PharmMT: A Neural Machine Translation Approach to Simplify Prescription Directions

Jiazhao Li,[1] **Corey Lester,**[2] **Xinyan Zhao,**[1] **Yuting Ding,**[2]
Yun Jiang,[3] **V.G.Vinod Vydiswaran**[4,1]

[1]School of Information; [2]Department of Clinical Pharmacy, College of Pharmacy;
[3]Department of Systems, Populations, and Leadership, School of Nursing;
[4]Department of Learning Health Sciences, Medical School
University of Michigan, Ann Arbor, MI
{jiazhaol,lesterca,zhaoxy,dingyt,jiangyu,vgvinodv}@umich.edu

Abstract

The language used by physicians and health professionals in prescription directions includes medical jargon and implicit directives and causes much confusion among patients. Human intervention to simplify the language at the pharmacies may introduce additional errors that can lead to potentially severe health outcomes. We propose a novel machine translation-based approach, PharmMT, to automatically and reliably simplify prescription directions into patient-friendly language, thereby significantly reducing pharmacist workload. We evaluate the proposed approach over a dataset consisting of over 530K prescriptions obtained from a large mail-order pharmacy. The end-to-end system achieves a **BLEU score of 60.27** against the reference directions generated by pharmacists, a 39.6% relative improvement over the rule-based normalization. Pharmacists judged **94.3%** of the simplified directions as usable as-is or with minimal changes. This work demonstrates the feasibility of a machine translation-based tool for simplifying prescription directions in real-life.

1 Introduction

Adverse drug events stemming from medication errors are a vital cause of concern in patient care and are estimated to cost US$42 billion annually or roughly 1% of total global expenditure. In the US alone, medication errors cause one death every day and are responsible for over 700,000 visits to the emergency department and over 100,000 hospitalizations each year (Budnitz et al., 2006, 2011; WHO, 2017).

One of the frequent sources of medication errors in the US is the directions on the 1.91 billion electronic prescriptions (e-prescriptions) transmitted annually (Moniz et al., 2011; Odukoya et al., 2014, 2015). The style and language used in e-prescriptions are highly variable and often filled with medical jargon. For example, in a recent study (Yang et al., 2018), the authors noted that the direction "*Take 1 tablet by mouth once daily*" was represented in 832 different ways. The study also found that 10.1% of e-prescriptions contained incorrect or confusing language. Pharmacists play a vital role as intermediaries between physicians and patients by translating the rich medical jargon in the e-prescriptions written by physicians to patient-comprehensible directions on the prescription labels printed on pill bottles. However, human translation is time-consuming and subject to errors, potentially leading to medication errors and other patient safety risks due to prescription ambiguity.

In this paper, we propose a machine translation-based system, called PharmMT, to simplify the e-prescription directions authored by physicians into patient-friendly language. The system aims to automate the translation and normalization of e-prescription directions and reduce the pharmacists' overall workload. We investigate multiple neural network-based models, including transformer-based models and bi-directional LSTM models, rule-based approaches, and a hybrid model combining neural network-based models with a rule-based backoff. We train and evaluate the proposed system over a dataset of over 530K paired e-prescriptions and their human-translated text, obtained from a large mail-order pharmacy. Using automated measures to evaluate machine translation output, we compare the PharmMT system against a rule-based approach developed based on domain knowledge from pharmacists. Our results show that PharmMT performs significantly better than the rule-based baseline. Manual evaluation by pharmacists also shows a high potential to directly apply the proposed approach in pharmacies to translate e-prescription directions.

The contributions of this work are:

1. We develop a neural machine translation model for simplifying e-prescriptions and build an **end-to-end system** to generate normalized, patient-friendly, and usable translations. The model achieves a BLEU score of 60.27 against the reference directions generated by pharmacists, and 94.3% of simplified prescriptions are judged as usable as-is or with minimal changes by pharmacists.

2. To the best of our knowledge, our work is the **first systematic effort** to study neural network models to simplify e-prescription directions. We also developed a rule-based approach as the baseline of this task.

3. Our work adds **additional insights** into the limitations of purely automated evaluation metrics of machine translation for domain-adaptive tasks, and offers alternative modes of evaluation.

2   Related work

Prior work on automated approaches for translating e-prescription directions mainly focused on information extraction models relied on handwritten rules or linguistic signals found in prescription free text. Tools such as MetaMap (Aronson and Lang, 2010) and MedLEE (Friedman et al., 1996) extract and organize clinical information in text documents using external knowledge sources, such as the Unified Medical Language System (UMLS). Other systems, such as FABLE (Tao et al., 2018), employed a conditional random fields-based model to recognize medication entities. Other researchers have proposed rule-based approaches to normalize and simplify directions using task-specific knowledge such as common abbreviations used in prescriptions (Qenam et al., 2017; Kandula et al., 2010).

While machine translation-based approaches have not yet been proposed for translating e-prescription directions, prior works such as (Yolchuyeva et al., 2018; Shardlow and Nawaz, 2019; Van den Bercken et al., 2019) have suggested solving machine translation tasks without the need for explicitly-defined rules. Neural machine translation (NMT) models have been shown to be able to learn contextual rules automatically from large corpora and produce higher quality translations (Bahdanau et al., 2014; Wu et al., 2016b; Lee et al., 2017). Other researchers, such as (Aw et al., 2006; Xu et al., 2016), have shown that while statistical machine translation methods mainly focused on lexical rules to minimize sentence complexity,

NMT models could capture richer syntactic information (Shi et al., 2016).

Researchers studying deep neural network models have explored multiple encoder-decoder frameworks, such as Transformer-based networks (Vaswani et al., 2017), and Recurrent Neural Networks (RNN), including Long Short-Term Memory (LSTM) models (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Chung et al., 2014). RNN units have been used to encode source sentences into fixed-length representations and then decoded into reference sentences (Cho et al., 2014). Models with deep LSTM-based RNN units have shown the benefits of using deeper structure (Wu et al., 2016a). In other works, (Gehring et al., 2017) introduced a convolution neural network with an attention-based mechanism to learn long-range dependency. In recent work, (Vaswani et al., 2017) developed a novel Transformer-based architecture using attention mechanism without recurrence or convolution, resulting in a state-of-the-art performance for many natural language processing tasks, such as recognizing textual entailment, sentiment analysis, and natural language inference. (Raffel et al., 2019; Lan et al., 2019; Devlin et al., 2018)

3   Simplifying e-prescription directions

We frame the challenge of simplifying e-prescription directions as a machine translation task from physician-authored directions ("**source**") to patient-facing text authored by pharmacists ("**reference**"). This monolingual translation task focuses on replacing highly-abbreviated medical jargon with patient-friendly vocabulary, simplifying cryptic expressions, and normalizing them so that they can be used with minimal changes by the pharmacists. Table 1 (top panel) shows three examples of e-prescriptions and their corresponding simplified directions.

E-prescription directions consists of specific components related to the prescribed drug, viz., *dosage*, *form*, *route*, *duration*, *frequency*, and *reason* for prescribed use. For example, the e-prescription "*2 puffs orally q 4 hrs x90 dys wheeze*" specifies that the patient should inhale 2 (*dosage*) puffs (*form*) by mouth (*route*) every 4 hours (*frequency*) for 90 days (*duration*) for wheezing (*reason*). While not all components are present in every e-prescription direction, some components are critical and need to be stated explicitly. The *name* and

E-prescription direction	Simplified direction
2 puffs orally q 4 hrs x90 dys wheeze	Inhale 2 puffs by mouth every 4 hours for 90 days for wheeze
1 g vaginal mon/tu/th/fr	Insert 1 gram vaginally monday, tuesday, thursday and friday
as needed prn; 1 po qd prn	Take 1 tablet by mouth **once a day** as needed
oral one tab po qd prn	Take 1 tablet by mouth **every day** as needed
	Take 1 tablet by mouth **daily** as needed
Drug name and strength	**Normalized drug name and strength**
albuterol 90 mcg/inh inhalation aerosol	PROAIR HFA AER
0.1 mg/g vaginal cream	ESTRADIOL CRE 0.01%
traMADol 50 mg tablet	TRAMADOL HCL TAB 50MG

Table 1: **Top**: Example pairs of e-prescription directions and corresponding simplified text. Variations of these directions exist on both sides. **Bottom**: Drug name and strength information in the original and normalized forms.
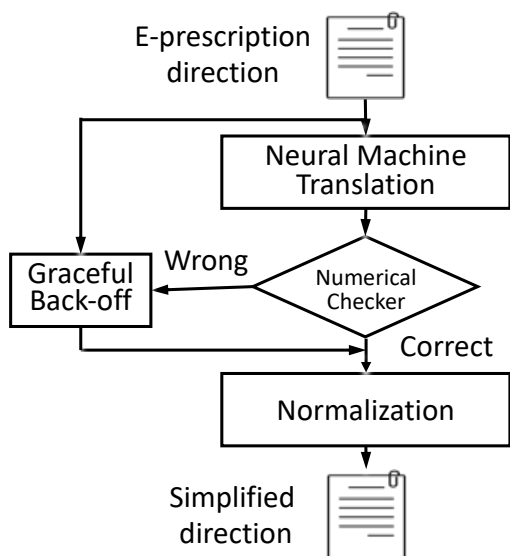

Figure 1: Schematic diagram of the PharmMT system

strength of the prescribed drug are also available as auxiliary information. Examples of drug names and strengths are shown in the bottom panel of Table 1.

While we formulate the challenge as a machine translation task in the pharmacy domain, key desiderata for the automated approach are to preserve the accuracy and consistency of the critical components in a prescription. To achieve this, we develop an end-to-end system called **PharmMT**, consisting of three stages: neural machine translation, numerical check and graceful backoff, and normalization, as depicted in Figure 1.

3.1 Neural Machine Translation (NMT)

The primary component of the proposed approach is a sequential model that "translates" physician-authored e-prescription text to normalized, patient-friendly language using an NMT framework. NMT models map a source sequence, $\mathbf{x} : x_1, x_2, \ldots, x_n$ into a reference sequence, $\mathbf{y} : y_1, y_2, \ldots, y_m$ by maximizing the conditional probability $p(\mathbf{y}|\mathbf{x})$ using an Encoder-Decoder framework (Neubig, 2017). One such model is a recurrent sequence-to-sequence model, which consists of a bidirectional LSTM model (Schuster and Paliwal, 1997) with global attention as the encoder and a forward-sequence LSTM model (Hochreiter and Schmidhuber, 1997) as the decoder. Both encoder and decoder stages are configured as multi-layer models, with a hidden state in each layer, to sufficiently capture the deep semantic components in the e-prescription (Barone et al., 2017).

To compare against the performance of the recurrent sequence-to-sequence model, we also trained an attention-based transformer model (Vaswani et al., 2017). Position embedding was enabled to capture sequence information and provide similar architectural complexity as a recurrent network. Both models were developed using the Open-NMT framework (Klein et al., 2017, 2018), and the dropout probability was adjusted to prevent over-fitting (Srivastava et al., 2014). Additional experimental details can be found in Section 4.2.

3.1.1 Augmenting auxiliary information

As described in Section 3, e-prescriptions contain auxiliary information on the drug name and strength. While the primary task is to simplify just the direction, access to the auxiliary information may help distinguish directions based on the context associated with drugs. We hypothesize that the auxiliary information will improve the neural machine translation models to simplify directions. This hypothesis also matches with real-life information available to pharmacists. To test this hypothesis, we prepend the drug name and strength

information to the "source" direction before training the models. We evaluate the updated model on the original task of simplifying just the directions.

3.1.2 Pre-trained word embeddings

The input representation is a pre-trained word embedding layer to allow for similar representations of words in similar contexts. Word embeddings can capture fine-grained semantic and syntactic word relationships and, in turn, allow for a better initialization for gradient optimization. We explored static pre-trained embedding models and compared them against a randomly-initialized representation vector. The first one was the general-domain GloVe word embeddings, pre-trained on the Wikipedia and Gigaword corpora (Pennington et al., 2014). The second one was clinical domain-adaptive word embeddings that we trained on MIMIC-III, a large corpus of clinical notes (Johnson et al., 2016) and a dataset of pharmacy directions (Zheng et al., 2020). We hypothesize that domain-adaptive word embeddings would outperform both the general-domain embeddings and randomly-initialized vector embeddings.

3.1.3 Learning ensemble models

The primary motivation of ensemble learning over neural network models is to improve the final model's robustness to the variations introduced in parameterized modules because of dropout probability and random seeding. In an ensemble model, the final distribution of the output dictionary is computed by averaging the output distributions from the trained models in the ensemble during the inference phase.

3.2 Numerical check

Once the machine translation module generates the simplified candidate directions, the candidates are checked for consistency of key components of the prescription. Numerical components – including *dosage*, *frequency*, and *duration* – are critical in prescriptions. Medication under-dosage often leads to poorer health outcomes, while over-dosage can be severe, even fatal.

The correctness of the numerical components in the simplified directions is checked by comparing against the source e-prescription. We incorporated two different numerical checking strategies – **Token-based** and **NER-based**. In the token-based checking, all numeric tokens that appear in the simplified direction were checked against numeric

Fields	Original	Normalized
Action	(missing)	take inject inhale . . .
Dosage	one and half 1 1/2 one (1)	1.5 1.5 1
Form	tab, tabs cap in, inj, injctor	tablet capsule injection
Route	orally, by oral sq, subcutaneous	by mouth under the skin
Frequency	qd bid	every day twice a day
Duration	x3 week	for 3 weeks

Table 2: Sample normalization and simplification rules

tokens in the source direction. The bag of tokens approach helps tag any simplified direction that "makes up" numeric values in a key component.

On the other hand, the token-based checker can also generate faulty consistency claims, for example, when the simplified direction swaps a *dosage* term with the *frequency*. To overcome this, we incorporated a pre-trained medication NER model (Zhao and Vydiswaran, 2020) to tag *dosage*, *frequency*, and *duration* components in both source and simplified directions, and compared them component-by-component. The NER model was trained over a medication extraction task (Henry et al., 2020) and achieved an overall F1 score of 0.9571 over all medication components.

3.3 Graceful backoff

If the simplified direction is deemed consistent after the numeric check, the direction is considered as the final candidate for normalization. However, if the numeric check fails, the NMT output is discarded and the original source direction is used as the final candidate. This graceful backoff represents a trade-off between information accuracy and good language model performance.

3.4 Normalization

Before the candidate direction from the numerical check phase is finalized, the candidate text undergoes pharmacy-specific post-processing and simplification. Two pharmacists identified common linguistic patterns in pharmacy directions, which were coded into normalization rules. Highly-abbreviated

Data set	Train	Validation	Test
Original	318,594	79,648	132,747
Deduplicated	318,594	15,625	36,652

Table 3: Data set sizes after removing duplicates.

Embeddings	Vocab. size	Source (n = 11,643)	Reference (n = 7,358)
Clinical	498,677	10%	6.69%
General	400,000	73.75%	60.29%

Table 4: Out-of-vocabulary (oov) ratio for pre-trained word embedding models. General-domain word embeddings show a large oov gap against our dataset.

medical jargon was replaced, for example, by replacing the Latin term *bid* with its synonymous phrase *twice a day*. Action verbs appropriate for the form of the drug, such as *inject* (syringe), *inhale* (nebulizer), and *take* (capsule), were added. Numerical values in words or fractions were converted to digits (e.g. *1 1/2* was converted to *1.5*). Abbreviations and other common medical expressions were normalized into standard and patient-friendly variants, e.g. *inj* or *injector* to *injection*. Overall, the normalization step included more than 300 rules. Table 2 shows examples of these normalization and simplification rules. The normalization module also served as the rule-based baseline.

4 Experimental setup

4.1 Data set description

The e-prescription corpus used in this study consists of all e-prescriptions dispensed by an online outpatient mail-order pharmacy from their dispensing software from January 2017 to October 2018. The corpus consists of 530,988 e-prescriptions received from 65,139 unique physicians from all fifty US states (Zheng et al., 2020).

Each e-prescription direction in the data set is paired with the corresponding simplified text authored by a mail-order pharmacy team member. Table 1 shows some example e-prescription directions and the corresponding simplified text.

In addition to the e-prescription, each direction in the corpus also contains auxiliary information about the name and strength of the drug. However, similar to the directions, physician-authored information often included chemical or ingredient names for the drug, while the pharmacist-translated direction contained generic drug names or brand names.

In all, there are 120,402 unique e-prescription directions and 83,823 unique pharmacist-authored directions in the dataset. The difference in these numbers is due to the diverse writing styles of physicians and pharmacists. On an average, there were 6.33 e-prescription directions mapped to a single pharmacist-authored direction; while one e-prescription direction mapped, on an average, to

4.41 different pharmacist-authored directions.

We split our data into train, validation, and test sets. To avoid information leak during the evaluation, we remove all duplicates appearing in more than one set, but retain those appearing within a single set. Table 3 summarizes the distribution of the instances over the three sets. None of the instances in the validation or test sets were used during the training phase.

4.2 Training process

To prepare the data for training, the source e-prescription directions and reference pharmacist-authored directions are prepended with the drug name and strength, as described in Section 3.1.1.

Pre-trained word embeddings: We tested two static pre-trained word embeddings – general-domain GloVe embeddings and a second one trained explicitly over two clinical domain corpora – against a randomly-initialized representation. The out-of-vocabulary rate is shown in Table 4. While only 10% of the source words and 6.69% of the target words had no word embeddings in the clinical-domain word embeddings, the out-of-vocabulary ratio was 7 to 9 times higher in the general-domain word embeddings.

Model configuration: The number of layers in the encoder and decoder stages of the Bi-LSTM, LSTM-based, and Transformer-based models was empirically chosen from the set $\{2, 4, 6, 8\}$. The length of hidden states was chosen from the set $\{128, 256, 512\}$. In the following description, we denote the number of layers (i.e., Transformer blocks) as L, the hidden size as H, and the number of self-attention heads as A, consistent with the BERT notation (Devlin et al., 2018). After choosing the hyper-parameter based on highest BLEU score, we primarily report results on two best performance architectures: **Bi-LSTM/LSTM-based** model (L = 4, H = 256, Dropout = 0.4) and **Transformed-based** model (L = 4, H = 128, A = 2, Dropout = 0.2). We trained our settled models

E-prescription direction	BLEU	METEOR
location: both eyes. 1 drop into each eye at bedtime. *[Source]* apply 1 drop into each eye at bedtime. *[NMT]* instill 1 drop into both eyes at bedtime. *[Reference]*	0	0.77
1 puff once daily per dr. jones. *[Source]* inhale the contents of one capsule by mouth once daily using handihaler. *[NMT]* use 1 inhalation by mouth once daily. *[Reference]*	0.38	0.90
Evaluation Limitation		
take **1** tablet by mouth every morning **and** every evening. *[Reference]*		
take 1 tablet by mouth every morning **&** every evening. *[NMT$_1$]*	0.70	0.91
take **10** tablets by mouth every morning and every evening. *[NMT$_2$]*	0.70	0.98

Table 5: **Top**: Comparison of BLEU and METEOR scores on two examples. METEOR scores show higher correlation with human judgement due to flexible matching. **Bottom**: Both variations of NMT model outputs have similar BLEU and METEOR scores. However, The first replaced an '*&*' with '*and*', while the second had a critical dosage error ('*10*', instead of '*1*').

on a single Tesla V100. It took 4.67 hours to train the LSTM-based models with 12.66 million parameters, and 3.33 hours for the Transformer-based model with 9.27 million parameters.

4.3 Evaluation metrics

Automatic evaluation: We evaluated the translation model using two automated candidate-reference comparison metrics: BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). The BLEU-4 score is the most popular metric used to evaluate the similarity between the candidate text and human reference in machine translation tasks (Sutskever et al., 2014; Koehn et al., 2003; Lipton et al., 2015). It is computed as the geometric mean of precision of unigram, bigram, trigram, and 4-gram matches between the candidate and reference texts. In contrast to the ngram-based overlap in BLEU-4, the METEOR score considers unigram alignment between candidate and reference texts. This results in more flexible matching, including stem match and synonym match using WordNet (Miller, 1995). METEOR scores are known to achieve a better correlation with human judgment (Banerjee and Lavie, 2005).

Sample comparisons of two metrics in Table 5 (top panel) show the limitation of using the BLEU score to compare pharmacy instructions, while also showing the feasibility of using METEOR score as a viable alternative. The bottom panel of Table 5 highlights the limitation of both metrics in checking the consistency of prescription components, and is discussed in more detail in Section 6.4.

NMT Models	BLEU	METEOR
TransF-TransF : Random	59.09±0.08	80.10±0.06
TransF-TransF : GloVe	62.36±0.10	80.45±0.05
TransF-TransF : Clinical	63.23±0.10	80.89±0.17
+ Ensemble	64.19±0.31	81.05±0.12
BiLSTM-LSTM : Random	64.63±0.06	81.14±0.30
BiLSTM-LSTM : GloVe	65.78±0.07	81.62±0.11
BiLSTM-LSTM : Clinical	66.02±0.03	82.32±0.20
+ Ensemble	**66.61±0.29**	**82.73±0.10**
Rule-based baseline	43.19	68.59

Table 6: Comparison of NMT models with different word embeddings approaches, against a rule-based baseline. An ensemble BiLSTM-LSTM model with clinical word embeddings achieves the highest BLEU and METEOR scores.

Manual evaluation: While automated metrics such as BLEU and METEOR are commonly used to evaluate machine translation tasks, they do not sufficiently evaluate the clinical usability of the candidate texts. Hence, in addition to the automated evaluation, we asked two pharmacist trainees to evaluate 300 pairs of e-prescription directions and corresponding simplified directions randomly sampled from the test set. The pharmacist trainees were asked to classify the direction pairs into one of three categories – **Correct:** all information in the simplified direction was correct; **Missing:** the simplified direction was correct, but missed some essential information; and **Wrong:** the simplified direction contained fundamental errors that needed to be corrected. A pharmacist expert resolved labeling disagreements. This manual evaluation simulates the human effort undertaken in real-life at the pharmacies to simplify e-prescription directions.

#	Issues	E-prescription	Rule-based baseline	PharmMT
1	Contextual ambiguity	1/2 tab bid orally **90**.	Take 0.5 tablet by mouth twice a day **90**	Take 0.5 tablet by mouth twice a day **for 90 days.**
2	Word sense ambiguity	**spray** 1 **spray(s)** 4 times a day by intranasal route as needed for 90 days.	Use **spray** 1 **spray** 4 times a day in the nose route as needed for 90 days.	Use 1 **spray** in the nose 4 times a day as needed .
3	Re-ordering components	tablets by mouth daily; **3.5 tab 7 mg.**	Take tablets by mouth daily; **3.5 tablet 7 mg**.	Take **3.5 tablets** by mouth daily.
4	Handling misspelling	one tablet by mouth **oce** daily.	Take one tablet by mouth **oce** daily.	Take 1 tablet by mouth **once** a day.
5	Informal abbreviations	1 puff **aero pow br act** bid.	Inhale 1 puff **aero pow br act** twice a day.	Inhale 1 puff **by mouth** twice a day.

Table 7: Examples comparing PharmMT model against a rule-based baseline, highlighting potential issues with rule-based approaches.

5 Results

We present our results first by evaluating the NMT module by comparing the two proposed architectures using the automated BLEU and METEOR scores. After finalizing the best performing NMT model, we compare its results against a rule-based baseline and demonstrate the significance of the individual stages through an ablation study. Finally, we report the performance of the end-to-end system based on the manual evaluation.

5.1 Evaluating Neural Machine Translation module

We compared two classes of NMT models – **Transformer-based** model and **BiLSTM/LSTM-based** model under different pre-trained word embeddings. The results are summarized in Table 6. The reported values are mean and standard deviation over ten independent iterations of training and validation using the same model hyper-parameters.

Both automated metrics were consistent in their ranking of the systems. On both metrics, LSTM-based models outperformed transformer-based models. One possible explanation for these results is that although transformer-based models are better at capturing long-range dependencies, their advantage is nullified by the relatively short sentences in this task. The average length of e-prescription directions is 10.42±4.55 tokens.

Models using the pre-trained clinical-domain word embeddings led to the highest performance on both metrics and were statistically better than models that used general-domain word embeddings. Models using the randomly-initiated word representation performed the worst. The model performance improved further when ensemble learning was applied. The ensemble BiLSTM / LSTM model with clinical domain-adaptive word embeddings achieved the highest overall BLEU score of 66.61±0.29 and the METEOR score of 82.73±0.10.

We also note that in Section 3.1.2, we stated our hypothesis that domain-adaptive word embeddings would outperform both the general-domain embeddings and randomly-initialized vector representations. This hypothesis was shown to be valid in our results in Table 6. Instead of the static, but domain-adaptive, word embeddings explored in this work, other alternatives such as contextual word embeddings could also be used, including BioBERT (Lee et al., 2019) and Clinical-BERT (Huang et al., 2019).

5.2 Comparison to rule-based baseline

Next, we compared the translated directions generated by the PharmMT model against a rule-based baseline in which the e-prescriptions are passed directly through the Normalization module to produce the outputs. Table 7 shows some examples of rule-based translation against PharmMT output. These examples highlight three potential issues of rule-based systems, viz., handling ambiguity, re-ordering of direction components, and sensitivity to misspelled tokens and abbreviations.

In the first example, the rule-based approach fails to recognize '*90*' as *duration* without any contextual clues, whereas PharmMT correctly normalizes it as '*for 90 days*'. Similarly, in the second example, the rule-based approach fails to identify '*spray*' as both an action verb and the *form* token. This highlights the difficulty in manually curating rules to cover all ambiguous cases.

Second, while rule-based approaches are very effective when the order of prescription components is as expected, they do not handle reordered components well. The third example in Table 7 shows one such instance, where unlike the rule-based approach, the PharmMT model correctly reorders the

Model variations	BLEU	METEOR
(1): Rule-based Baseline (Normalizer)	43.19	68.59
(2): Best NMT model	67.21	82.90
(3): Best NMT model w/o Auxiliary	63.59	80.83
(4): (2) + Backoff	66.24	79.59
(5): **PharmMT** [same as (4) + (1)]	**60.27**	**76.11**
(6): (5) - Backoff [same as (2) + (1)]	60.29	76.20

Table 8: Ablation study results showing model performance taking components out one-at-a-time, from the best NMT model to the end-to-end system, PharmMT (in bold).

direction components by inserting *dosage* tokens ('*3.5 tablets*') before the *form* token, '*tablet*'.

Finally, rule-based approaches are also more sensitive to misspelled tokens and informal abbreviations, compared to PharmMT, as shown in the final two examples in Table 7.

5.3 Results of the ablation study

The best performing NMT model achieved a BLEU score of 67.21 and a METEOR score of 82.90. After finalizing the best NMT model, the significance of the remaining components is shown using an ablation study. The results are summarized in Table 8.

Removing auxiliary information: Without augmenting auxiliary drug information, the performance of the NMT model drops by 5.4% on the BLEU score and 2.5% on the METEOR score. These results indicate that augmenting directions with auxiliary drug information helps improve the overall performance, as hypothesized in Section 3.1.1.

Graceful backoff and normalization: Adding the NER-based numeric check and resorting to graceful backoff, when necessary, reduces the overall scores on the automated metrics (BLEU: 66.24, METEOR: 79.59). Adding normalization decreases both metrics even further (BLEU: 60.27, METEOR: 76.11). This performance drop is because the normalization-based approaches tend to produce reference direction texts influenced heavily by preference rules from pharmacists. So, while the resultant directions are more readable, patient-friendly, and preferred by pharmacists, they do not accurately reflect the preferred style in the original reference corpus. We expand on this further in Section 6.3.

	Dosage (n=22,099)	Frequency (n=8,256)	Duration (n=2,879)	Combined (n=23,237)
NER	2,570	1,451	238	4,027
Token	-	-	-	1,390

Table 9: Number of instances marked as inconsistent by the numeric checkers. NER-based checker flagged more inconsistent instances than NER-based checker.

5.4 Manual evaluation of end-to-end system

The final output of the end-to-end system was evaluated by domain experts. Of the 300 pairs of e-prescription directions and their corresponding simplified texts, 86.7% (n=260) were labeled as Correct, 7.6% (n=23) as Missing; and 5.7% (n=17) as Wrong. The missing errors were primarily related to missing adjectives and adverbs (e.g., *transdermal*, *slowly*), or typographic omissions, such as brackets. The incorrect errors were primarily related to special directions (e.g. taking medications *before meals*, *with food*), formatting issues with dosage (e.g. *10-12*), or complex directions based on days of the week (e.g. *every day except Sundays*). These results show that in **94.3% instances**, the simplified output can be used as-is or after minimal changes to add the missing elements.

6 Discussion

6.1 Error analysis

We further analyzed all non-'Correct' instances (n=17; 5.7%) identified during the manual evaluation. A major class of errors was complicated instruction and language patterns in the e-prescription. On average, these directions were 16.86 words long, compared to an average of 12.57 words for 'Correct' instances. For example, one direction that was not simplified correctly was: *30 units with meals plus ssi 150-200 2 units, 201-250 4 units, 251-300 6 units, 301-350 8 units, greater than 351 10 units.* This direction instructed patients to change dosage depending on the sliding scale for insulin (*ssi*).

Based on the hypothesis that shorter directions will have simpler language patterns, we evaluated a subset of test instances (n=11,977; 32.6%) that were under 12 words long. This 'shorter length' subset achieved an aggregate BLEU score of 71.14, while the complementary 'longer length' subset managed an aggregate BLEU score of 64.02.

Against normalized reference	BLEU	Ratio
Rule-based Baseline (Normalizer)	47.60	86.58%
Best NMT model	62.68	-
Best NMT model + Normalizer	71.33	30.81%
PharmMT	71.14	36.01%
Reference (upper bound)	82.48	42.36%

Table 10: Performance against normalized reference

6.2 Numeric checker and Graceful backoff

We further investigated the performance of the numeric checker. The number of instances marked as inconsistent by the two approaches is summarized in Table 9. The stricter, NER-based numeric checker flagged 4,027 (17.33%) instances as inconsistent, while the token-based checker flagged only 1,390 (5.98%) of instances as inconsistent.

6.3 Need for normalized reference

The normalization process is heavily influenced by the style preferred by pharmacists coding the normalization rules. Since the preferred style of the team that generated the original reference differed from the expert pharmacists in our team, the original reference data were themselves not normalized. This led to a reduction in BLEU scores when normalization was added (see Table 8).

To understand how well our trained models could perform on this corpus, we created a normalized version of the reference corpus. Using this as the gold reference, the original reference corpus has a BLEU score of 82.48, and that of the PharmMT system was 62.68 (see Table 10). The table also shows the ratio of test instances that are normalized to indicate how *close* the output is to the normalization rules: the lower the ratio, the closer it is. The results indicate that the NMT model learned more latent rules from the train data set than the hand-crafted normalization rules, while having a higher BLEU score.

6.4 Limitation of BLEU and METEOR

Automated metrics such as BLEU and METEOR can only evaluate the translation results using a linguistic, token-level approach, but fail to capture the nuanced semantic-level information. However, in prescription directions, different words contain **unequal** useful information, and hence should be re-weighted during the evaluation process. As we noted in Section 3.2, consistency of key information is vital for patient safety. For example, in the translation shown in bottom of Table 5, both machine translation outputs NMT_1 and NMT_2 have only one token different from the reference. But, NMT_1 is labeled as 'Correct' in the manual evaluation while NMT_2 is labeled as 'Wrong' because of a serious error on *dosage*. However, both translations got similar BLEU and METEOR scores. In the future, we will focus on improving information consistency while maintaining high model performance.

7 Conclusion

We proposed and developed a machine translation-based approach, called PharmMT, to simplify e-prescription directions. We systematically evaluated the individual stages and the overall approach over a large mail-order pharmacy data corpus. Our results showed that an ensemble model with a bi-directional LSTM encoder and an LSTM decoder, trained over a clinical-domain word embedding representations, achieved the best overall BLEU score of 60.27. NER-based numeric check and graceful backoff ensure information consistency and the normalization stage helps generate patient-friendly directions. Qualitative evaluation by domain experts showed that 94.3% of the simplified directions could be used as-is or with minimal changes. These results indicate that the proposed approach could be deployed in practice to automate the simplification of prescription directions.

Acknowledgment

References

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.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 33–40. Association for Computational Linguistics.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*.

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. ACM.

Daniel S. Budnitz, Maribeth C. Lovegrove, Nadine Shehab, and Chesley L. Richards. 2011. Emergency hospitalizations for adverse drug events in older Americans. *New England Journal of Medicine*, 365(21):2002–2012.

Daniel S. Budnitz, Daniel A. Pollock, Kelly N. Weidenbach, Aaron B. Mendelsohn, Thomas J. Schroeder, and Joseph L. Annest. 2006. National surveillance of emergency department visits for outpatient adverse drug events. *JAMA*, 296(15):1858–1866.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Carol Friedman, Lyudmila Shagina, Socrates A Socratous, and Xiao Zeng. 1996. A web-based version of medlee: A medical language extraction and encoding system. In *Proceedings of the AMIA Annual Fall Symposium*, page 938. American Medical Informatics Association.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

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

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

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*, 3:160035.

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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Zachary C Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Thomas T Moniz, Andrew C Seger, Carol A Keohane, Dianel Lew Seger, David W Bates, and Jeffrey M Rothschild. 2011. Addition of electronic prescription transmission to computerized prescriber order entry: effect on dispensing errors in community pharmacies. *American Journal of Health-System Pharmacy*, 68(2):158–163.

Graham Neubig. 2017. Neural machine translation and sequence-to-sequence models: A tutorial. *CoRR*, abs/1703.01619.

Olufunmilola Odukoya, Loren Schleiden, and Michelle Chui. 2015. The hidden role of community pharmacy technicians in ensuring patient safety with the use of e-prescribing. *Pharmacy*, 3(4):330–343.

Olufunmilola K Odukoya, Jamie A Stone, and Michelle A Chui. 2014. E-prescribing errors in community pharmacies: exploring consequences and contributing factors. *International journal of medical informatics*, 83(6):427–437.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389, Florence, Italy. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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.

Carson Tao, Michele Filannino, and Özlem Uzuner. 2018. Fable: A semi-supervised prescription information extraction system. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1534. American Medical Informatics Association.

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

WHO. 2017. Medication without harm: WHO global patient safety challenge.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, and Taku Kudo and. 2016a. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016b. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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.

Yuze Yang, Stacy Ward-Charlerie, Ajit A. Dhavle, Michael T. Rupp, and James Green. 2018. Quality

and variability of patient directions in electronic prescriptions in the ambulatory care setting. *Journal of managed care specialty pharmacy*, 24(7):691–699.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2018. Text normalization with convolutional neural networks. *International Journal of Speech Technology*, 21(3):589–600.

Xinyan Zhao and V.G.Vinod Vydiswaran. 2020. LEx-Transfer: Break data discrepancy with lexically enhanced extraction transfer learning for prescription parsing. *Under review at the Journal of the American Medical Informatics Association*.

Yifan Zheng, Yun Jiang, Michael P. Dorsch, Yuting Ding, V.G. Vinod Vydiswaran, and Corey A. Lester. 2020. Work effort, readability and quality of pharmacy transcription of patient directions from electronic prescriptions: a retrospective observational cohort analysis. In *BMJ Quality & Safety*.