Context-Aware Automatic Text Simplification of Health Materials in Low-Resource Domains

Tarek Sakakini, Jong Yoon Lee, Aditya Duri, Renato F.L. Azevedo, Kuangxiao Gu
University of Illinois at Urbana-Champaign
{sakakini, jlee642, duri2, razeved2, kgu3}@illinois.edu

Suma Bhat, Dan Morrow, Mark Hasegawa-Johnson, Thomas Huang
University of Illinois at Urbana-Champaign
{spbhat2, dgm, jhasegaw, t-huang1}@illinois.edu

Victor Sadauskas, James Graumlich, Saqib Walayat
University of Illinois College of Medicine, Peoria
{vsadau2, jfg, swalayat}@uic.edu

Ann Willemsen-Dunlap, and Donald Halpin
Jump Simulation Center, Peoria, Illinois
{ann.m.willemsen-dunlap, donald.j.halpin}@jumpsimulation.org

Abstract

Healthcare systems have increased patients' exposure to their own health materials to enhance patients' health levels, but this has been impeded by patients' lack of understanding of their health material. We address potential barriers to their comprehension by developing a context-aware text simplification system for health material. Given the scarcity of annotated parallel corpora in healthcare domains, we design our system to be independent of a parallel corpus, complementing the availability of data-driven neural methods when such corpora are available. Our system compensates for the lack of direct supervision using a biomedical lexical database: Unified Medical Language System (UMLS). Compared to a competitive prior approach that uses a tool for identifying biomedical concepts and a consumer-directed vocabulary list, we empirically show the enhanced accuracy of our system due to improved handling of ambiguous terms. We also show the enhanced accuracy of our system over directly-supervised neural methods in this low-resource setting. Finally, we show the direct impact of our system on laypeople's comprehension of health material via a human subjects' study ($n = 160$).

1 Introduction

Healthcare practices have granted patients increased access to their health information to support self-care (Davis et al., 2005; Detmer et al., 2008). But, the benefits have been hindered by patients' low comprehension of their own health data (Irizarry et al., 2015), as a study shows that readability measures of online health information is significantly higher than patient health literacy abilities (Mcinnes and Haglund, 2011). Moreover, older adults, the largest demographic group interacting with the healthcare system, are often the least health-literate (Kessels, 2003; Kutner et al., 2006). With low levels of health literacy resulting in worse health outcomes (Ha and Longnecker, 2010; Kindig et al., 2004), there is an urgent need to reduce the gap between the health literacy of patients and the health literacy demands of healthcare systems. Accordingly, we explore text simplification methods for health text, taking medication instructions as a use case (see Table 1).

This mismatch in patient literacy levels and health documents is due in part to the differing language used by healthcare professionals and patients (Rotegard et al., 2006). For example, what professionals refer to as "PO", patients might refer

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 115–126

https://doi.org/10.18653/v1/P17

Original:
Take 50 mg PO daily for 2 days. Hold for SBP $<$ 90.
Gold Simplification:
Take 50 milligrams by mouth daily for 2 days. Hold for systolic blood pressure $<$ 90.
Dr. Babel Fish:
Take 50 milligrams orally daily for 2 days. Hold for systolic blood pressure $<$ 90.

Table 1: Example medication instruction, its target simplification, and our system's simplification. Color coding reflects replacement.

to as "by mouth". While previous works have addressed this by performing local word replacement (Kandula et al., 2010), their context-free frameworks lacked the accuracy. In a health document, "Mg" could mean "milligrams" or "Magnesium", and harnessing the contextual information, for example in "Take 50 Mg" or "Mg reacts with", aids accurate simplification.

Our approach is a context-aware medical text simplification system, named Dr. Babel Fish (DBF). We design our system to be independent of the availability of annotated datasets as scarcity of such data is expected due to privacy and proprietary concerns. To compensate for annotated datasets, we instead rely on a structured knowledge base in the form of the Unified Medical Language System (UMLS) (Bodenreider, 2004; Lindberg et al., 1993). Taking inspiration from the modular and context-aware frameworks of Phrase-Based Statistical Machine Translation (PBSMT) systems (Koehn et al., 2003), our system, DBF, first identifies hard (low frequency) words such as "SBP", then collects possible simplifications of these words from the UMLS such as {systolic blood pressure, and serotonin-binding protein}, and finally chooses the simplification that best reflects patients' preferred medical terms and best fits the context ("systolic blood pressure" in this case), by relying on a patient language model trained on a suitable monolingual corpus.

Although Neural Machine Translation (NMT) frameworks (Bahdanau et al., 2014; Sutskever et al., 2014) constitute the state-of-the-art, they suffer in the low-resource settings of the clinical (medical) domains, and we accordingly present our system to complement neural methods in domains lacking the appropriate parallel corpus. Although we take medication instructions as a use case, our system is general enough by construction, to handle any medical text. All code and materials associated with this study are released to the public[1]. This paper makes the following contributions:

- It studies a knowledge-aware text simplification model that does not rely on parallel text.
- The study empirically demonstrates the higher precision simplification output of the proposed model compared to previous methods.
- It makes a parallel corpus of medication instructions available to foster future research.
- It provides a comprehensive and comparative study of NMT models applied to healthcare text simplification, previously impossible due to the lack of a parallel corpus.
- Via a human subjects' study, it shows the positive impact of DBF on patient comprehension.

2 Previous Work

Efforts to improve patient comprehension of health information in the biomedical informatics community can be categorized into: developing standards (Atreja et al., 2005; Wolf et al., 2011), curating dictionaries (Zeng and Tse, 2006), annotating text with additional information (Tupper, 2008; Mohan et al., 2013; Zheng and Yu, 2016; Martin-Hammond and Gilbert, 2016), normalizing terms (Mowery et al., 2016), syntactic simplification (Jonnalagadda and Gonzalez, 2010; Kandula et al., 2010; Peng et al., 2012), and finally, lexical simplification (Chen et al., 2018; Kandula et al., 2010; Qenam et al., 2017). Our work belongs to the final category.

One popular previous attempt (Kandula et al., 2010) of health material text simplification relies on the Consumer Health Vocabulary (CHV) (Zeng and Tse, 2006) for mapping the hard term to its simpler counterpart, disregarding context information. Other word-replacement systems (Chen et al., 2018; Qenam et al., 2017) have relied on MetaMap (Aronson, 2001) to map medical terms to their simpler counterparts by either utilizing CHV as a thesaurus (Qenam et al., 2017), or relying on an in-house equivalent resource (CoDeMed) (Chen et al., 2018). Although, MetaMap performs word sense disambiguation (WSD) by relying on the context, its creators admit its low WSD quality (Aronson and Lang, 2010). Therefore, we rely on a language model instead of MetaMap. Nonetheless, since

[1] http://bit.ly/dbf-public-access

MetaMap followed by a CHV (or another dictionary) is a popular method in previous works, we include it as a baseline in our experiments.

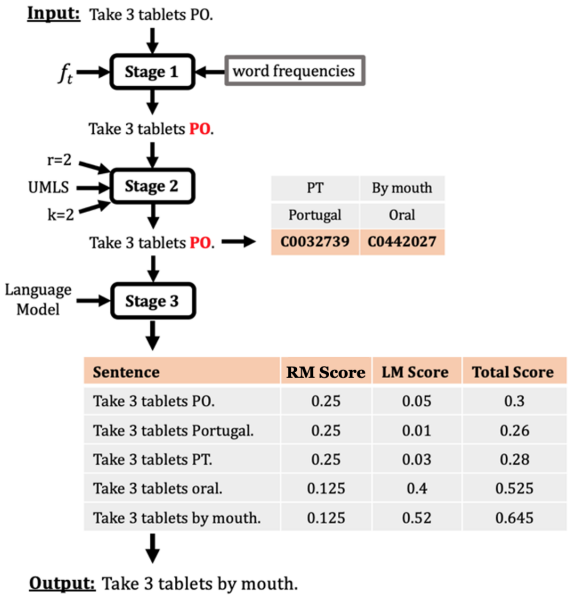
Beyond health materials, lexical simplification is highly researched. A thorough survey of this field is present in (Paetzold and Specia, 2017), which divides work in this field into four stages of a pipeline: (1) Complex Word Identification, (2) Substitution Generation, (3) Substitution Selection, and (4) Substitution Ranking. We also perform Complex Word Identification as a first stage by relying on word frequencies, which is a popular method among previous work (Bott et al., 2012; Leroy et al., 2013; Shardlow, 2013; Wróbel, 2016). We also generate substitutions as a second stage by relying on UMLS, similar to how previous work relied on word taxonomies (Carroll et al., 1998; Devlin, 1998). The last two stages are performed in one shot in DBF, where instead of finding which candidate substitutions fit the context and then select the simplest based on a certain metric, we let the language model decide which is the most probable substitution in terms of meaning and simplicity. Our work is the first work to combine these stages in a context-aware method tailored for the healthcare domain. Finally, text simplification has been modeled previously as a machine translation task where parallel corpora are available (Wang et al., 2016; Wubben et al., 2012; Van den Bercken et al., 2019). Accordingly, we compare against these methods in this study to assess their capacity in the low resource setting and their capability to generalize across healthcare domains.

3 Methods

In this section, we describe our method, DBF, along with the established baselines it was quantitatively evaluated against: MetaMap+CHV, Seq2Seq-w-Attention, and Pointer-Generator.

3.1 Dr. Babel Fish

For reproducibility purposes, following is a detailed system description. DBF is designed as a 3-stage pipeline. First, hard (and easy) words are identified based on their frequency of usage. Then, in the second stage, candidate simplifications of a given hard word are collected and each given a replacement probability (p_{rm}). In the final stage, every candidate output simplification is assigned a language model score and a replacement model score. We will refer to this system as an "unsupervised" system due its independence of annotated datasets, as well as "knowledge-aware" due to its reliance on a knowledge base in the form of UMLS. The highest scoring simplification is then selected as the output of DBF. We describe the 3 stages in the following subsections (see Figure 1).



PT	By mouth
Portugal	Oral
C0032739	C0442027

Sentence	RM Score	LM Score	Total Score
Take 3 tablets PO.	0.25	0.05	0.3
Take 3 tablets Portugal.	0.25	0.01	0.26
Take 3 tablets PT.	0.25	0.03	0.28
Take 3 tablets oral.	0.125	0.4	0.525
Take 3 tablets by mouth.	0.125	0.52	0.645

Figure 1: Block diagram of DBF for the sample sentence: "Take 3 tablets PO." RM (replacement model) score represents p_{rm} for the whole system, and LM (language model) score represents p_{lm} for the whole sentence.

3.1.1 Stage 1: Identification of Hard Words:

In the first stage, the task is to identify the hard words to be translated from the source sentence and to retain the easy words. Let C_t be a corpus of patient-facing health text. Accordingly, we devise a simple statistical model which checks a word's frequency of usage in C_t. We consider the high usage frequency of a word by patients (or targeted towards patients) to be a strong indicator that it is easy for patients to understand, and vice versa. Thus, if a given word has a frequency lower than a tunable frequency threshold (f_t), DBF labels it as hard.

3.1.2 Stage 2: Candidate Generation:

Next, DBF relies on the UMLS to collect all candidate replacements of each hard word, and estimates the probability of each candidate.

A salient feature of the UMLS is its groupings of words/phrases into clusters, where each cluster represents one concept. In Table 2, we present three example concepts, each headed by its "Pre-

Oral	Twice a day	Milligram
PO	BID	Mg
Orally	Twice daily	Milligramos
By mouth	Two times daily	Milligrams

Table 2: Example UMLS concepts

Query	PO	BID
1^{st} Result	Portugal	BID Protein
2^{nd} Result	Oral	Twice a day
3^{rd} Result	Positive	BID gene

Table 3: Example UMLS queries

ferred Name", followed by three example atoms (the UMLS term for a phrase in a given concept). We note the variability of atoms in a concept in terms of complexity, and language.

A second feature of the UMLS is its ability to return an ordered list of concepts to best match a search query. In Table 3, we see the top three concepts returned for two example queries: "PO", and "BID". The correct concept for "PO" appears only second in the results, as is the case for "BID". This suggests that just relying on the top result of such a context-insensitive static search of the UMLS is insufficient for accurate simplifications.

Leveraging these two features, DBF uses the hard word from the input sentence as a query to the UMLS search function. Then, all atoms of the top k returned concepts are considered as candidate simplifications, with concepts ranked higher assigned higher probabilities.

Formally, let $\{C_1, C_2, ..., C_k\}$ be the top k concepts returned by the UMLS search feature when using the hard word c as a query. Also, let $C_i = \{a_{i1}, a_{i2}, ..., a_{in}\}$ be all the atoms of the i^{th} concept. Then, the probability of atom a_{ij} being the simpler replacement of c is assigned $p_{rm}(a_{ij}|c) \propto \frac{1}{r^i}$, where $r \geq 1$. Thus, an atom of the i^{th} concept is allocated a probability r times that of an atom of the $(i+1)^{th}$ concept. In this setup, r and k are tunable hyperparameters of the system. To allow for possibly keeping a hard word c unaltered on the output side, we also assign the probability $p_{rm}(c|c)$ equal to that of an atom in the 1^{st} concept. This helps in cases where a word was wrongly identified as hard, or a simpler alternative does not exist for it. For an easy word e, we assign $p(e|e) = 1$ to force retention of easy words.

3.1.3 Stage 3: Decoding via Language Model:

Finally, we consider all possible combinations of simplifications and choose that with the highest product of replacement probability and language model probability.

Formally, we identify $T(c) = \{t_1, t_2, ..., t_m\}$ which is the set of possible simplifications for a word c. Using this, the set of possible simplifications of the input sentence becomes $H = T(c_1) \times T(c_2) \times T(c_n)$, where $\times$ refers to the Cartesian product of sets.

Now consider a sentence $t \in H$ and let $t = t_1 t_2 ... t_T$ where t_i is the i^{th} word of t. Then, $P(t|c) \propto \prod_{i=1}^{T} p_{rm}(t_i|c_i) * p_{lm}(t_i|t_{i-1:i-5})$, where $p_{lm}(t_i|t_{i-1:i-5})$ is the probability assigned by the 6-gram language model (Brown et al., 1992), for the word t_i occurring after the sequence of words $t_{i-5}t_{i-4}t_{i-3}t_{i-2}t_{i-1}$. The 6-gram language model is trained on the patient-friendly corpus to model the target language. Finally, the sentence t with the highest assigned probability $P(t|c)$ is selected as the output simplification of the system. To further advance this knowledge-aware framework, one can resort to more sophisticated language models such as the recent masked language models (Ghazvininejad et al., 2019).

The significance of the language model is, first, it utilizes the context in which a word like "PO" appears to reward a simplification like "Oral", and penalize a simplification like "Portugal", especially considering that "Portugal" is assigned a higher replacement probability (p_{rm}). Second, it encodes word usage preferences–such as "by mouth" being preferred over "Oral"– even though they both had equal replacement probabilities (p_{rm}).

From an implementation perspective, sentences with high count of hard words would lead to a large H and exponentially slower execution. Hence, we approximate maximizing $P(t|c)$ over H by employing beam search. Traversing H from 1 to n, we maintain only the top 5 sequences as evaluated by $P(t|c)$ up to the respective index.

3.2 MetaMap+CHV:

To compare DBF to the majority of previously used methods for simplifying health materials, we implement the following baseline. Text is first passed through MetaMap, which maps phrases in the text to their respective UMLS concepts. Then, for every phrase identified, we first check if it includes at least one hard word. If it does, and if that UMLS

concept is covered by CHV, we replace it by CHV's most preferred term for that concept, otherwise, we replace it with the UMLS preferred term for that concept. With that being said, all phrases identified by MetaMap, which are not contiguous, are ignored to avoid errors in sentence structure when performing the phrase replacement.

3.3 NMT Baselines:

Representing state-of-the-art approaches, our last set of baselines are two supervised NMT architectures (Jhamtani et al., 2017; Sutskever et al., 2014), requiring training data.

One NMT baseline we consider is a Seq2Seq-with-Attention architecture (Sutskever et al., 2014). In this deep learning architecture, a Long Short-Term Memory (LSTM) encoder maps the input sentence to a fixed length vector, and generates contextualized representations of the input words. Then, an LSTM decoder generates the output words sequentially based on the fixed length vector, and the contextualized representations, while the attention mechanism indicates which input words influence each output decision. For this baseline, we utilize Google's open source implementation (Developers, 2017) with default parameters.

Due to the large overlap in the vocabulary of the source and target sentences, particularly the "easy" words, we consider a second NMT baseline called Pointer-Generator capable of copying words as is from the source sentence (Jhamtani et al., 2017). It differs from Seq2Seq-with-Attention in that at every decode step, it estimates a probability g of generating a new word rather than copying a word from the source sentence. If g is low, the model relies more heavily on the estimated attention distribution over the input source words, which increases the chances of copying the word that is most highly weighted by the attention mechanism. To implement the system, we use the author's original open-source implementation (Jhamtani et al., 2017) with default parameters, except for using the Proximal Adagrad (Singer and Duchi, 2009) optimization algorithm to maximize performance.

4 Experiments

This section describes the results of two studies. The first study uses automated evaluation metrics to assess DBF's output in comparison to the baselines considered. The second study evaluates the impact of DBF on laypeople comprehension. We first describe the data used in our experiments and then present the results.

4.1 Data

Parallel Corpus: For the purpose of training supervised NMT methods, and evaluating all systems, we collected 4554 unique and de-identified medication instructions for diabetic patients from the electronic health records of a collaborating healthcare institution. They were of two types: (1) Structured– automatically populated using three drop-down fields: Dose, Route, Frequency (2) Free-text– manually typed. Free-text instructions tend to have more hard words due to their uncontrolled nature.

Then, for every instruction, a physician, with expertise in standard practices for increasing patient comprehension, annotated each instruction with its accurate simplification. Although one other physician was hired for the same task to ensure a high-quality parallel corpus, it was evident by manual inspection, as well as, overall statistics (such as sentence length and word frequencies used in translation), that the former physician's annotations had superior quality. We thus relied solely on the former physician's annotations.

As shown in Table 4, the ratio of structured instructions to free-text ones is around 2:1. On average, structured instructions are slightly lengthier due to the consistency of their length, while free-text instructions vary between the long and short instructions. In terms of novelty—average count of unique new words added in the simplification—simplifications of free-text instructions introduce more novelty due to the complex nature of these instructions. Finally, more free-text instructions were left after simplification. The high level of novelty despite the high number of unchanged instances shows the disparity in instances between long ones that require significant simplification and short ones that require no simplification. We finally check for the average frequency of words (as estimated by the monolingual corpus) in the complicated and simplified side of both types of medication instructions. We notice that, as expected, average frequency increases after simplification. Also, average frequency of words is lower on the free-text side due to their complexity, and the impact of simplification is larger on the free-text side as well. We also include in Appendix A the top 20 most frequent hard words for both types of instructions for a better understanding of the dataset.

	Structured	Free-Text
Instances	3013	1541
Words	11.17	9.94
Novelties	1.44	3.67
Unchanged	200	406
Avg Freq: Compl	68536	59180
Avg Freq: Simpl	69138	65337

Table 4: Parallel Corpus Statistics

Monolingual Corpus: Next, in order to develop a corpus representative of the target language (accessible to patients), we scraped medication-related pages from five medicine-related websites[2] targeted for laypeople. We selected the five websites to be: (1) medication-related, and (2) patient-facing. This corpus C_t ($\approx 11M$ words) was used to: (1) train a language model, and (2) estimate usage frequency of words by DBF's target audience.

Human Subjects' Study: Finally, we designed an online human subjects' study (via Mechanical Turk) that presents medication instructions to participants and tests their comprehension of the instructions, before and after simplification, using multiple-choice questions.

Accordingly, we randomly choose 100 of the free-text medication instructions of varying levels of hardness (1: 29 instructions, 2: 29 instructions, and 3: 42 instructions) as measured by the number of hard (low frequency) words. Then, we simplify every instruction using DBF, and pair both the original and simplified versions of the instruction with the same multiple-choice question.

4.2 Automated Evaluation:

One standard measure for machine translation tasks is BLEU score (Papineni et al., 2002), which measures the overlap in words and phrases between a system's output and a reference output. It is also used frequently in other sequence-to-sequence problems such as text simplification. Nevertheless, BLEU has been shown insufficient for text simplification tasks due to the large overlap between the source and target vocabulary (Xu et al., 2016). Therefore, we instead consider the SARI metric, which showed better correlation than BLEU with human judgement on text simplification tasks (Xu et al., 2016). SARI, similarly to BLEU, measures the overlap of the system's output with a reference output, but also measures the amount of novelty introduced by the system. The novelty component in the metric rectifies BLEU's shortcoming in measuring the performance of a text simplification system. Moreover, we also use the PINC metric (Chen and Dolan, 2011) to measure, in isolation, the amount of novelty introduced by a system. Readability measures such as the Flesch-Kincaid index (Flesch, 1948), are not suitable for our experiments for at least 2 reasons: (1) they penalize higher word and syllable counts, when most simplifications will increase word and syllable counts such as "PO" to "by mouth", and (2) they are designed for document-level instead of sentence-level assessment.

As for the experimental setup, to avoid evaluating systems on a limited dataset size, we perform 5-fold cross validation to utilize the full dataset for evaluation. For every fold, we take 20% of the training data for tuning.

We present in Table 5 the average performance of all systems on the evaluation portion of the dataset for all five folds. Results of neural baselines were averaged over 3 runs. We also distinguish between the performance on the full dataset and the more critical subset – free-text instructions, and include the results in Table 6. For reference, we also include a baseline system that performs no change. For sample simplifications, the reader is referred to Appendix B.

Method	Supervision Type	PINC	SARI
No Change	N/A	0.00	32.83
MetaMap+CHV	Knowledge-Aware	25.84	45.64
DBF	Knowledge-Aware	19.61	55.33
Pointer-Generator	Direct Supervision	32.25	54.75
Seq2Seq-w-Att	Direct Supervision	50.81	**79.26**

Table 5: Performance of the simplification systems on all medication instructions

Method	Supervision Type	PINC	SARI
No Change	N/A	0.00	39.29
MetaMap+CHV	Knowledge-Aware	26.32	54.35
DBF	Knowledge-Aware	21.52	**56.51**
Pointer-Generator	Direct Supervision	36.34	40.01
Seq2Seq-w-Att	Direct Supervision	78.35	48.27

Table 6: Performance of the simplification systems on the free-text subset of the medication instructions

We first compare the two knowledge-aware systems: MetaMap+CHV, and DBF. First, and confirming our main hypothesis, the context-aware framework of DBF led to higher quality simplifications gaining an absolute 9.7% improvement in SARI scores over MetaMap+CHV, and a 22.5%

[2] medlineplus.gov; nia.nih.gov; umm.edu; mayoclinic.org; medicinenet.com

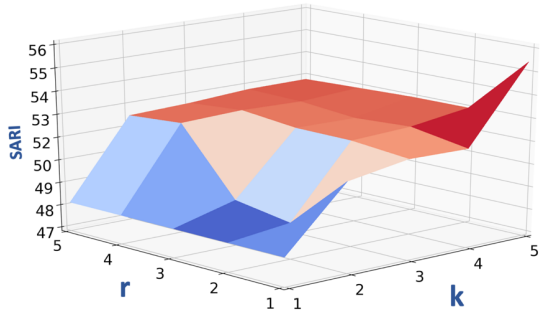


Figure 2: Effect of the hyperparameters k and r on the performance of DBF

gain compared to the No-Change case. Even though MetaMap has the added flexibility to operate on a phrase level, we attribute its comparatively lower quality to its poor WSD. Second, and by comparing PINC scores, we notice that DBF is more conservative in its changes, making it less likely to mistakenly alter key information, arguably a desired behavior in a critical domain such as healthcare. This is mainly due to it considering the identity replacement as a possible simplification, and letting the context decide whether to attempt simplification or not. These observations are also consistent on the free-text subset of the evaluation data, though we note that the gap shrinks between the two systems. We hypothesize that this is due to MetaMap+CHV committing consistent errors over one or more highly repeated terms in the structured subset of the medication instructions.

Next, we observe that, including the supervised deep learning methods, Seq2Seq-w-Attention performs significantly better than all systems. The high performance of the Seq2Seq-w-Attention is an expected result, due to the advantage of direct supervision in general, but also because direct supervision would allow it to memorize the annotator's style as well. The poor performance of the Pointer-Generator was unexpected considering its mechanism to pass easy words. Upon further inspection, we noticed two factors that degraded performance. First, the copy mechanism led to meaningless repetition of words as previously noted in the literature (See et al., 2017). Second, the copy mechanism led to copying hard words as is.

Finally, we focus our attention on how performance levels are affected when considering free-text instructions only, which are more representative of complicated health material. We notice that all the systems show more activity (higher PINC scores) in their simplifications, as these systems encounter more hard words in the original instructions. This provides further evidence that the free-text instructions constitute a critical component of the evaluation. Second, we notice that the performance of the supervised systems suffers significantly on the free-text instructions (compared to that on All Instructions), while that of the knowledge-aware (utilizing background knowledge such as UMLS and CHV) systems remain comparable, to the extent that DBF becomes the best performing approach on free-text instructions. This reflects the robustness of the knowledge-aware systems in a low-resource setting. In a setting where a sufficient parallel corpus is available, neural machine translation systems are recommended, most notably Transformer-based (Vaswani et al., 2017) for future endeavours. But in the absence of a sizable in-domain corpus, DBF achieves better performance.

Furthermore, it is notable that DBF was the best performing system despite it being the only one limited to lexical simplification. A concrete future direction would be to extend DBF's capabilities to perform phrase-level simplification.

4.3 Simplification Effects on Patient Comprehension:

We also investigated whether DBF's simplification efficacy helped improve laypeople comprehension, by measuring their ability to answer multiple choice questions (percent correct) on medication instructions before and after simplification (see Figure 3). Participants, on Amazon Mechanical Turk, were 160 adults diverse in age, cultural and academic background, and gender. 100 instructions were randomly selected from the free-text subsample of our original set of medication instructions (see Materials Section), along with their DBF simplifications and their respective multiple choice questions. Each participant read 50 instructions and answered the corresponding questions. A counterbalancing scheme ensured that each participant read 25 original instructions (as written by the physician) and 25 instructions simplified by DBF. No participant encountered both the original and the simplified version of the same instruction. Also, hardness levels of medication instructions were balanced for each participant.

The key result of this experiment was that the participants understood the simplified instructions 24.4% better than the original instructions

Medication Instruction: 10 mg PO 6pm daily.

Question: How should you take this medicine?

a) By needle
b) Into the butt hole
c) By mouth
d) On the skin
e) It was not indicated in the medication instruction

Medication Instruction: 10 mg orally 6pm daily.

Question: How should you take this medicine?

a) By needle
b) Into the butt hole
c) By mouth
d) On the skin
e) It was not indicated in the medication instruction

Figure 3: Example questions from the online human subjects study

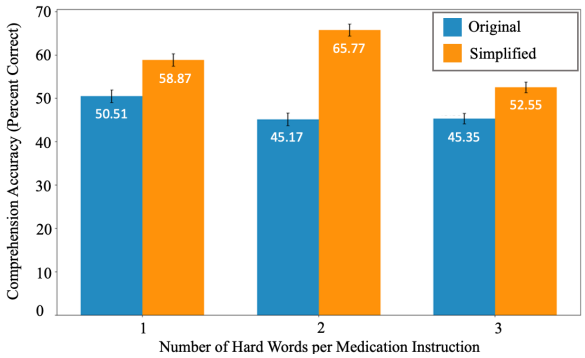


Figure 4: Impact of DBF on the different hardness levels of medication instructions

($F(1, 7973) = 112.3$, $p < .0001$, 58.23% vs 46.80%). Hardness level also influenced comprehension ($F(2, 7973) = 14.6$, $p < .0001$; see Figure 4). The simplification benefit was largest when there were two difficult words (45.63% relative), rather than one (16.55% relative) or three difficult words (15.87% relative). It is possible that having two rather than one difficult word gave more potential for DBF's simplification to increase comprehension. However, when simplification involved three words, the propagation of error led to a decrease in the quality of the simplification, and this may have negatively impacted comprehension.

5 Discussion

To better understand the functioning of the knowledge-aware systems, we study the effect of f_t on their first stage of identifying hard words. Upon tuning the systems on the validation dataset, f_t was set to 672 for both systems coincidentally. Based on Figure 5, we deduce that around 18% of words in the original instructions were attempted for translation, reflecting a high recall of hard words.

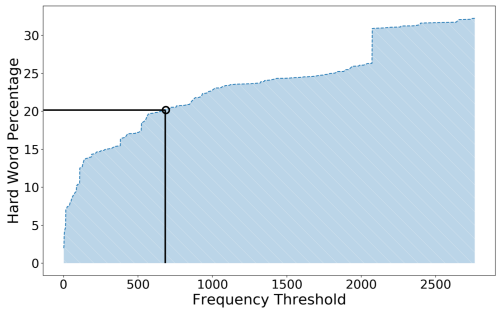


Figure 5: Percentage of words considered hard as we vary the frequency threshold

Along the same lines, we check the effect of f_t on DBF and MetaMap+CHV in terms of the two evaluation scores (see Figure 6). In terms of PINC scores, we observe an expected pattern of increase as we increase f_t for both systems. As f_t increases, both systems attempt to modify more of the original sentence (including easy words) leading to a lower overlap with reference sentences. MetaMap+CHV introduces more novelty as we increase f_t since DBF can retain easy words even if they were identified as hard, unlike MetaMap+CHV. As for SARI scores, we observe the significance of tuning the first stage, where too low of an f_t results in reduced performance due to lack of attempted translations (low PINC scores), and too high of an f_t results in reduced performance due to translating easy words. Moreover, we observe consistent enhances in performance for DBF over MetaMap+CHV for all f_t considered.

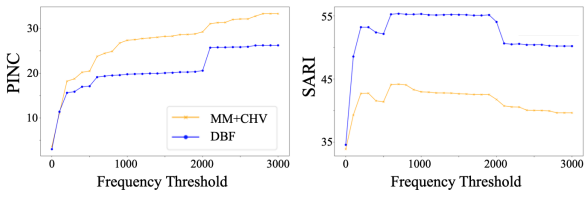


Figure 6: Effect of the frequency threshold on the performance of DBF, and MetaMap+CHV

Moving our attention to the effect of k and r on the performance of DBF, we show in Figure 2, DBF's SARI score when varying k and r from 1 to 5, and fixing f_t to 672. Our first observation is a positive trend as we increase k, particularly for $r = 1$. This shows the aptitude of the language model at selecting the best translation even when faced with a plethora of options given equal trans-

lation probabilities. As for r, we notice reduced performances for any r value different from 1. We thus conclude that the model we use for estimating translation probabilities is not benefiting translation quality. Moreover, the ranking of the concepts returned by the UMLS search function has insignificant value, when an appropriate language model is present.

6 Conclusion

Despite significant efforts to keep patients informed of their health condition, the gap in health literacy remains an issue, calling for a necessary text simplification system to bridge the gap. In this work, we suggest a context-aware framework to ensure high accuracy in such a critical domain, while also showing its positive impact on comprehension through a human subjects' study. We also conclude that, while supervised NMT methods are well-suited for the task, several healthcare subdomains lack suitable parallel corpora, which limits the performance of these supervised methods. To overcome this, we offer a knowledge-aware text simplification system to robustly operate in a low-resource setting.

Looking forward, we see great potential in adapting deep learning architectures to utilize the rich and highly curated content of UMLS, and exploring better methods or implementations to copy easy words to the target side. These two adaptations would make the high performing supervised methods of NMT less dependent on direct supervision and more generally applicable to the multiple healthcare domains. This would also address the limitation of DBF to word-level simplifications.

Acknowledgment

This work was partly supported by the IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) — a research collaboration as part of the IBM AI Horizons Network. This work was also partly supported by the Jump Applied Research for Community Health through Engineering and Simulation (ARCHES) program, UIUC/OSF Hospital, Peoria IL.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

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

Ashish Atreja, Naresh Bellam, and Susan R Levy. 2005. Strategies to enhance patient adherence: making it simple. *Medscape General Medicine*, 7(1):4.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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.

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

Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of COLING 2012*, pages 357–374.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics.

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.

Karen Davis, Stephen C Schoenbaum, and Anne-Marie Audet. 2005. A 2020 vision of patient-centered primary care. *Journal of general internal medicine*, 20(10):953–957.

Don Detmer, Meryl Bloomrosen, Brian Raymond, and Paul Tang. 2008. Integrated personal health records: transformative tools for consumer-centric care. *BMC medical informatics and decision making*, 8(1):45.

TensorFlow Developers. 2017. Tensorflow neural machine translation tutorial.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

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

Taya Irizarry, Annette DeVito Dabbs, and Christine R Curran. 2015. Patient portals and patient engagement: a state of the science review. *Journal of medical Internet research*, 17(6):e148.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Siddhartha Jonnalagadda and Graciela Gonzalez. 2010. Biosimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. In *AMIA Annual Symposium Proceedings*, volume 2010, page 351. American Medical Informatics Association.

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. American Medical Informatics Association.

Roy PC Kessels. 2003. Patients' memory for medical information. *Journal of the Royal Society of Medicine*, 96(5):219–222.

David A Kindig, Allison M Panzer, Lynn Nielsen-Bohlman, et al. 2004. *Health literacy: a prescription to end confusion*. National Academies Press.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Mark Kutner, Elizabeth Greenburg, Ying Jin, and Christine Paulsen. 2006. The health literacy of america's adults: Results from the 2003 national assessment of adult literacy. nces 2006-483. *National Center for Education Statistics*.

Gondy Leroy, James E Endicott, David Kauchak, Obay Mouradi, and Melissa Just. 2013. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of medical Internet research*, 15(7):e144.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of Medical Informatics*, 2(01):41–51.

Aqueasha Martin-Hammond and Juan E Gilbert. 2016. 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.

Nicholas Mcinnes and Bo JA Haglund. 2011. Readability of online health information: implications for health literacy. *Informatics for health and social care*, 36(4):173–189.

Arun V Mohan, M Brian Riley, Dane R Boyington, and Sunil Kripalani. 2013. Illustrated medication instructions as a strategy to improve medication management among latinos: a qualitative analysis. *Journal of health psychology*, 18(2):187–197.

Danielle L Mowery, Brett R South, Lee Christensen, Jianwei Leng, Laura-Maria Peltonen, Sanna Salanterä, Hanna Suominen, David Martinez, Sumithra Velupillai, Noémie Elhadad, et al. 2016. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: Share/clef ehealth challenge 2013, task 2. *Journal of biomedical semantics*, 7(1):43.

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

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

Yifan Peng, Catalina O Tudor, Manabu Torii, Cathy H Wu, and K Vijay-Shanker. 2012. isimp: A sentence simplification system for biomedicail text. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–6. IEEE.

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.

AK Rotegard, Laura Slaughter, and Cornelia M Ruland. 2006. Mapping nurses' natural language to oncology patients' symptom expressions. *Studies in health technology and informatics*, 122:987.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Matthew Shardlow. 2013. 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.

Yoram Singer and John C Duchi. 2009. Efficient learning using forward-backward splitting. In *Advances in Neural Information Processing Systems*, pages 495–503.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

J RC Tupper. 2008. Plain language thesaurus for health communications.

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.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Michael S Wolf, Laura M Curtis, Katherine Waite, Stacy Cooper Bailey, Laurie A Hedlund, Terry C Davis, William H Shrank, Ruth M Parker, and Alastair JJ Wood. 2011. Helping patients simplify and safely use complex prescription regimens. *Archives of internal medicine*, 171(4):300–305.

Krzysztof Wróbel. 2016. Plujagh at semeval-2016 task 11: Simple system for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 953–957.

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*, pages 1015–1024. Association for Computational Linguistics.

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.

Structured	**Count**	**Free-text**	**Count**
route	711	indications	332
tab	439	units	160
ml	319	tab	129
tabs	286	tabs	115
subcutaneous	228	x	108
nightly	191	po	97
units	164	evening	38
cap	149	friday	36
g	115	monday	35
nebulization	78	am	31
evening	74	wednesday	27
mcg	68	ml	26
inhalation	68	scale	25
caps	59	q	24
admin	46	sunday	23
spasms	40	saturday	23
puffs	38	sliding	22
wheezing	35	pt	22
transdermal	28	bedtime	21
breakfast	27	tuesday	21

Table 7: Top 20 Most frequent words identified by our system as hard for structured and free-text medication instructions

Jiaping Zheng and Hong Yu. 2016. Methods for linking ehr notes to education materials. *Information Retrieval Journal*, 19(1-2):174–188.

A Hard Words

To better understand the dataset and the transformations DBF performs, we include in Table 7 the top 20 most frequent hard words in both structured and free-text medication instructions. A word is identified as hard if its frequency in the monolingual corpus C_t is less than the tuned hyperparameter $f_t = 672$.

On the structured side, we notice a concentration of words describing the route: “subcutaneous”, “nebulization”, “inhilation”, ”transdermal”. We also find a concentration of units: “ml”, “g”, “mcg”. On the free-text side, we observe an abundance of weekdays, as well as non-standard abbreviations such as: “po” for “by mouth”, “pt” for “patient”, and “q” for “every”.

B Example Simplifications

Here, we analyze several example simplifications from the various systems in Table 8. The first example shows the incapability of Seq2Seq-w-Att to recover from a wrongly generated

Source:	Total 90 mg QAM.
Gold:	Total 90 milligrams every morning.
DBF:	Total 90 mg every morning.
MetaMap+CHV:	Total 90 mg every morning.
Seq2Seq-w-Att:	Wheeled systolic blood sugar test result is between 301 and 180,
Pointer-Generator:	Total 90 mg QAM.
Source:	Every 4-6 hours PRN thoracic back pain.
Gold:	Every 4 up to 6 hours as needed for chest back pain.
DBF:	Every 4-6 hours as needed thoracic back pain.
MetaMap+CHV:	Every 4-6 hours PRN thoracic back pain.
Seq2Seq-w-Att:	Every 6 hours as needed for back pain.
Pointer-Generator:	Every 4-6 hours PRN back pain.
Source:	Take 15 g by mouth 2 times daily as needed.
Gold:	Take 15 grams by mouth 2 times daily as needed.
DBF:	Take 15 grams by mouth 2 times daily as needed.
MetaMap+CHV:	Take 15 gram per deciliter by mouth 2 times daily as needed.
Seq2Seq-w-Att:	Take 15 grams by mouth 2 times daily as needed.
Pointer-Generator:	Take 15 grams by mouth 2 times daily as needed.
Source:	For better hearing with the ear, avoid cleaning your cerumen.
Gold:	For better hearing with the ear, avoid cleaning your earwax.
DBF:	For better hearing with the ear, avoid cleaning your wax.
MetaMap+CHV:	For better hearing with the ear, avoid cleaning your earwax.
Seq2Seq-w-Att:	Provide syringes dressings with the month, and Sunday more Lantus.
Pointer-Generator:	For UNK UNK with the UNK UNK

Table 8: Sample output simplifications from the different systems considered

first word (Wheeled). Moreover, we notice Pointer-Generator's tendency to even pass hard words. In the second example, we notice how MetaMap+CHV retains "PRN" despite being a hard word, due to MetaMap not mapping it to any UMLS concept. Additionally, we see Seq2Seq-w-Attention's mishandling of numbers since it does not have a mechanism for passing easy words. We also notice how Pointer-Generator wrongly eliminates words (thoracic) essential to the meaning of the sentence. On the other hand, the next example shows the shortcomings of MetaMap+CHV's disambiguation algorithms, while DBF was able to accurately map "g" to "grams". Whereas both deep learning methods get the full mark on this example since it is a structured medication instruction.

The last point we would like to address is the last example in Table 8. This example, contrary to the previous ones, was not taken from the medication instruction dataset, but rather created by us to portray a complicated sentence from another medical domain, in this case: online health tips. As can be seen from the systems' outputs, the robustness of knowledge-aware systems is evident in comparison to the supervised deep learning methods, which are completely off the mark.