Benchmarking Automated Clinical Language Simplification: Dataset, Algorithm, and Evaluation

Junyu Luo[1], Junxian Lin[2,3], Chi Lin[2,3], Cao Xiao[4], Xinning Gui[1], Fenglong Ma[1*]

[1]College of Information Sciences and Technology, Pennsylvania State University, USA
[2]School of Software Technology, Dalian University of Technology, China
[3]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China
[4]Relativity, USA

{junyu, xinninggui, fenglong}@psu.edu, linjunxian@mail.dlut.edu.cn,
c.lin@dlut.edu.cn, cao.xiao@relativity.com

Abstract

Patients with low health literacy usually have difficulty understanding medical jargon and the complex structure of professional medical language. Although some studies are proposed to automatically translate expert language into layperson-understandable language, only a few of them focus on both accuracy and readability aspects simultaneously in the clinical domain. Thus, simplification of the clinical language is still a challenging task, but unfortunately, it is not yet fully addressed in previous work. To benchmark this task, we construct a new dataset named `MedLane` to support the development and evaluation of automated clinical language simplification approaches. Besides, we propose a new model called DECLARE that follows the human annotation procedure and achieves state-of-the-art performance compared with eight strong baselines. To fairly evaluate the performance, we also propose three specific evaluation metrics. Experimental results demonstrate the utility of the annotated `MedLane` dataset and the effectiveness of the proposed model DECLARE[1]

1 Introduction

Health literacy is generally defined as the ability of patients to obtain, process, understand, and communicate basic health information (Parker et al., 1999), which is significantly important for making health decisions and ensuring treatment outcomes. The increasing accessibility of technology information makes patients have more opportunities to access health information. However, it is challenging for patients, especially with limited health literacy, to read and understand health materials such as clinical notes written by doctors, with medical jargon (Korsch et al., 1968), abbreviations, and professional language (Friedman et al., 2002).

The lack of proper communication between doctors and patients not only results in a tense doctor-patient relationship (Ha and Longnecker, 2010) but also increases the risk of adverse health outcomes over time (Sudore et al., 2006). Therefore, there is a great need to *simplify professional clinical language to layperson-understandable language*.

1.1 Why We Need a New Dataset?

To implement an automated clinical language simplification system, the first step is to prepare the dataset for model training. Although there are several annotated medical datasets that are summarized in Table 1, most of them do not focus on the clinical domain. As we discussed before, the writing style of clinical materials is significantly different from that used in the general biomedical area. Thus, those datasets cannot be used to train a clinical simplification model.

To the best of our knowledge, there are only two datasets focusing on clinical language simplification tasks. The n2c2-track3 dataset (Henry et al., 2020) focuses on clinical term normalization, i.e., recovering the full-term expressions for those abbreviations and acronyms in clinical notes. This term-level recovery cannot guarantee the simplicity of the translated sentences because the full expressions may still be hard to be understood by patients with low levels of health literacy. For example, the full expression of the acronym "*NC*" is "*nasal cannula*", which is still a professional medical term instead of plain language.

In (Sakakini et al., 2020), the authors annotate a clinical language simplification dataset, but it is **private**. Besides, *the number of the annotated free-text parallel sentences/instances is only 1,541*, which is too small to be enough for evaluating the real performance of machine learning models, especially for deep learning-based models. Finally, this dataset only focuses on sentence-level simplification and *does not provide term-level annotation*,

---

*Corresponding author.
[1]The source code of DECLARE and the `MedLane` dataset can be found in the https://github.com/machinelearning4health/MedLane.

Text Source	Dataset Name	Accessible	Term Normalization			Sentence Simplification	
			Abbreviations	Acronyms	Complex Phrase	Style Transfer	Training Annotations
Biomedical Article	CLEF (Elhadad et al., 2013)	✓	✓	✓	-	-	-
	MSD (Cao et al., 2020)	✓	-	-	-	✓	-
	CDSR (Guo et al., 2021)	✓	-	-	-	✓	✓
	(Devaraj et al., 2021)	✓	-	-	-	✓	✓
Perscription	e-perscription (Zheng et al., 2021)	✓	-	-	-	✓	✓
Wiki	(Van den Bercken et al., 2019)	-	-	-	-	✓	✓
	AutoMeTS (Van et al., 2020)	✓	-	-	-	✓	✓
Clinical Note	n2c2-track3 (Henry et al., 2020)	✓	✓	✓	-	-	-
	(Sakakini et al., 2020)	-	-	-	-	✓	✓
	MedLane (Ours)	✓	✓	✓	✓	✓	✓

Table 1: Dataset comparison.

which leads to the difficulty of evaluating whether professional medical terms can be correctly translated and further decreases the reliability of machine learning models. Thus, it is essential to create a **publicly available, large-scale yet fully-annotated** dataset for the clinical language simplification task.

1.2 Why We Need a New Model?

Existing models for automatic text simplification (ATS) in the biomedical domain are mainly designed based on the available datasets, which either mainly focus on term normalization or directly apply neural machine translation techniques.

The term normalization technique aims to recover the full-term expressions for medical abbreviations and acronyms using a dictionary (Vydiswaran et al., 2014; Elhadad and Sutaria, 2007; Deléger and Zweigenbaum, 2008; Qenam et al., 2017; Rahimi et al., 2020; Liu et al., 2021), i.e., only targeting the *term-level simplification* and without considering the readability and understandability of the whole sentences. Another line of work treats the original sentences as the source language and the simplified sentences as the target language, i.e., the *sentence-level simplification*. They usually borrow the ideas from neural machine translation models to only learn the style mapping function between the original and simplified sentences (Weng et al., 2019; Pattisapu et al., 2020; Li et al., 2020; Devaraj et al., 2021) but ignore the term-level translation.

In fact, the drawbacks of existing studies make them impossible to be applied to the new annotated dataset. Therefore, we need to design a new model to *achieve term-level simplification and layperson-understandable sentence generation simultaneously*.

1.3 Our Contributions

• **Dataset**. We manually annotate a new Medical Language simplification dataset named MedLane,

which not only provides aligned sentence pairs but also offers term-level annotations. The dataset consists of 12,801 training samples, 1,015 validation samples, and 1,016 testing samples.

• **Approach**. Following the human annotation procedure, we design a novel end-to-end Dictionary-enhanced clinical language simplifier (shorten for DECLARE), which consists of three parts, including *a complex word locator, a neural lexical interpreter*, and *a restricted syntactic polisher*. The locator is in charge of automatically recognizing medical jargon and abbreviations in the input sentences. The neural lexical interpreter aims to replace the located terms with appropriate professional expressions selected from a predefined dictionary. Finally, a syntactic polisher is implemented to simplify the modified sentences by the interpreter and further increases the readability and understandability of original sentences, which should be significantly helpful for users with low health literacy.

• **Baseline**. We compare the proposed DECLARE against *nine baselines*, including a dictionary-based approach, a statistical machine translation approach, three neural machine translation approaches (i.e., Seq2Seq (Bahdanau et al., 2015) and its two variants), a modified text summarization model PointerNet (Vinyals et al., 2015), two state-of-the-art transformer-based pre-trained Seq2Seq models (i.e., T5 (Raffel et al., 2019) and BART (Lewis et al., 2019)), and PMBERT-MT that is built upon the pre-trained language model PubMedBERT (Gu et al., 2020), for validating the usability of the MedLane dataset and the effectiveness of our model. We also list EditNTS (Dong et al., 2019), an approach in the general ATS domain, as a baseline.

• **Evaluation**. Different from bilingual translation tasks, our task is to translate professional medical language to layperson-understandable language. We not only require the translated results to be readable but also accurate and easily understandable. Thus, we design three new yet general evaluation

metrics for the clinical simplification task. Besides, we still evaluate the results with commonly-used evaluation metrics, including BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and SARI (Xu et al., 2016).

2 Related Work

2.1 Medical Text Simplification Datasets

We summarize the widely-adopted clinical-related text simplification datasets and make a comparison with our proposed `MedLane` from different angles as shown in Table 1.

The first type of datasets focus on the normalization of abbreviations and acronyms by choosing proper explanations for them from a predefined dictionary, e.g., n2c2-track3 (Henry et al., 2020) and CLEF (Elhadad et al., 2013). The datasets are designed for a *classification* problem and do not include any sentence-level polishing and complex phrase translation. Thus, the readability of the simplified sentences may still be low even after the term-level normalization.

The second type of datasets, including MSD (Cao et al., 2020), CDSR (Guo et al., 2021), e-prescription (Zheng et al., 2021) and (Van den Bercken et al., 2019), focus on the style translation setting, which ignores the term-level simplification. The most similar dataset is the work (Sakakini et al., 2020). Except for missing the term-level annotation, it is not publicly available and only contains a small number of annotated free-text sentence pairs, which cannot be used by deep learning models.

2.2 Medical Text Simplification Method

Medical text simplification is a sub-task of automatic text simplification (ATS) (Laban et al., 2021; Dong et al., 2019; Stahlberg and Kumar, 2020; Paetzold and Specia, 2016, 2017), whose goal is to reduce the linguistic complexity of the original text to improve the readability. Besides increasing the readability, another target of the medical text simplification task is to accurately simplify professional medical terms. Existing approaches for medical text simplification can be roughly classified into two categories.

Dictionary-based approaches rely on using the expert-curated medical dictionaries to simplify the professional medical sentences (Kandula et al., 2010; Zeng and Tse, 2006; Zeng-Treitler et al., 2007; Chen et al., 2017; Lalor et al., 2019) or

link medical terms with lay definitions (Chen et al., 2018) and definitions in controlled vocabularies (Polepalli Ramesh et al., 2013). These approaches are highly reliable yet cannot manage the case of polysemant, i.e., and a term can have multiple correct explanations under different cases. To solve this issue, (Sakakini et al., 2020) utilizes a pre-trained language model to select the most possible answer and then replaces the selection with the located hard terms. This approach is an advanced version of the dictionary-based approach. However, this simple replacement may lead to a decrease in sentence readability, which further makes the simplified sentence still difficult to be understood by users or patients.

Researchers also try to borrow ideas from *machine translation*, like aligning word embeddings between professional terms and daily expressions to achieve the translation (Kang et al., 2016; Weng and Szolovits, 2018), or further using back-translation procedures (Weng et al., 2019) and denoising autoencoder (Pattisapu et al., 2020) to improve the simplification results under unsupervised conditions. For supervised approaches, PharmMT (Li et al., 2020) uses the Seq2seq (Bahdanau et al., 2015) model pulsing a numerical checker to perform the simplification. (Devaraj et al., 2021) improves the BART(Lewis et al., 2019) model using the unlikelihood constraint (Welleck et al., 2019) to penalize the model generating technical words. Although the readability of these methods is higher compared to the dictionary ones by directly modifying sentences, the accuracy of term-level translation cannot be guaranteed. To address these problems, in the design of the DECLARE, we absorb the advantages of both machine translation and dictionary-based approaches to simplify clinical text from both sentence and term levels.

3 `MedLane`: A New Benchmark Dataset

3.1 Data Collection

The MIMIC-III database (Johnson et al., 2016) contains de-identified data from 58,976 ICU patient admissions, which includes several types of medical information such as demographics, medications, clinical notes, and so on. We first select clinical notes from the NOTEEVENTS table of the MIMIC-III v1.4 dataset[2] focusing on the following three sections: (1) History of present ill-

---

[2]https://mimic.physionet.org/mimictables/noteevents/

Algorithm 1: Sentence Selection Algorithm

Input: Target sentence s, top-3000 word set T, medical abbreviation set A
Output: Selected sentence set

1 Tokenize s into words $[w_1, ..., w_n]$;
2 **for** $i = 1$ *to* n **do**
3   **if** $w_i \in A$ **then**
4        $abb = abb + 1$;
5   **end**
6     $w_i' = lemmatize(w_i)$;
7   **if** $w_i' \notin T$ **then**
8        $unc = unc + 1$;
9   **end**
10 **end**
11 **if** $n < 10$ *or* $\frac{unc+abb}{n} > 0.5$ *or* $\frac{unc+abb}{n} < 0.1$ **then**
12   **return** False;
13 **else**
14   **return** True;
15 **end**

ness, (2) Brief summary of hospital course, and (3) Brief hospital course. These three sections contain thoughts and reasoning for the communication between clinicians, which are usually written with professional medical jargon and abbreviations. However, they still contain many plain sentences such as *"She now also reports of being hunger."*. To avoid translating them again, we design a heuristic feature-based sentence selection approach to filter out such sentences. In particular, we first tokenize each sentence into a set of words and then use a dictionary-based approach to match medical abbreviations. We also count the number of commonly-used English words within a given list[3]. Based on the length of the sentences (greater than 10), the ratio of medical abbreviations (smaller than 0.5), and the ratio of top-3000 words (greater than 0.1), we can automatically select candidate sentences. The sentence selection algorithm is summarized in Algorithm 1.

3.2 Data Annotation

After we collect a set of source sentences, the next step is to annotate them. However, annotating medical sentences is different from creating a parallel bilingual translation corpus. The medical language translation task aims to "translate" professional and clinical jargon to layperson-understandable language, which is still from the same language but uses different expressions. Besides, annotating bilingual translation data focuses on readability and accuracy. Except for those two perspectives,

annotating medical sentences also considers understandability, which is from the perspective of patients or customers. Based on the above guidance, we invited six researchers to annotate the data. All of them are familiar with medical data. For each sentence, there are two extra senior researchers holding the Doctor of Medicine (M.D.) degree to check the annotation quality. For the translated sentences with low quality, senior researchers need to re-translate them.

The purpose of this work is to create a benchmark for the automated clinical language understanding task, which is not only used for training translation models but also for fair evaluation of different approaches. Thus, we set different requirements for workers when annotating the training data and validation/testing data. In general, they use two steps when annotating each source sentence. The first step is to paraphrase the abbreviations with the whole words or phrases. For each abbreviation, there may be several full forms. For example, *"TLC"* has two full forms[4]. One is *"thin-layer chromatography"*, and the other is *"total lung capacity"*. Therefore, it is important for workers to understand the context in which the abbreviation or term has been used. Note that we do not provide a dictionary for workers, and they search the full expressions on the Internet if they are not sure about the abbreviations. The second step is to use simple words to replace professorial medical expressions. Take the word *"hematocrit"*[5] as an example, which means the ratio of the volume of red blood cells to the volume of the whole blood. If we use the expression, *"the proportion of red blood cells in the blood"*, it is much more understandable for patients compared with directly using professional clinical jargon. An example in Figure 1 illustrates the annotating procedure.

When annotating the training data, workers are asked to return the final understandable sentences, i.e., the simplified ones, which will be checked by experts again to guarantee the annotation quality. For validation and testing data, we require workers to return both rephrasing and simplifying forms for each source sentence.

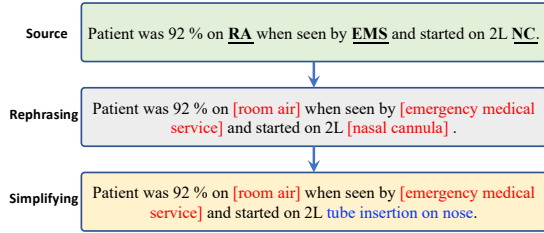Note that for all the training, validation, and testing data, there is a special case that we do not need

Figure 1: An example of annotating a sentence by a worker using two steps, i.e., rephrasing and simplifying. In the rephrasing step, three abbreviations are replaced by full forms. In the simplifying step, the full form "nasal cannula" is replaced by "tube insertion on nose".

to translate the source sentence. For example, it is easy to understand the sentence "*He had a set of surveillance blood cultures drawn last week, which were negative.*". These sentences are extremely useful when training an understandable translation model because they can be considered as important indicators for guiding model learning. In the validation and testing data, another special case is that the sentence may not be simplified any more. For example, the source sentence is "*She also had subjective SOB with CXR suggesting fluid overload.*", and the rephrased and simplified sentences are the same, which is "*She also had subjective [shortness of breath] with [chest x-ray] suggesting fluid overload.*".

# of tokens in the source sentences	14,780
# of tokens in the target sentences	14,278
# of overlapped tokens between source & target	12,501
Avg. length of the source sentences	20.6
Avg. length of the target sentences	24.0
Avg. # of abbreviations in validation & testing sets	1.2

Table 2: MedLane data statistics.

3.3 Dataset Statistics

The MedLane dataset contains 12,801 sentences for training, 1,015 sentences for validation, and 1,016 sentences for testing. Table 2 shows the statistics of the MedLane dataset, which are different from those of traditional machine translation datasets. First, the way of annotation is different, as we discussed in the previous section. Second, there are a large number of overlapped tokens between source and target sentences, which is also different from traditional machine translation. These differences make our task unique and challenging.

4 DECLARE: An Effective Approach

Motivated by the annotation steps, we propose an effective end-to-end model named DECLARE for the automated clinical language simplification task.

In particular, we collect a medical dictionary to estimate the possible full-term expressions of abbreviations in the input sentences. The proposed model is shown in Figure 2, which consists of three main modules, i.e., a complex word locator (CWL), a dictionary-based neural lexical interpreter (DNLI), and a restricted syntactic simplification polisher (RSSP).

Given a tokenized professional medical sentence $W = [w_1, w_2, \cdots, w_n]$, where n denotes the number of tokens, the locator aims to dig out possible phrases that need to be simplified or translated. In the neural lexical interpreter, the chosen phrases will be replaced with full-term expressions selected from the medical dictionary. Finally, the replaced sentence will pass the polisher to generate the final output Y. These three parts tightly work together and enhance each other. Next, we introduce the design details of each module, respectively.

4.1 Complex Word Locator (CWL)

As we discussed in Section 3, clinical language understanding is different from the traditional machine translation task, and we only need to modify a part of professional medical jargon in the input sentence $W = [w_1, w_2, \cdots, w_n]$ and keep the remaining words. Towards this end, we design a complex word locator to find out the tokens that need to be modified. Note that in the annotated dataset, we have such indicators that which tokens are modified. In particular, we use a pre-trained BERT model (Devlin et al., 2019) with PubMed data, i.e., PubMedBERT (Gu et al., 2020), to learn a representation for each input token w_i, and a fully-connected layer (FC) followed by the softmax function is used to calculate the probability of each token to be modified or not. Let $\hat{\mathbf{p}}_i$ denote the binary probability vector, and we have

$$\hat{\mathbf{p}}_i = \text{softmax}(\text{FC}(\text{PubMedBERT}(w_i))). \quad (1)$$

Let $\mathbf{p}_i \in \{0, 1\}^2$ be the ground truth vector for the i-th token and $\mathcal{L}_p$ denote the average cross-entropy (CE) loss function, i.e.,

$$\mathcal{L}_p = \frac{1}{n} \sum_{i=1}^{n} \text{CE}(\mathbf{p}_i, \hat{\mathbf{p}}_i). \quad (2)$$

4.2 Dictionary-based Neural Lexical Interpreter (DNLI)

When workers annotate the source sentence, the first step is to rephrase the abbreviations. Using the
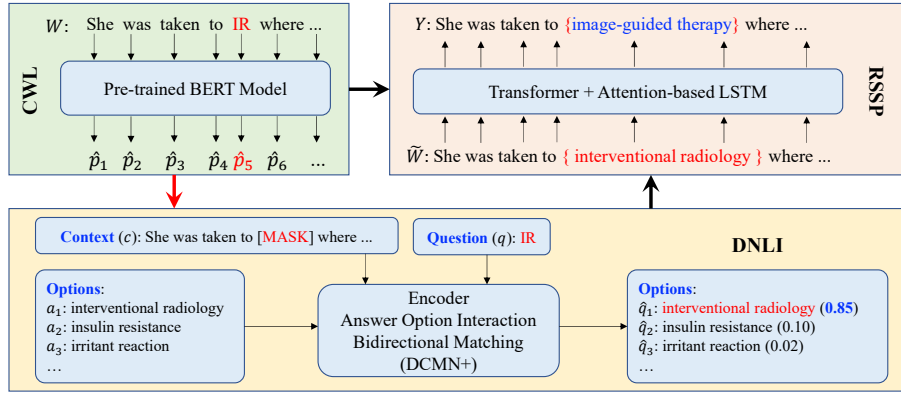
Figure 2: Overview of the proposed DECLARE model.

designed locator, DECLARE is able to identify the possible abbreviations. Then we use a dictionary-based neural lexical interpreter to automatically substitute it with an appropriate full form for each located token. Note that if the located token does not have any full-term expression, we will keep it in the sentence. Some abbreviations have several full-term expressions, and the proposed neural lexical interpreter will automatically choose one with the highest probability based on the context information, i.e., the remaining tokens in the sentence. The details of constructing the dictionary can be found in Section 5.1.

To recover the most appropriate full version of located token w_i, we borrow the idea of DCMN+ (Zhang et al., 2020) to design the dictionary-based neural lexical interpreter. Specifically, we mask the located token w_i from the input sentence, i.e., the masked sentence is $W' = [w_1, \cdots, w_{i-1}, [\text{MASK}], w_{i+1}, \cdots, w_n]$, which can be considered as the context information. Then the located token w_i can be regarded as the question, and all the full terms $A = \{a_1, \cdots, a_m\}$ are considered as options, where m is the number of possible full versions. The goal is to select the best candidate $\hat{a}_j$ from the options A when given the context vector W' and the question w_i.

We first encode W' with PubMedBERT and denote the encoded vector as $\mathbf{c}$. Similarly, the question w_i can also be mapped to a vector $\mathbf{q}$, and each option a_j can be converted to a vector $\mathbf{a}_j$. Then using answer option interaction and bidirectional matching modules in DCMN+, we can estimate the probability of each full-term expression to be selected, which is denoted as $\mathbf{q} = [q_1, \cdots, q_m]$.

Note that there may be multiple located tokens in a sentence, and we will generate the corresponding number of {context, question, options} pairs. Besides, when annotating the MedLane dataset, the

annotators use square brackets "[]" to indicate the correct full term expressions as shown in Figure 1. Thus, there are ground truths that are denoted as $\hat{\mathbf{q}}$ for {context, question, options} pairs. Assume that there are l located tokens in the input sentence, then we will have the following loss function:

$$\mathcal{L}_q = \frac{1}{l} \sum_{j=1}^{l} \text{CE}(\mathbf{q}_j, \hat{\mathbf{q}}_j). \quad (3)$$

The output of the dictionary-based neural lexical interpreter is a new sentence by replacing located tokens in the locator with the best candidates selected from the dictionary, which is denoted as $\tilde{W} = [\tilde{w}_1, \cdots, \tilde{w}_{n'}]$, where n' is the number of tokens in the new sentence $\tilde{W}$.

4.3 Restricted Syntactic Simplification Polisher (RSSP)

The second step of annotation is to polish the sentences to make them more fluent, simple, and understandable by patients with low health literacy. An easy way is to directly "translate" the new sentence $\tilde{W}$ to the target sentence with a neural machine translation model. However, as we mentioned before, the clinical language understanding task is different from traditional neural machine translation, which aims to make the professional clinical jargon understandable by patients. In fact, most of the tokens in the sentences can be kept and do not translate again, and we only polish the terms labeled by the locator and replace them with the interpreter. Thus, this is a *partial translation task*.

First, we add special markers to $\tilde{W}$ to indicate the tokens to be polished, i.e., $\tilde{W} = [\tilde{w}_1, \cdots, \{\tilde{w}_i, \cdots, \tilde{w}_j\}, \cdots, \tilde{w}_{n'}]$, where the tokens from $\tilde{w}_i$ and $\tilde{w}_j$ will be "translated" by the polisher. The polisher first encodes the input sentence $\tilde{W}$ via Transformer (Vaswani et al., 2017),

i.e., $\mathbf{v}_i = \text{transformer}(w_i|w_1, \cdots, w_{n'})$, where $\mathbf{v}_i$ is the representation of the i-the token. Then an attention-based LSTM (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014; Bahdanau et al., 2015; Zhou et al., 2021) is used as the decoder to generate the t-th word y_t within braces. Let $\mathbf{h}_t$ represent the hidden state outputted by the decoder LSTM, and then we can obtain the weighted context vector $\mathbf{s}_t$ using the attention mechanism as follows:

$$\mathbf{s}_t = \sum_{i=1}^{n} \alpha_{ti}\mathbf{v}_i, \alpha_{ti} = \frac{\exp(o_{ti})}{\sum_{j=1}^{n}(\exp(o_{tj}))}, o_{ti} = \mathbf{h}_t\mathbf{v}_i^\top.$$

Finally, we can obtain the probability of the t-th word: $\mathbf{r}_t = \text{softmax}(\mathbf{W}[\mathbf{h}_t; \mathbf{s}_t] + \mathbf{b})$, where $\mathbf{W}$ and $\mathbf{b}$ are two parameters. Assume that the k-the element of $\mathbf{r}_t$ corresponds to the truth token, and then we have the loss of the polisher as follows:

$$\mathcal{L}_r = -\frac{1}{j - i + 1}\sum_{t=i}^{j}\log(\mathbf{r}_t[k]). \quad (4)$$

4.4 Training

The proposed DECLARE is an end-to-end model, and we can train the model using the following loss function:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_q + \mathcal{L}_r. \quad (5)$$

In the evaluation stage, we can run these three modules one by one to generate the final understandable sentences.

5 Experiments

5.1 Experimental Setups

Dictionary Construction. We construct the mapping dictionary by collecting medical abbreviations and their corresponding full forms from the book (Dorland, 2016) and online sources, including Charleston Area Medical Center[6], Taber's Medical Dictionary[7], your dictionary[8], MedicineNet[9], and Wikipedia[10].

Baselines. We use the following approaches as baselines: a simple term replacement approach named *Dictionary-based model*, a statistical machine translation (SMT) system *Moses*[11], neural machine translation approaches, including

Seq2Seq (Bahdanau et al., 2015) and its two variants *Seq2Seq−* and *Seq2Seq-S*, a modified version of the pointer network (Vinyals et al., 2015) *Pointer-Net*, state-of-the-art language models *BART* (Lewis et al., 2019) and *T5* (Raffel et al., 2019), and a modified BERT-based simplifier *BERT-MT*. We also include *EditNTS* (Dong et al., 2019) – a method proposed for the general ATS – as a baseline. Although there are other more advanced methods (Martin et al., 2020, 2019; Maddela et al., 2020) in the general ATS domain, they all require special grammar level information like the part of speech and syntactic tree information, which is hard to obtain for our clinical domain dataset. Thus, they are not compared in our experiments.

Parameter Settings. For the dictionary method, we use the pre-constructed dictionary as the same as the DECLARE model. For the statistical model Moses, we follow the training procedure listed on the User Manual and Code Guide file[12]. For neural machine translation models and text summarization baseline, we all conduct a grid search to find the optimal parameters. For Seq2Seq, Seq2Seq−, Seq2Seq-S, and PointerNet, the hidden size is set to 256 for both encoder and decoder by greedy search, and the learning rate is set to $1e - 3$. We use Adam (Kingma and Ba, 2014) as the optimizer. Tokenization is performed using NLTK word tokenizer (Bird et al., 2009). The early stop is also applied by checking the BLEU score (Papineni et al., 2002) on the validation set, and the training batch size is set to 30.

For EditNTS, we use the original default parameter setting with the learning rate of $1e - 3$ for the Adam optimizer. For the BERT-MT model, the hidden size is the same as that of PubMedBERT, which is 786. We also use the default AdamW optimizer used by PubMedBERT with the learning rate as $5e - 5$, the warm-up method, the default PubMedBERT vocabulary, and tokenization are applied. For BART and T5, the setting of the optimizer and training procedure is the same as the BERT-MT.

For the proposed DECLARE, the locator is based on PubMedBERT to perform token-level classification, and we use the default setting of PubMedBERT to train the locator. For the dictionary-based neural interpreter, we use the same parameter setting as (Zhang et al., 2020). The max size of the

---

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU	METEOR	ROUGE-L	CIDEr	SARI	HIT	CWR	AScore
Dictionary	0.7158	0.6364	0.5684	0.5076	0.6070	0.3933	0.7308	4.2037	37.3391	0.5572	0.6407	0.5948
Moses	0.7880	0.7130	0.6530	0.6016	0.6889	0.4237	0.8188	5.1046	51.6827	0.6823	0.7543	0.6859
Seq2seq	0.7136	0.6322	0.5969	0.5160	0.6147	0.3533	0.7609	4.1299	46.1328	0.7388	0.7980	0.6648
Seq2seq-	0.5066	0.3315	0.2373	0.1787	0.3135	0.1859	0.4948	1.2670	24.5346	0.6427	**0.8367**	0.4070
Seq2seq-S	0.7180	0.6386	0.5778	0.5267	0.6153	0.3604	0.7683	4.2635	46.5085	0.7331	0.7953	0.6630
PointerNet	0.6870	0.5904	0.5158	0.4541	0.5618	0.3338	0.7285	3.9458	42.2857	0.6414	0.7555	0.5949
EditNTS	0.8213	0.7801	0.7452	0.7132	0.7649	0.4674	0.7401	5.9508	62.6036	0.6405	0.6915	0.7116
BART	0.7148	0.6755	0.6396	0.6060	0.6590	**0.5320**	0.7616	4.9783	70.3058	0.5266	0.7311	0.6191
T5	0.7223	0.6812	0.6445	0.6103	0.6646	0.5305	0.7645	5.0629	71.3255	0.5262	0.7342	0.6220
BERT-MT	0.8003	0.7428	0.6952	0.6531	0.7228	0.4566	0.8218	5.3293	**72.2260**	0.7808	0.7358	0.7417
DECLARE	**0.8624**	**0.8291**	**0.8004**	**0.7737**	**0.8165**	0.5290	**0.8894**	**6.7212**	70.8583	**0.7986**	0.7328	**0.7983**
↑	+5.0%	+6.3%	+7.4%	+8.5%	+6.7%	-0.5%	+8.2%	+26.1%	-1.9%	+2.2%	-12.4%	+7.6%

Table 3: Performance evaluation of all the baselines with different metrics. ↑ denotes the percentage of performance gain compared with the best baselines.

answers is set to 8. The maximum length of the input sentence is set to 64 during training. The learning rate is set to $5e-5$ with ten epochs, and an early stop is adopted. For the restricted polisher, the setting is the same as the BERT-MT model except for the restricted translation setting. PubMedBERT is applied with an LSTM decoder that has the same hidden size.

In the evaluation stage, the same NLTK word tokenizer is applied as baselines to break the sentences into words for calculating the scores for a fair comparison. All models are trained on Ubuntu 16.04 with 128GB memory and an Nvidia Tesla P100 GPU.

Evaluation Metrics. We use BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) scores as the evaluation metrics, which are widely-used for the machine translation task. Besides, we use SARI (Xu et al., 2016), which is specially designed for the general ATS task. It combines the n-gram evaluation method of the BLEU score and rewards the replacement of the input words. However, it fails to cover the accuracy requirement of the professional medical term simplification.

Since our task is different from traditional machine translation and ATS tasks, directly applying existing evaluation metrics cannot fairly evaluate the performance of different models. Since our task is different from traditional machine translation tasks, directly applying existing evaluation metrics cannot fairly evaluate the performance of different models. Here, we use an example in Figure 3 to demonstrate the failure of existing evaluation metrics, such as the BLEU score. If we directly copy the original source sentence as the answer, a very high BLEU score can be obtained, which is 0.85. However, the critical term "*pt*" is not translated.

Without translating this term, patients or customers may not totally understand the meaning of this sentence. Thus, it is necessary to design task-specific metrics.

```
Source: vascular saw the pt and
did not feel that there was an acute
need for an invasive procedure.
Target: vascular saw the patient
and did not feel that there was
an acute need for an invasive
procedure.
```

Figure 3: Example of the failure of existing metrics.

- *Hit Ratio (HIT)*. A key challenge of medical language translation is to translate professional medical jargon into layperson-understandable words. Let n_p denote the number of professional medical terms in a source sentence and n_t be the number of correctly translated terms in the target. We then have the HIT ratio, which is $HIT = \frac{n_t}{n_p}$.

- *Common Word Ratio (CWR)*. To evaluate the simplicity of the translated sentences, we follow the work of Dale–Chall readability (McClure, 1987) to calculate the common word ratio for each output sentence. We first lemmatize each word of the output. If the lemmatized form is in the top-3000 commonly-used English words, then it is a common word. Otherwise, it is not a common word. Let n_c denote the number of common words in the translated sentence, and n represents the length of the translated sentence. The common word ratio score is $CWR = \frac{n_c}{n}$.

- *Aggregation Score (AScore)*. The quality of the translated sentences is not only decided by the readability (BLEU) but also related to the correctness (HIT) and simplicity (CWR). Among these three perspectives, readability and correctness should be more important than simplicity. Thus, we design a

new score to model them jointly, which is

$$Ascore = \frac{1 + \alpha^2 + \beta^2}{\frac{\alpha^2}{BLEU} + \frac{\beta^2}{HIT} + \frac{1}{CWR}},$$

where α and β are parameters for controlling the importance of BLEU and HIT scores. If any of the three metrics is 0, then it will be added to a very small number such as 10^{-8} to avoid AScore being 0. We take $\alpha = 2$ and $\beta = 1.5$ in our experiments.

5.2 Experimental Result Analysis

Table 3 shows the experimental results of all baselines and DECLARE on different evaluation metrics. The Dictionary method is the simplest approach, but its performance is not the worst compared with other baselines in terms of transitional machine translation evaluation metrics such as BLEUs. The reason is that there are many overlapping tokens in both source and target sentences, which is the main difference between the traditional machine translation task and our clinical language understanding task. These results also confirm that we need to design new evaluation metrics for this new task.

General neural network-based approaches, including Seq2Seq, its variants, and PointerNet, have a relatively low BLEU score. The reason is that the labeled data is insufficient for them to train a powerful translation model from scratch. For the general ATS method EditNTS, we can observe a relatively high score for BLEU. EditNTS is good at keeping the information. However, due to the lack of external knowledge support, the HIT score is still not satisfactory enough, proving the importance of the DNLI module. In addition, EditNTS is designed for the general ATS domain. Many attributes are not suitable for our task, which can also contribute to the bad performance. For the transformer-based pre-trained sequence to sequence models T5 and BART, the BLEU score is relatively high but with a much lower HIT score compared to Seq2Seq and PointerNet. The unsatisfied results may be related to the domain shifting problem since T5 or BART models are not pre-trained for the medical text-domain. On the contrary, BERT-MT conducting pre-training on a large medical language corpus significantly increases the ability of model learning. Hence, the performance of BERT-MT is the best among all the baselines. However, the proposed DECLARE mimics the human annotation procedure and employs a mapping dictionary with a novel model design, which leads to achieving the best performance compared with all the baselines.

Using the pre-trained language models is helpful to attain a higher SARI score because the pre-training technique can increase the models' generalization ability and benefit the word replacement rewards of the SARI metric. However, SARI does not focus on the accuracy of simplifying professional medical abbreviations, which makes it unable to comprehensively and fairly evaluate the results.

From the view of the HIT score, we can find a sufficient gap between generation-based methods (including Seq2Seq, Seq2Seq-S, BERT-MT, and DECLARE) and other methods. To achieve a high HIT score, the accurate recognition of abbreviations is necessary. Moreover, the models should understand the context correctly, which is an advantage of neural network-based models. Besides, another requirement is the generation ability. The use of a reference mechanism probably limits the generation ability of the PointerNet model, and thus, it does not achieve a high HIT score.

The CWR score can reflect the simplicity of sentences in a scene. However, we should notice that the higher CWR scores do not mean better performance. The reason is that translating professional medical terms will inevitably result in some uncommon words. Thus, we should attach less importance to the CWR score when calculating the comprehensive rank.

The top 3 models in the view of AScore are DECLARE, BERT-MT, and EditNTS, which is reasonable. The AScore attaches the highest importance to BLEU, followed by the HIT and CWR scores. Using the harmonic mean can make sure that we penalize the tendency of going overboard on one subject and guarantee good general performance.

6 Conclusion

This paper aims to benchmark a new, challenging, yet practical task of automated clinical language simplification by constructing a high-quality `MedLane` dataset and proposing a new model DECLARE that mimics the human annotation procedure. We conducted experiments on the annotated `MedLane` dataset by implementing nine strong baselines against DECLARE. Experimental results confirmed the utility of the constructed `MedLane` dataset and the effectiveness of the proposed DECLARE for addressing the automated clinical language simplification task.

7 Broader Impacts

Current health care service and health information technology (HIT) system design strive to provide accessible ways for patients to be engaged in their own care and make informed decisions. One instance is to make personal health records accessible to patients through patient portals (the electronic personal health record systems connected to organizations' electronic health record systems). However, only providing access to health records is insufficient to fully empower patients. Patients may struggle to understand those records due to such reasons as low health literacy, unfamiliarity with medical jargon and clinical abbreviations, or difficulty in understanding the complex structure of medical language. Currently, only a few studies target this practical issue. In this paper, we not only provide a human-annotated dataset, design a new model but also propose three evaluation metrics for benchmarking the clinical language simplification task. Our work will expedite research in multiple domains, including but not limited to, natural language processing, machine learning, and healthcare informatics.

8 Ethical Consideration

The original data of our study are directly extracted from the MIMIC-III database (Johnson et al., 2016), where all private health information was de-identified. The MIMIC-III database was performed under Health Insurance Portability and Accountability Act (HIPAA) standards, which require the removal of all the identifying data elements in the list of HIPAA (e.g., name, phone number, address, and so on). Thus, this is no privacy issue for the data that we use. When annotating the dataset, all annotators submitted all required consent forms. Since this work only focuses on simplifying clinical text, and no additional identified and private information is added. As a result, the protection of privacy is preserved. For disseminating our dataset to be publicly available, we will follow the same requirement of the MIMIC-III data. In other words, the requester must complete a recognized training for protecting human research participants and sign an agreement to protect the data privacy following the requirement of the PhysioNet[13] (Goldberger et al., 2000).

---

[13] https://physionet.org/

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015 : International Conference on Learning Representations 2015*.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.

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

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of ACL*, pages 1061–1071.

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):e26.

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):e8531.

Louise Deléger and Pierre Zweigenbaum. 2008. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA Annual Symposium Proceedings*, volume 2008, page 146.

Ashwin Devaraj, Iain Marshall, Byron C Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. Editnts: An neural programmer-interpreter model for sentence simplification through explicit editing. *arXiv preprint arXiv:1906.08104*.

Dorland. 2016. *Dorland's Dictionary of medical acronyms & abbreviations, Seventh Edition.* Elsevier, Inc.

N Elhadad, WW Chapman, T O'Gorman, M Palmer, and G Savova. 2013. The share schema for the syntactic and semantic annotation of clinical texts. *Under Review*.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of biomedical informatics*, 35(4):222–235.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of AAAI*.

Jennifer Fong Ha and Nancy Longnecker. 2010. Doctor-patient communication: a review. *Ochsner Journal*, 10(1):38–43.

Sam Henry, Yanshan Wang, Feichen Shen, and Ozlem Uzuner. 2020. The 2019 national natural language processing (nlp) clinical challenges (n2c2)/open health nlp (ohnlp) shared task on clinical concept normalization for clinical records. *Journal of the American Medical Informatics Association*, 27(10):1529–1537.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, volume 2010, page 366.

Hong Jin Kang, Tao Chen, Muthu Kumar Chandrasekaran, and Min Yen Kan. 2016. A comparison of word embeddings for english and cross-lingual chinese word sense disambiguation.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Barbara M Korsch, Ethel K Gozzi, and Vida Francis. 1968. Gaps in doctor-patient communication: I. doctor-patient interaction and patient satisfaction. *Pediatrics*, 42(5):855–871.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378.

John P Lalor, Beverly Woolf, and Hong Yu. 2019. Improving electronic health record note comprehension with noteaid: randomized trial of electronic health record note comprehension interventions with crowdsourced workers. *Journal of medical Internet research*, 21(1):e10793.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiazhao Li, Corey Lester, Xinyan Zhao, Yuting Ding, Yun Jiang, and VG Vinod Vydiswaran. 2020. Pharmmt: A neural machine translation approach to simplify prescription directions. In *Proceedings of EMNLP: Findings*, pages 2785–2796.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lisa Grossman Liu, Raymond H Grossman, Elliot G Mitchell, Chunhua Weng, Karthik Natarajan, George Hripcsak, and David K Vawdrey. 2021. A deep database of medical abbreviations and acronyms for natural language processing. *Scientific Data*, 8(1):1–9.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2020. Controllable text simplification with explicit paraphrasing. *arXiv preprint arXiv:2010.11004*.

3560

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Louis Martin, Benoît Sagot, Eric de la Clergerie, and Antoine Bordes. 2019. Controllable sentence simplification. *arXiv preprint arXiv:1910.02677*.

Glenda M McClure. 1987. Readability formulas: Useful or useless? *IEEE transactions on professional communication*, (1):12–15.

Gustavo Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Gustavo Paetzold and Lucia Specia. 2017. 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*, pages 34–40.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Ruth M Parker, Mark V Williams, Barry D Weiss, David W Baker, Terry C Davis, Cecilia C Doak, Leonard G Doak, Karen Hein, Cathy D Meade, Joann Nurss, et al. 1999. Health literacy-report of the council on scientific affairs. *Jama-Journal of the American Medical Association*, 281(6):552–557.

Nikhil Pattisapu, Nishant Prabhu, Smriti Bhati, and Vasudeva Varma. 2020. Leveraging social media for medical text simplification. In *Proceedings of SIGIR*, pages 851–860.

Balaji Polepalli Ramesh, Thomas Houston, Cynthia Brandt, Hua Fang, and Hong Yu. 2013. Improving patients' electronic health record comprehension with noteaid. In *MEDINFO 2013*, pages 714–718. IOS Press.

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):e417.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Afshin Rahimi, Timothy Baldwin, and Karin Verspoor. 2020. Wikiumls: Aligning umls to wikipedia via cross-lingual neural ranking. *arXiv preprint arXiv:2005.01281*.

Tarek Sakakini, Jong Yoon Lee, Aditya Duri, Renato FL Azevedo, Victor Sadauskas, Kuangxiao Gu, Suma Bhat, Dan Morrow, James Graumlich, Saqib Walayat, et al. 2020. Context-aware automatic text simplification of health materials in low-resource domains. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 115–126.

Felix Stahlberg and Shankar Kumar. 2020. Seq2edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159.

Rebecca L Sudore, Kristine Yaffe, Suzanne Satterfield, Tamara B Harris, Kala M Mehta, Eleanor M Simonsick, Anne B Newman, Caterina Rosano, Ronica Rooks, Susan M Rubin, et al. 2006. Limited literacy and mortality in the elderly. *Journal of general internal medicine*, 21(8):806–812.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27:3104–3112.

Hoang Van, David Kauchak, and Gondy Leroy. 2020. Automets: The autocomplete for medical text simplification. *arXiv preprint arXiv:2010.10573*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of CVPR*, pages 4566–4575.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700.

V. G. Vinod Vydiswaran, Qiaozhu Mei, David A Hanauer, and Kai Zheng. 2014. Mining consumer health vocabulary from community-generated text. *AMIA*, 2014:1150.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Wei-Hung Weng, Yu-An Chung, and Peter Szolovits. 2019. Unsupervised clinical language translation. In *Proceedings of SIGKDD*, pages 3121–3131.

Wei-Hung Weng and Peter Szolovits. 2018. Mapping unparalleled clinical professional and consumer languages with embedding alignment.

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*, 4:401–415.

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

Qing Zeng-Treitler, Sergey Goryachev, Hyeoneui Kim, Alla Keselman, and Douglas Rosendale. 2007. Making texts in electronic health records comprehensible to consumers: a prototype translator. In *AMIA Annual Symposium Proceedings*, volume 2007, page 846.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2020. Dcmn+: Dual co-matching network for multi-choice reading comprehension. In *Proceedings of AAAI*, volume 34, pages 9563–9570.

Yifan Zheng, Yun Jiang, Michael P Dorsch, Yuting Ding, VG Vinod Vydiswaran, and Corey A Lester. 2021. Work effort, readability and quality of pharmacy transcription of patient directions from electronic prescriptions: a retrospective observational cohort analysis. *BMJ quality & safety*, 30(4):311–319.

Yao Zhou, Haonan Wang, Jingrui He, and Haixun Wang. 2021. From intrinsic to counterfactual: On the explainability of contextualized recommender systems. *CoRR*, abs/2110.14844.