Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification

Viraj Adduru[1*], Sadid A. Hasan[2], Joey Liu[2], Yuan Ling[2], Vivek Datla[2], Kathy Lee[2], Ashequl Qadir[2], Oladimeji Farri[2]
[1]Rochester Institute of Technology, Rochester, NY, USA
[2]Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA
{vra2128}@rit.edu,
{firstname.lastname, kathy.lee_1, dimeji.farri}@philips.com

Abstract

A paraphrase is a restatement of a text while retaining the meaning. Clinical paraphrasing involves restatement of sentences, paragraphs, or documents containing complex vocabulary used by clinicians. Paraphrasing can result in an alternative text that is either simple or complex form of the original input text. Simplification is a form of paraphrasing in which a sentence is restated into a linguistically simpler sentence yet retaining the meaning of the original sentence. Clinical text simplification has potential applications such as simplification of clinical reports for patients towards better understanding of their clinical conditions. Deep learning has emerged as a successful technique for various natural language understanding tasks preconditioned with large annotated datasets. In this paper, we propose a methodology to create preliminary datasets for clinical paraphrasing, and clinical text simplification to foster training of deep learning-based clinical paraphrase generation and simplification models.

1 Introduction and related work

Paraphrasing (a.k.a. paraphrase generation) is transforming a text that can be a word, phrase, sentence, paragraph, or a document, while retaining the meaning and content. For example, the sentence *'I am very well'* can be paraphrased as *'I am doing great'*. Paraphrasing can lead to a new text which may be simpler, more complex or at the same complexity level as the source text. The task of paraphrasing text to a simpler form is called simplification. In simplification, the output text is a linguistically simplified version of the input text. Paraphrasing and simplification may have numerous applications such as document summarization, text simplification for target audience e.g. children, and question answering [Madnani and Dorr, 2010].

In the clinical context, health care systems and medical knowledge-bases contain large collections of texts that are often not comprehensible to the layman population. For example, clinical texts like radiology reports are used by radiologists to professionally communicate their findings to other physicians [Qenam et al., 2017]. They contain complex medical terminologies that the patients are not familiar with. A recent study reported that allowing patients to access their clinical notes has showed an improvement in their health care process [Kosten et al., 2012]. Realizing the need for increased inclusion of patients in their health care process, large health care systems have allowed for the patients to access their medical records [Delbanco et al., 2015]. However, these medical records contain raw complex clinical text intended for the communication between medical professionals. Paraphrasing or simplification of clinical text will improve the patients' understanding of their health conditions and thereby play an important role in connecting patients and caregivers across the clinical continuum towards better patient outcome.

Traditional clinical paraphrasing and simplification approaches use lexical methods [Kandula et al., 2010; Pivovarov and Elhadad, 2015; Qenam et al., 2017], which are typically focused on identifying complex clinical words, phrases, or sentences and replace them with their alternatives in case of paraphrasing or simpler versions in case of simplification. Lexical methods take advantage of knowledge sources like Unified Medical Language System (UMLS) metathesaurus [Lindberg et al., 1993] which contains grouped words and phrases that describe various medical concepts. Simplification is traditionally performed by mapping UMLS concepts to their alternatives provided in consumer health vocabulary (CHV) [Qenam et al., 2017].

Recently, paraphrase generation was casted as a monolingual machine translation problem resulting in the development of data-driven methods using statistical machine translation (SMT) [Koehn, 2010], and neural machine translation (NMT) principles [Koehn, 2017]. SMT methods [Quirk et al., 2004; Wubben et al., 2010; Zhao et al., 2009] model the conditional distributions of the words and phrases and replace the phrases in the source text with the phrases that maximize the probability of the resulting text. However, syntactic relationships are difficult to model using SMT methods. Monolingual NMT systems use neural network architectures to model complex relationships by automatically learning from large datasets containing source and target text pairs, both belonging to the same language. Current NMT systems for paraphrase generation or simplification [Brad and Rebedea, 2017; Hasan et al., 2016; Prakash et al., 2016] use sequence-to-sequence networks based on encoder-decoder architectures. Unlike traditional methods,

*This work was conducted as part of an internship program at Philips Research.

NMTs do not need semantic or syntactic rules to be explicitly defined. However, they need carefully constructed datasets that contain sufficient information to robustly train the deep neural networks.

Existing clinical paraphrasing and simplification datasets are limited to short phrases. Hasan et al., (2016) trained an attention-based encoder-decoder model [Bahdanau et al., 2015] using a dataset created by merging two word and phrase level datasets: the paraphrase database (PPDB) [Pavlick et al., 2015] and the UMLS metathesaurus. They showed that their model outperformed an upper bound paraphrasing baseline. However, they used a phrasal dataset that does not contain more complex contextual knowledge like a sentential dataset, and the ability of the network to simplify the clinical text was not explored. In contrast to paraphrasing, simplification is a harder problem and may involve addition, deletion or splitting of sentences to suite the target audience. These operations require additional knowledge that a dataset with longer sequences like sentences or paragraphs could provide. Other studies [Brad and Rebedea, 2017; Prakash et al., 2016] have trained encoder-decoder architectures with attention for paraphrasing using general domain sentence level datasets like Microsoft Common Objects in Context (MSCOCO) [Lin et al., 2014], Newsela [Xu et al., 2015] and Wikianswers [Fader et al., 2013]. They demonstrated that neural machine translation models successfully captured the complex semantic relationships from the general domain datasets. However, it is unclear how these networks would perform on complex clinical text.

In this paper, our aim is to pioneer the creation of parallel (with source and target pairs) sentential datasets for clinical paraphrase generation and simplification. Web-based unstructured knowledge sources like `www.mayoclinic.com` contain articles on various medical topics. We obtain articles with matching titles from different web-based knowledge sources and align the sentences using various metrics to create paraphrase and simplification pairs. Additionally, we train NMT models using the prepared clinical datasets and present baseline performance metrics for both clinical paraphrase generation and simplification.

Next section outlines our approach to create clinical paraphrase generation and simplification datasets. First, we discuss our proposed methodology for extracting sentence pairs from web-based clinical knowledge sources. Then we describe various metrics to align the pairs of related sentences for dataset creation. Section 3 discusses the neural network architectures used for establishing baselines. Sections 4 and 5 present the performance evaluation of the models and in section 6 we conclude and discuss the future work.

2 Approach

2.1 Paraphrase pairs from web-based resources

Web-based textual resources contain large collections of articles for various medical topics related to diseases, anatomy, treatment, symptoms etc. These articles are often targeted for general (non-clinician) users and are easier to understand unlike the complex clinical reports written by clinicians. We crawl the articles with same topics from two or more web-based knowledge sources. Each sentence in a topic (i.e. in an article) from one resource is mapped to sentences belonging to the same topic from another resource(s) using a one-to-many scheme to create all possible sentence pair combinations. These sentence pairs essentially contain a large number of unrelated pairs from which meaningful paraphrasing pairs are identified.

Manual identification of the relevant paraphrase pairs is a tedious task as the sentence pair combinations (as discussed above) contain a large number (in millions) of unrelated sentence pairs. Therefore, we use an automated approach to identify the paraphrase pairs from the sentence pair combinations. Our method is similar to the approach by [Zhu et al., 2010]. They use TF-IDF [M. Shieber and Nelken, 2006] metric to align sentences between Wikipedia and Simple-Wikipedia knowledge sources to create sentence pairs for the text simplification task. However, some studies, e.g. Xu et al., 2015, reveal the noisy nature of such datasets, which motivated us to explore various textual similarity/distance metrics instead of relying on one single metric for sentence alignment. Our intuition is that the strengths of a collection of diverse metrics may be useful for better sentence alignment. In addition to various existing metrics, we train a neural paraphrase identification model to estimate a similarity score between two sentences, which is also used as a supplementary sentence alignment metric.

2.2 Sentence alignment

Paraphrase pairs can be identified by computing various sentence similarity/distance metrics between the two sentences in a pair. Various character-level and word-level metrics that we used are described below.

Levenshtein distance

Levenshtein distance [Levenshtein, 1966] is defined as the minimum number of string operations consisting of additions, deletions, and substitutions of symbols that are necessary to transform one string into another. Normalized Levenshtein distance (LDN) is computed by dividing the number of string operations required by the length of the longer string. Character- or word-level LDN is calculated by treating characters or words as symbols respectively:

$$LDN = \frac{N}{max\,(n,\ m)} \qquad (1)$$

where N is the minimum number of string operations to transform a text x to y or vice versa, and n and m are the number of symbols in the texts x and y respectively.

Damerau-Levenshtein distance

Damerau-Levenshtein distance [Damerau, 1964] is similar to LDN and is defined as the minimum number of string operations needed to transform one string into the other. In addition to the string operations in Levenshtein distance, Damerau-Levenshtein distance further includes transposition of two adjacent symbols. Normalized Demerau-Levenshtein distance (DLDN) is calculated by dividing the number of string operations by the number of symbols in the longer string.

Optimal string alignment distance

Optimal string alignment distance [Herranz et al., 2011] is a variant of DLDN but under a restriction that no substring is edited more than once. The normalized form is computed similarly as in DLDN.

Jaro-Winkler distance

Jaro-Winkler distance (JWD) [Winkler, 1990] computes the distance between two strings, where the substitution of two close symbols is considered more important than the substitution of two symbols that are far from each other. The Jaro-Winkler distance JWD is given by:

$$JWD = \begin{cases} d_j, & if\ d_j < 0 \\ d_j + k\ p\ (1 - d_j), & otherwise \end{cases} \qquad (2)$$

where k is the length of the common prefix at the start of the string up to 4 symbols, p is the constant usually set to 0.1 and d_j is the Jaro distance given by:

$$d_j = \begin{cases} 0 & if\ q = 0 \\ \frac{1}{3}\left(\frac{q}{n} + \frac{q}{m} + \frac{q - t}{q}\right), & otherwise \end{cases} \qquad (3)$$

where q is the number of matching words between the two texts x and y with lengths n and m respectively and t is half of the number of transpositions. Jaro-Winkler distance is a normalized quantity ranging from 0 to 1.

Longest common subsequence

Longest common subsequence distance (LCSD) [Bakkelund, 2009] is computed using the following equation:

$$LCSD = 1 - \frac{LCS(x, y)}{max(n, m)} \qquad (4)$$

where LCS (longest common subsequence) is the longest subsequence common to strings x and y with lengths n and m respectively.

N-gram distance

N-gram is a contiguous sequence of n items from a given sample of a text. N-gram distance [Kondrak, 2005] is similar to computing LCS but in this case the symbols are n-grams. We used n = 4 in this paper.

Cosine similarity

Cosine similarity between two strings is computed as the cosine of the angle between the vector representation of two strings (x and y) and is given by the equation:

$$CS = \frac{V_x . V_y}{|V_x| . |V_y|} \qquad (5)$$

Jaccard similarity

Jaccard similarity is calculated as the ratio of the intersection to the union of the items in the two strings.

Sorensen similarity

Sorensen similarity (also called Sorensen-Dice coefficient) [Sørensen, 1948] is similar to Jaccard similarity and it is computed as the ratio of twice the number of common items (intersection) and the sum of number of items in the two strings.

All the above metrics are used in their normalized forms (values between 0 to 1). These metrics calculate the similarity/distance between the sentence pairs using the character- or word-level overlap and the pattern of their occurrences in the sentences. However, these metrics do not consider the presence of concepts (e.g. words or phrases) that are paraphrased using a different vocabulary (e.g. 'glioma' can be paraphrased with its synonym 'brain tumor') and also do not perform well for sentences that differ by a few words resulting in contradicting sentences. Therefore, we need a similarity metric that can consider complex semantic relationships between the concepts represented in the sentences. Deep neural network architectures with recurrent neural networks (RNNs), and convolution neural networks (CNNs) have so far demonstrated state-of-the-art performance [Conneau et al., 2017] in learning semantic associations between the sentences. Therefore, deep-learning based systems are increasingly being used for advanced natural language inferencing tasks like paraphrase identification, and textual entailment [Ghaeini et al., 2018], which motivated us to create a neural paraphrase identification model for the purpose of supplementing our sentence similarity measures for better sentence pair alignment.

2.3 Paraphrase identification metric

Neural paraphrase identification can be stated as a binary classification task in which a neural network model estimates the probability that the two sentences are paraphrases. This estimated probability can be used as a similarity metric to align the sentence pairs.

Neural paraphrase identification

The network consists of stacked bidirectional long short-term memory (BiLSTM) layers in a Siamese architecture [Dadashov et al., 2017] (Figure 1). Each arm of the Siamese network consists of three stacked BiLSTM layers. The outputs of the final BiLSTM layers of both the arms are concatenated and fed into the dense layer with ReLU activation followed by a second dense layer with a sigmoid activation function. We use a depth of 300 for all the BiLSTM layers and the dense layers. The maximum sequence length of the BiLSTM layers is set to 30. The words in the input sentences are embedded using Word2Vec embeddings pre-trained on the Google news corpus.

Hybrid dataset for paraphrase identification

Our paraphrase identification model is trained using a hybrid corpus created by merging two paraphrase corpora: Quora question pairs, and Paralex question pairs. The Quora question pair corpus [Iyer et al., 2017] consists of 404289 question pairs with 149263 paraphrase pairs and 255027 non-paraphrase pairs. The Paralex dataset [Fader et al., 2013] consists of 35,692,309 question pairs, where all the question pairs are paraphrases of each other. The Paralex dataset is unbalanced as it does not contain any non-paraphrase pairs. After merging the sentence pairs from both the corpora, we have 35692309 sentence pairs with 35437283 paraphrase pairs and only 255027 non-paraphrase pairs. To balance the dataset, we identify the list of unique questions and then randomly select two questions from the unique questions list and add the pair to the merged corpus as a non-paraphrase pair if the pair does not already exist. Non-paraphrase pairs are created until the non-paraphrase, and paraphrase pairs are equal in number, resulting in a balanced dataset of 70 million pairs.

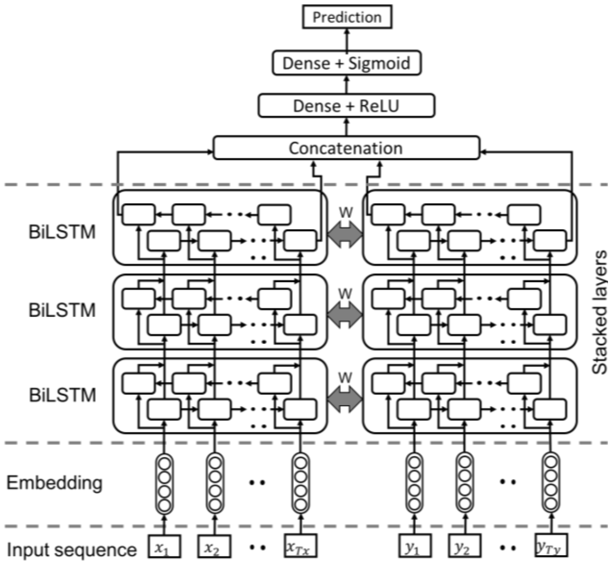

Figure 1. Paraphrase identification architecture. Gray arrows represent weight-sharing between the left and right BiLSTM.

Training

The dataset is preprocessed by removing punctuations, normalization with respect to case, and standard tokenization. The tokens are embedded using Word2Vec embeddings pre-trained on Google news corpus [Mikolov et al., 2013]. Words that are not found in the pre-trained vocabulary are embedded with a zero-vector representing an UNK token. Longer sentences (> 30 words) are truncated, and smaller sentences (< 30 words) are padded with UNK tokens. As the sentences are in a bidirectional relationship to each other, the training pairs are swapped to increase the dataset size. The dataset is split into 80%, 10% and 10% for training, validation and testing respectively.

The paraphrase identification model is trained using Adam optimizer [Kingma and Ba, 2014] with Nesterov momentum [Nesterov, 1983] to optimize a binary cross entropy loss. The update direction is calculated using a batch size of 512. We utilize early stopping using validation error with patience of 3 epochs to prevent overfitting.

The network is trained for 18 epochs before early stopping at 22 minutes per epoch. The validation accuracy of our model is 95% and test accuracy is 97%.

Paraphrase identification model for sentence alignment

The probability score from our paraphrase identification model for the predicted class is used along with the word- and character-level similarity/distance metrics to calculate a mean similarity score. Note that, all normalized distance metrics are converted into similarity metrics by subtracting the corresponding score from 1, thereby obtaining 12 different similarity metrics. The mean similarity score is computed using the formula given below:

$$mean\ similarity\ score = mean\ (all\ similarity\ scores) \tag{6}$$

Minimum and maximum thresholds of 0.5 and 0.8 are empirically selected by observing the sentence pairs. Sentence pairs with a similarity score within these thresholds are considered paraphrase pairs. We use a maximum threshold of 0.8 to avoid selection of identical sentences.

2.4 Paraphrase generation dataset

The paraphrase sentence pairs are obtained from three web-based unstructured knowledge sources: Wikipedia, SimpleWikipedia, and MayoClinic. These sentence pairs form our clinical paraphrase generation dataset, which is later used to train a baseline neural paraphrase generation model.

Wikipedia and SimpleWikipedia

SimpleWikipedia contains simplified versions of pages from the original Wikipedia. However, the text in the corresponding documents is unaligned (no sentence-to-sentence matching). Pairing sentences from Wikipedia to those in SimpleWikipedia leads to a parallel corpus for the simplification dataset as the later mostly contains simplified versions of the former. However, in the case of paraphrase generation, the resultant pairs can be swapped as paraphrasing applies in both directions. The swapping also helps augmenting the dataset. We create a parallel corpus using 164 matched titles from Wikipedia from clinically relevant categories such as anatomy, brain, disease, medical condition etc. Sentences from each of 164 Wikipedia documents are paired with all the sentences from the documents with identical titles from SimpleWiki. Thus, we obtain 818520 sentence pairs for which we compute similarity scores as discussed in the previous subsections. We finally obtain 1491 related sentence pairs after thresholding the mean similarity score and we name this parallel corpus as WikiSWiki.

Mayoclinic

Mayoclinic contains pages for 48 identically matched titles from the 164 titles identified from Wikipedia. Unique sentences from WikiSWiki were paired with the sentences obtained from the pages with matched titles from Mayoclinic and similarity scores are computed. Using the same thresholds as above, 3203 sentence pairs are selected. These pairs are added to the WikiSWiki corpus to form a corpus containing 4694 sentence pairs; we name it WikiSwikiMayo.

2.5 Simplification dataset

The WikiSWiki is a simplification corpus as it mostly contains sentences mapped to their simpler forms. However, the small number of sentence pairs may be insufficient for training the network to learn complex relationships required for clinical text simplification. Therefore, we use additional web-based knowledge sources to increase the dataset size. Web-based knowledge sources: `www.webmd.com` (webmd) and `www.medicinenet.com` (medicinenet), are other clinical knowledge sources that are similar to MayoClinic. Through manual inspection, we found that webmd contains simpler sentences than medicinenet in many topics that we have examined, which is reasonable as medicinenet content is curated by clinicians. Therefore, we use them as additional knowledge sources to create our simplification dataset. For 164 topics from the WikiSWiki dataset we perform a google search with 'webmd' and 'medicinenet' as additional search terms. The search returns 61314 sentences from webmd and medicinenet for all 164 topics. Sentences from medicinenet are paired to the sentences from SimpleWiki and webmd from the articles with matched titles. Sentences from Wikipedia articles are paired with sentences from webmd separately as they are already paired with SimpleWikipedia. We obtain 714608 new sentence pairs resulting in 1002 final pairs after computing similarity scores and thresholding. These sentence pairs are merged with WikiSWiki dataset to create the monolingual clinical simplification dataset containing 2493 sentence pairs. Although our final corpus contains a small number of sentence pairs, our main contribution in this paper is to introduce an automated method to create sentence pairs from web-based knowledge sources, towards creating a large clinical simplification corpus in the future.

3 Paraphrase generation and simplification

3.1 Model

Sequence-to-sequence models using encoder-decoder architecture with attention [Vinyals et al., 2015] (Figure 2) are trained for both paraphrase generation and simplification tasks. The encoder and decoder are made of three stacked RNN layers using BiLSTM cells and LSTM cells respectively. We use a cell depth of 1024 for all the layers in the encoder and the decoder. The maximum sequence length is set to 50. The sentences are preprocessed, and the words are encoded using one-hot vector encoding. The outputs of the decoder are projected onto the output vocabulary space using a dense layer with a softmax activation function.

3.2 Training

The network parameters are optimized by minimizing a sampled softmax loss function. The gradients are truncated by limiting the global norm to 1. The network is trained using mini-batch gradient descent algorithm with batch size of 128. An initial learning rate of 0.5 is used with a decay of 0.99 for every step. The training set is shuffled for every epoch. The networks are trained using 80% of the sentence pairs and validated on 10% and tested on 10%. Both models are developed using Tensorflow, version 1.2, and two Tesla K20 GPUs.

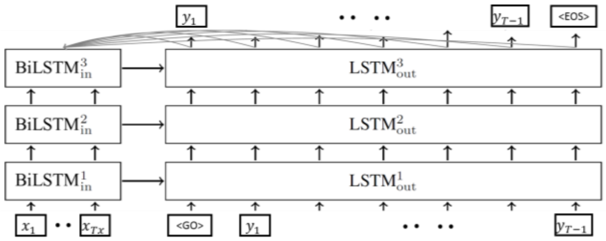
y_1
• • • •
y_{T-1}
<EOS>
$\mathrm{BiLSTM}^3_{\mathrm{in}}$
$\mathrm{LSTM}^3_{\mathrm{out}}$
$\mathrm{BiLSTM}^2_{\mathrm{in}}$
$\mathrm{LSTM}^2_{\mathrm{out}}$
$\mathrm{BiLSTM}^1_{\mathrm{in}}$
$\mathrm{LSTM}^1_{\mathrm{out}}$
x_1
• •
x_{Tx}
<GO>
y_1
• • • •
y_{T-1}

Figure 2. Encoder-decoder architecture. x and y are the source and target sequences respectively.

For paraphrase generation, the network is trained using the WikiSwikiMayo corpus containing 4694 sentence pairs. The source and target sentences are swapped as paraphrasing is bidirectional thereby, doubling the number of sentence pairs to 9388. The dataset is divided into training, validation and test sets. The training sentence pairs that contain sentences from the source side of the test are removed to prevent data leak issues. Same is repeated for the validation set. Using this we make sure that any sentence occurs as a source sentence in exactly one of the sets (training, test or validation). The number of sentence pairs in training, test and validation datasets are 6095, 611 and 611 respectively. The paraphrase generation network is trained for 10000 steps with a batch size of 128 samples per step.

The simplification corpus containing 2493 sentence pairs is used to train the simplification network. Vocabularies for source and target are created separately in case of simplification. The source and target vocabularies are different in case of text simplification. As simplification is a unidirectional task, we do not use data swapping. We prevent data leak issues using the same procedure as paraphrase generation while splitting the data. The training, test and validation sets contain 1918, 187 and 187 sentence pairs respectively. The simplification network is trained for 3500 steps.

4 Evaluation metrics

BLEU [Papineni et al., 2002], METEOR [Banerjee and Lavie, 2005] and translation error rate (TER) [Snover et al., 2006] are used to evaluate our models. These metrics are shown to correlate with human judgements for evaluating paraphrase generation models [Wubben et al., 2010]. BLEU looks for exact string matching using n-gram overlaps to evaluate the similarity between two sentences. METEOR uses WordNet to obtain synonymously related words to evaluate sentence similarity. Higher BLEU and METEOR scores indicate higher similarity. TER score measures the number of edits necessary to transform the source sentence to the target. Lower TER score indicates higher similarity.

5 Results and discussion

5.1 Sentence alignment

Table 1 presents a few examples of the aligned sentence pairs for both clinical paraphrase generation and simplification.

Clinical Paraphrase Generation	Mean Sim. Score
Example 1: Good **S1:** No drug is currently approved for the treatment of smallpox. **S2:** No cure or treatment for smallpox exists	0.52
Example 2: Acceptable **S1:** Worldwide, breast cancer is the most common invasive cancer in women. **S2:** After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States	0.62
Example 3: Bad **S1:** Gallbladder cancer is a rare type of cancer which forms in the gallbladder. **S2:** At this stage, gallbladder cancer is confined to the inner layers of the gallbladder	0.53
Clinical Text Simplification	
Example 1: Good **S1:** In Western cultures, ingestion of or exposure to peanuts, wheat, nuts, certain types of seafood like shellfish, milk, and eggs are the most prevalent causes. **S2:** In the Western world, the most common causes are eating or touching peanuts, wheat, tree nuts, shellfish, milk, and eggs.	0.54
Example 2: Acceptable **S1:** Together the bones in the body form the skeleton. **S2:** The bones are the framework of the body.	0.54
Example 3: Bad **S1:** There are two major types of diabetes, called type 1 and type 2 **S2:** There are other kinds of diabetes, like diabetes insipidus.	0.54

Table 1. Examples of aligned sentence pairs. Good represents that accepted sentences are paraphrases. Bad represents that accepted sentences are not paraphrases.

In Table 1, for both paraphrase generation and text simplification tasks, though the similarity score between the sentence pairs is similar across all the examples there is a large variability in the classification of the sentence pair. This means there is an overlap between the distributions of the mean similarity score of the paraphrase pairs and the non-paraphrase pairs. Therefore, the selection of minimum threshold less than 0.5 introduces more non-paraphrase pairs into the dataset and by selecting the threshold more than 0.5 we lose a large number of pairs that are paraphrases. One desirable approach is to train a linear regression or any multi-variate machine learning model to classify the paraphrase pairs using all the computed similarity metrics. However, training such machine learning systems requires ground-truth data and therefore is outside the scope of this paper.

Our paraphrase identification system uses a vocabulary from the Google News corpus dataset. The words that are not present in this vocabulary are assigned the UNK token. Therefore, the neural paraphrase identification network is not sensitive when two semantically similar sentences refer to different objects. However, this problem is minimized in our case as we pair the sentences from the pages belonging to the same topic. Furthermore, using other similarity metrics that are based on word matching helps in overcoming this problem in cases where the paraphrase identification metric is insensitive. We examined that this holds true in majority of the pairs by visual inspection of the selected sentence pairs, for both the datasets.

5.2 Paraphrase generation and simplification

Average quality scores on the test sets for the clinical paraphrase generation and the text simplification models are presented in Table 2. These scores serve as baselines for clinical paraphrase generation and text simplification for the datasets that we have created. The quality metrics are lower for clinical text simplification than the paraphrase generation. This is expected as in the case of paraphrase generation many of the words from the source sentence can be retained in the paraphrased sentence whereas simplification involves complex transformations which results in different words in the resulting sentence and hence the quality scores are low. Further human evaluations are required to better rate the performance of the simplification model.

Task	BLEU	METEOR	TER
Clinical Paraphrase Generation	9.4±0.5	15.1±0.3	108.7±1.5
Clinical Text Simplification	9.9±1.6	10.6±0.8	97.7±2.9

Table 2. Average scores computed over test sentence pairs.

Few example outputs of the clinical paraphrase generation and simplification system are presented in Table 3. The examples show that both paraphrase generation and simplification models retained the knowledge of the overall topic in the generated sentences. Example 2 in both models shows that, though the topic of the generated sentence matches with the source, the sentence is not a paraphrase or the simplification respectively, as the context in the resultant sentence is different from that of the source. This may be because of the failure in the alignment of the sentences while creating the datasets. This shows that the paraphrase identification model and the metrics were not fully sufficient to pair the sentences accurately. In particular, the paraphrase identification model trained on general domain question pairs may not generalize well to identify paraphrase pairs in case of clinical texts. The solution may be using transfer learning and training the paraphrase identification network on a subset of human rated clinical paraphrases.

Clinical Paraphrase Generation		Example 1	Example 2
	Source	dengue fever pronounced den gay is an infectious disease caused by the dengue virus	Lung cancer often spreads (metastasizes) to other parts of the body, such as the brain and the bones

Target	dengue fever is caused by any of the four dengue viruses spread by mosquitoes that thrive in and near human lodgings	Primary lung cancers themselves most commonly metastasize to the brain, bones, liver and adrenal glands
Generated	Dengue fever is a mosquito borne tropical disease caused by the dengue virus	Lung cancer staging is an assessment of the degree of spread of the cancer from its original source
Clinical Text Simplification		
Source	Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced	Ventricular tachycardia can be classified based on its morphology
Target	Diabetes is the condition that results from lack of insulin in a person blood or when their body has a problem using the insulin it produces insulin resistance	Ventricular tachycardia can be treated in a few different ways
Generated	Diabetes can occur when the pancreas produces very little to no insulin or when the body does not respond appropriately to insulin	Ventricular tachycardia can be caused by many different things

Table 3. Example outputs from clinical paraphrase generation and simplification models.

Our datasets consist of a small number of sentence pairs (few thousands) and may not be sufficient for the neural network models to learn complex clinical concepts. Furthermore, we use only 164 medical topics from Wikipedia for this work. Improving the efficiency of paraphrase identification and inclusion of more knowledge sources and topics will create larger and better training datasets. Many sentences that are paired contain text related to additional information that the other sentence does not contain. For example:

Source: “It isn’t clear why some people get asthma and others don’t, but it’s probably due to a combination of environmental and genetic factors”.

Target: “Asthma is thought to be caused by a combination of genetic and environmental factors”.

The removal of the additional text in the first part of the source sentence will improve the training of the neural network as it can focus more on the important text. The unwanted text in this example can be easily removed as it is clearly separated from the rest of the sentence. However, many sentences that contain unwanted text are not easily separable. Moreover, manual removal of unwanted text from thousands of sentences (if not millions) is not practical. Automated methods are needed to remove unwanted text during sentence alignment, which would help to create cleaner datasets.

Previous research has found that existing simplification datasets created using Wikipedia-like knowledge sources are noisy [Xu et al., 2015] as these knowledge sources are not created with a specific objective. However, task specific datasets for clinical paraphrase generation and simplification do not exist as of writing this paper. Therefore, we approached the creation of such datasets for clinical paraphrase generation and simplification using web-based knowledge sources. We hope that this serves as a starting point towards developing automated approaches for creating task specific datasets using unstructured knowledge sources.

6 Conclusion and future work

This paper presents a preliminary work on automated methodology to create clinical paraphrase generation and simplification datasets. We use web-based knowledge sources and automatically align sentence pairs from matching topics to create the datasets. Additionally, these datasets are used to train sequence-to-sequence models leveraging an encoder-decoder architecture with attention for paraphrase generation and simplification. Further research to improve string similarity metrics is required to accurately identify similar sentence pairs to create cleaner datasets. In future, we will include more knowledge sources and topics to create larger datasets and use automated methods to remove unrelated or unwanted text in the paired sentences.

References

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., Bengio, Y.,. Neural Machine Translation By Jointly Learning To Align and Translate, in: ICLR. pp. 1–15, 2015.

[Bakkelund, 2009] Bakkelund, D.,. An LCS-based string metric. University of Oslo, Oslo, Norway, 2009.

[Banerjee and Lavie, 2005] Banerjee, S., Lavie, A.,. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: ACL. pp. 65–72, 2005.

[Brad and Rebedea, 2017] Brad, F., Rebedea, T.,. Neural Paraphrase Generation using Transfer Learning, in: INLG. pp. 257–261, 2017.

[Conneau et al., 2017] Conneau, A., Kiela, D., Schwenk, H.,. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, in: CoRR. 2017.

[Dadashov et al., 2017] Dadashov, E., Sakshuwong, S., Yu, K.,. Quora Question Duplication 1–9, 2017.

[Damerau, 1964] Damerau, F.J.,. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* 7, 171–176, 1964.

[Delbanco et al., 2015] Delbanco, T., Walker, J., Darer, J.D., Elmore, J.G., Feldman, H.J.,. Open Notes: Doctors and Patients Signing On. *Ann. Intern. Med.* 153, 121–126, 2015.

[Fader et al., 2013] Fader, A., Zettlemoyer, L., Etzioni, O.,.

Paraphrase-Driven Learning for Open Question Answering, in: ACL. pp. 1608–1618, 2013.

[Ghaeini et al., 2018] Ghaeini, R., Hasan, S.A., Datla, V. et al. DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference, in: NAACL HTL, 2018.

[Hasan et al., 2016] Hasan, S.A., Liu, B., Liu, J. et al. Neural Clinical Paraphrase Generation with Attention, in: CNLP Workshop. pp. 42–53, 2016.

[Herranz et al., 2011] Herranz, J., Nin, J., Sole, M.,. Optimal Symbol Alignment Distance: A New Distance for Sequences of Symbols. *IEEE Trans. Knowl. Data Eng.* 23, 1541–1554, 2011.

[Iyer et al., 2017] Iyer, S., Dandekar, N., Csernai, K.,. Quora question pair dataset [WWW Document], 2017.

[Kandula et al., 2010] Kandula, S., Curtis, D., Zeng-Treitler, Q.,. A semantic and syntactic text simplification tool for health content., in: AMIA. pp. 366–70, 2010.

[Kingma and Ba, 2014] Kingma, D.P., Ba, J.,. Adam: A Method for Stochastic Optimization, in: ICLR. pp. 1–15, 2014.

[Koehn, 2017] Koehn, P.,. Neural Machine Translation. *CoRR*, 2017.

[Koehn, 2010] Koehn, P.,. Statistical Machine Translation, 1st ed. Cambridge University Press, NY, USA, 2010.

[Kondrak, 2005] Kondrak, G.,. N-gram similarity and distance. *SPIR* 115–126, 2005.

[Kosten et al., 2012] Kosten, T.R., Domingo, C.B., Shorter, D., Orson, F. et al. Inviting Patients to Read Their Doctors' Notes: A Quasi- experimental Study and a Look Ahead. *Ann. Intern. Med.* 157, 461–470, 2012.

[Levenshtein, 1966] Levenshtein, V.,. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Sov. Phys. Dokl.* 10, 707–710, 1966.

[Lin et al., 2014] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.,. Microsoft COCO: Common objects in context, in: ECCV. pp. 740–755, 2014.

[Lindberg et al., 1993] Lindberg, D.A., Humphreys, B.L., McCray, A.T.,. The Unified Medical Language System. *Methods Inf. Med.* 32, 281–291, 1993.

[M. Shieber and Nelken, 2006] M. Shieber, S., Nelken, R.,. Towards robust context-sensitive sentence alignment for monolingual corpora, 2006.

[Madnani and Dorr, 2010] Madnani, N., Dorr, B.J.,. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *Comput. Linguist.* 36, 341–387, 2010.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., Dean, J.,. Distributed Representations of Words and Phrases and their Compositionality. *NIPS* 1–9, 2013.

[Nesterov, 1983] Nesterov, Y.,. A method for unconstrained convex minimization problem with the rate of convergence o(1/k^2). *Dokl. AN USSR* 269, 543–547, 1983.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., Zhu, W.,. BLEU: a method for automatic evaluation of machine translation, in: ACL. pp. 311–318, 2002.

[Pavlick et al., 2015] Pavlick, E., Rastogi, P., Ganitkevitch, J., Durme, B. Van, Callison-Burch, C.,. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. *ACL* 425–430, 2015.

[Pivovarov and Elhadad, 2015] Pivovarov, R., Elhadad, N.,. Automated methods for the summarization of electronic health records. *J. Am. Med. Informatics Assoc.* 22, 938–947, 2015.

[Prakash et al., 2016] Prakash, A., Hasan, S.A., Lee, K., Datla, V., Qadir, A., Liu, J., Farri, O.,. Neural Paraphrase Generation with Stacked Residual LSTM Networks, in: COLING. pp. 2923–2934, 2016.

[Qenam et al., 2017] Qenam, B., Kim, T.Y., Carroll, M.J., Hogarth, M.,. Text Simplification Using Consumer Health Vocabulary to Generate Patient-Centered Radiology Reporting: Translation and Evaluation. *J. Med. Internet Res.* 19, e417, 2017.

[Quirk et al., 2004] Quirk, C., Brockett, C., Dolan, B.,. Monolingual Machine Translation for Paraphrase Generation, in: ACL, 2004.

[Snover et al., 2006] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.,. A Study of Translation Edit Rate with Targeted Human Annotation, in: AMTA. pp. 223–231, 2006.

[Sørensen, 1948] Sørensen, T.,. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.* 5, 1–34, 1948.

[Vinyals et al., 2015] Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I., Hinton, G.,. Grammar as a Foreign Language, in: NIPS, 2015.

[Winkler, 1990] Winkler, W.E.,. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, in: ASA. pp. 354–359, 1990.

[Wubben et al., 2010] Wubben, S., van den Bosch, A., Krahmer, E.,. Paraphrase Generation As Monolingual Translation: Data and Evaluation, in: INLG, INLG '10. pp. 203–207, 2010.

[Xu et al., 2015] Xu, W., Callison-Burch, C., Napoles, C.,. Problems in Current Text Simplification Research: New Data Can Help, in: ACL. pp. 283–297, 2015.

[Zhao et al., 2009] Zhao, S., Lan, X., Liu, T., Li, S.,. Application-driven statistical paraphrase generation, in: ACL. pp. 834–842, 2009.

[Zhu et al., 2010] Zhu, Z., Bernhard, D., Gurevych, I.,. A Monolingual Tree-based Translation Model for Sentence Simplification, in: COLING. pp. 1353–1361, 2010.