipedia/

[3] http://www.nltk.org

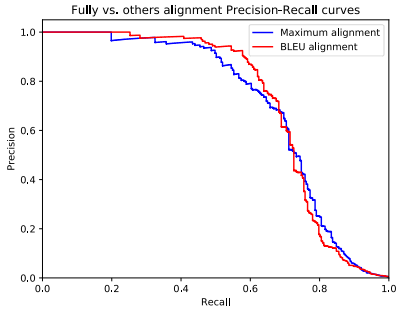


Figure 1: Precision-recall curves of Maximum alignment and BLEU alignment

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

performance of approaches that use such sources. BLEU only looks at overlapping n-grams, which makes it language independent.

In the AUTOMATED dataset, we only only include fully aligned sentence pairs, where BLEU alignment shows good performance in the evaluation. In the generation of the dataset, we include sentence pairs with a BLEU score above a treshold of 0.29, which provided the maximum F1 score during evaluation on the general domain set. After filtering out sentences with MediaWiki mathematical formulas, we include 3,797 fully aligned medical sentences. In table 2 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 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

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 2: Example alignments using BLEU alignment

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 pre-trained 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 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.

[4] http://opennmt.net/

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 4, 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 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.

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 3: 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 4: Evaluations with automatic metrics

Table 5 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 3. 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.

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.

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

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 et al. 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.