# Personalized Text Normalization of Electronic Health Records

ANONYMOUS AUTHOR(S)*

Text normalization standardizes the text by removing irrelevant information and replacing non-standard words with standard ones. Text normalization improves the efficiency of downstream applications on text in natural language processing tasks. The existing works on electronic health records (EHRs) clinical notes perform named entity recognition, acronym expansion but did not attempt the personalized text normalization. In this report we attempted to implement personalized text normalization of clinical notes using bert2bert encoder-decoder network. We pretrained the model on MIMIC-III clinical research notes and tested on external test set, n2c2 nlp research data. We have prepared the annotations (labels) for both training and validation datasests. So far, results have not shown promising results. Further experimental studies are in progress and will be reported soon. Implementation code can be found at: https://github.com/SanjeevaRDodlapati/Clinical-Text-Norm-BERT.

CCS Concepts: • **Personalized text normalization**; • **encoder-decoder**; • **Clinical-Text-Norm-BERT**;

Additional Key Words and Phrases: Electronic Health Records, neural networks, Text normalization

**ACM Reference Format:**

Anonymous Author(s). 2022. Personalized Text Normalization of Electronic Health Records. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, April 30–May 6, 2022, New Orleans, LA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

Text normalization is the process of standardizing the text by replacing non-standard words with standard words and removing irrelevant components. This process reduces randomness in the information and improves the efficiency of downstream applications such as information retrieval, text to speech and machine translation.

Earlier approaches for text normalization often relied on hand-crafted algorithms tailored to a specific context, while more recent approaches have focused on supervised machine learning[6]. Substitution list methods are the simplest form of normalization. They simply look up each historical variant in a precompiled list that maps it to its intended normalization. Rule-based approaches try to find mapping based on predefined rules hidden in the text. Distance-based methods have also been used to compare variants to entries in a full-form lexicon. Distance scores are also used in unsupervised methods for clustering historical variants of the same modern form. Statistical models take a probabilistic approach to normalization tasks to optimize the probability of replacement words. Recently, neural network architectures have been used for a variety of NLP tasks including text normalization.

Electronic health records (EHRs) data is crucial in developing personalized medicine. However, EHR data comes with many challenges. They contain both structured data like ICD-10 codes and unstructured free text data. Standardization of EHR data is needed before it can be exploited to develop targeted healthcare solutions.



Manuscript submitted to ACM

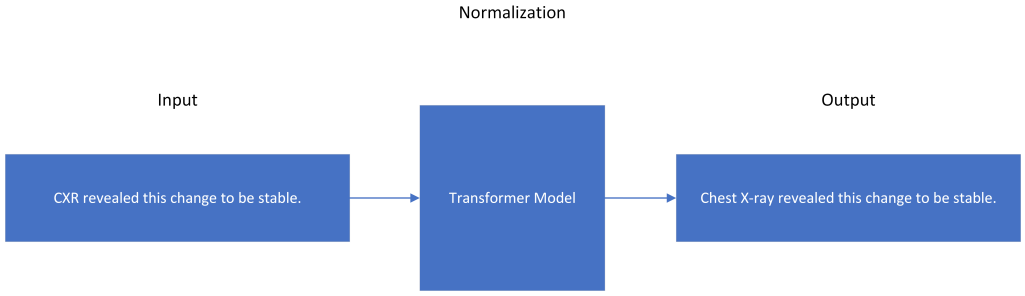


Fig. 1. Overview of the proposed method for personalized text normalization of electronic health records

There have been many efforts to develop natural language processing (NLP) methods to normalize text. Yolchuyeva et.al [40] used convolutional neural networks for text normalization. Arnaud et.al. [3] used pretrained transformer model to learn contextual embeddings from free-text triage notes recorded at the hospital emergency department.

However, personalized text normalization is under explored. In this project, we are proposing to develop personalized text normalization of electronic health records (EHR) by using NLP machine learning methods. Specifically, we will be using a transformer-based network pre-trained on EHR dataset (BEHRT)[18]. We use its feature extractor to extract important features from the EHR dataset and change the model to help normalize clinical records. We have collected the patient EHRs from n2c2 NLP Research Data Sets of Harvard medical school.

Our main contribution in this work is building a deep learning model for text normalization of clinical patients.

## 2 RELATED WORK

### 2.1 Text Normalization Techniques

Text normalization has a long history in speech technology, dating back to the earliest work on full TTS synthesis[2]. [33] provided a unifying model for most text normalization problems in terms of weighted finite-state transducers (WFSTs). The first work to treat the problem of text normalization as essentially a language modeling problem was[35]. More recent machine learning work specifically addressed to TTS text normalization include [28, 34, 36]. In the last few years there has been a lot of work that focuses on social media [4, 5, 7, 11, 16, 17, 19–21, 27, 38, 39], with some recent neural work in this space[8, 23]. Such work tends to focus on different problems from those of TTS: on the one hand one, in social media one often has to deal with odd spellings of words such as "cu 18r", "cooooooooooooooolllll", or "dat suxx", which are less of an issue in most applications of TTS; on the other, expansion of digit sequences into words is critical for TTS text normalization, but of no interest to the normalization of social media texts.

The standard approach of text normalization commonly used by industries involves complex hand-written grammars to verbalize input tokens, such as Google's Kestrel TTS text normalization system[9]. The system works by classifying nonstandard tokens into their respective semiotic classes (classification grammars) and taking their context to appropriately verbalize the token (verbalization grammars). The classification and verbalization grammars are then compiled into weighted finite-state transducers (WFSTs) which will be used to pass non-standard text into.

Within the last few years, deep learning has taken over the speech and language technology field. One recent work[37] decided to tackle the text normalization task as a supervised sequence to sequence deep learning task: given a large corpus of written text aligned to its normalized spoken form, train a recurrent neural network (RNN) to learn

the correct normalization function. Only recently, transformer models have been topping the charts in various tasks ranging over different fields, such as machine translation in natural language processing. Zhang et al. [41] suggested that we can treat the text normalization problem as a machine translation task, where the source language is raw text and the target language is normalized text, in the same language as the source language. They classified the errors into two types, one that preserves the meaning of the text, but contains the wrong form of a word, the other being an error that changes the meaning of the text and conveys an entirely different sense than what is intended. Using this classification, they built a neural network to normalize texts.

### 2.2 Personalized Text Normalization

There have been several methods for text normalization, however, personalized text normalization has not been widely explored. Aw et al, [4] proposed a probabilistic method called AsiaSpic to support the use of user-defined short-forms in a multilingual chat system by exploiting a personalized dictionary for each user to support user-defined short-forms. AsiaSpik is a web-based multi-lingual instant messaging system chat in one language readable in other language by other users. Experiments were conducted using 134 chat messages sent by high school students.

### 2.3 Characteristics of EHRs

Electronic Health Records (EHRs) consist many heterogeneous data[29] elements including: a) Demographic characteristics of patients such as age, gender, ethnicity and socioeconomic status etc; b) Vital signs such as body temperature, pulse rate, respiration rate, and blood pressure indicating the status of the body's vital functions; c) Medications information in the form of narratives or codes (RxNorm), for example, Tylanol 500 mg oral tablet; d) Diagnostic codes (ICDs) that represent disease and related health problems, for example 'acute respiratory failure = J96.00'; e) Procedures of medical, surgical and diagnostic processes such as eyelid skin biopsy, partial mastectomy, MRI, thoracic spine; f) Clinical notes (free-text written by clinical professionals) such as consultation notes, discharge summaries, procedure notes, progress notes and medical notes; g) Laboratory data from medical examination results either as narratives or codes (LOINC) to indicate, for example, red/white blood cell count, hemoglobin, glucose etc.; h) Hospitalization information including admission, length of stay, transfer record, discharge disposition, and observations.

As described above, EHRs contain abundant information that can be utilized to improve patient healthcare and develop precision medicine. However, handling of EHR data is very challenging because EHR data is collected from multiple sources and many healthcare professionals and different departments/hospitals, and it often lacks uniformity in following protocols and medication standards[14]. The medical codes can vary between differnt organizations and countries[31]. EHR data is complicated by multiple biases. Data inaccuracy arise due to data entry mistakes, inaccurate measurements, inconsistent units and protocols. The challenges are further compounded as EHRs consist of semi-structured lab reports, which contain measurement tables and text descriptions, and clinical notes that consist of unstructured free-text format. Figure 2 shows a sample EHR of Veterans Information Systems and Technology Architecture (VISTA) that most widely used EHR in the United States, which contains both structured and unstructured text[32]. Standardization of EHRs is required to exploit abundant information to develop various healthcare applications.

### 2.4 Normalization of EHRs

Normalization of EHRs involves, apart from standardizing natural language errors, identifying the medical concepts or codes within a document and map to natural language to improve interoperability. Medical concept normalization (MCN) is the task of assigning canonical identifiers to concept mentions, in order to unify different ways of referring to

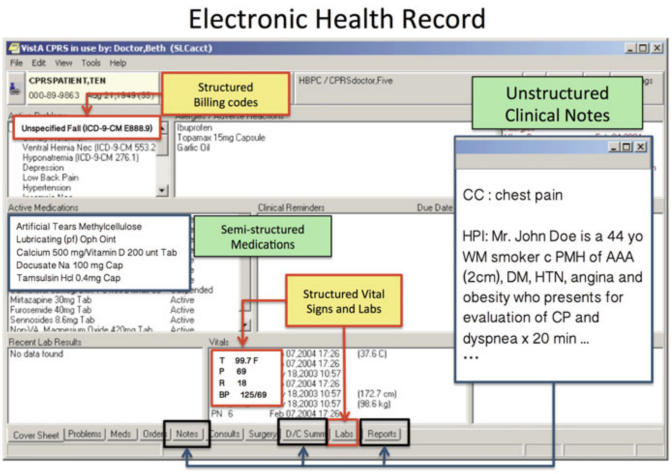


Fig. 2. Sample electronic health record representation

the same concept[24]. There have been many efforts to normalize EHR data. Medical concept normalization benchmark datasets such as ShARe corpus[10] and a dataset by Luo et al[22], have been developed. EHR data normalization pipeline, cTAKES implements a clinical elements model to create structural and semantic mappings[25]. Fast Healthcare Interoperability Resources (FHIR) was developed by HL7 (Health Level 7 - international standard). FHIR has been rapidly adopted by many institutions to bridge the gap in representing unstructured information in clinical narratives[12]. SHARPn consortium is one of the earliest clinical data normalization of syntactic and semantic data[26]. SHARPn pipeline adopts Mirth Connect, an open-source healthcare integration engine and Apache Unstructured Information Management Architecture (UIMA) as a software platform. DeepPhe[30] project utilized NLP method to develop EHR phenotype data normalization pipeline that adehere to FHIR specification. Hong et al[12], developed NLP2FHIR pipeline for FHIR-based EHR normalization using multiple extended datasets, resulted in 30 mapping rules, 62 normalization rules, and 11 NLP-specific FHIR extensions. Zhao et at[42], utilized deep leaning natural language processing model clinical data extraction and normalization of EHRs in Bulgarian languages. Dual embedding for encoding English and Bulgarian languages and subsequently aligned so that both were in same vector space.

## 3 DATA

### 3.1 Data Acquisition:

1. MIMIC-III data: MIMIC-III data has been downloaded from physionet.org after going through training for "CITI Data for Specimens Only Research" to get access. This dataset contains data from over forty thousand patients who stayed in critical care units. The database includes demographics information, measurements of vital signs, laboratory test results, procedures, medications, caregiver notes, imaging reports, and discharge notes[15]. We use this data to pretrain the ClnicalBERT model. We split the data into 80% of training data and rest of 20% as validation set.

2. n2c2(track-2) NLP Research Data: n2c2 data is downloaded from Harvard medical school after obtaining the data access licence. The data contains unstructured notes from research patient data registry. This data also, like MIMIC-III

data, contains clinical notes including medication and discharge notes from multiple visits of thousands of patient. This data is used as a test set to evaluate the ClinicalBERT encoder-decoder model.

### 3.2 Data Preprocess:

MIMIC dataset contains different categories of notes such as notes by nurses, notes about lab test results, scanning reports, and discharge notes etc. Combined different category of notes there are approximately 2.5 million notes in the dataset. The details can be seen figure 3. highest number of notes are for nursing/other category with 800k notes followed by radiology around 500k notes. All the other categories contain approximately 200k or less than that of notes. Due to constraints of resources and time, we have preprocessed discharge summary notes(as explained below) and prepared label sentences in which medical terminology abbreviations are expanded.

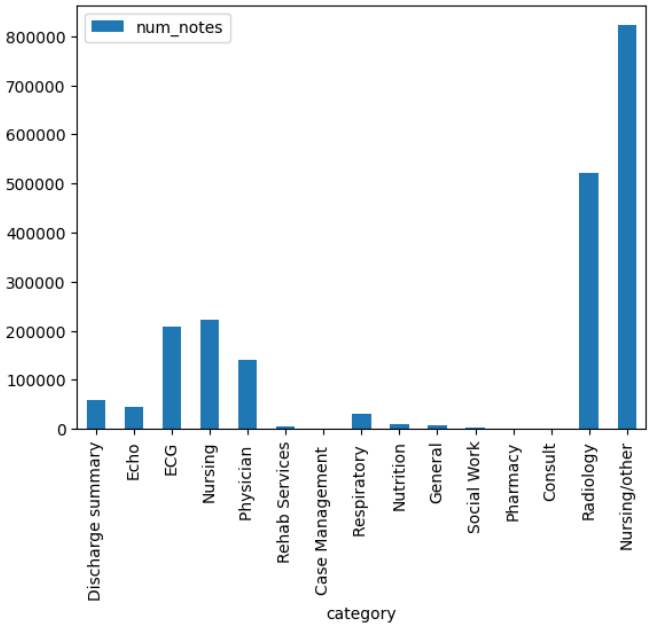


Fig. 3. Nu‘ qqqqqmber of clinical notes for each category in MIMIC-III EHR notes

The data preprocessing is performed on the texts extracted from MIMIC-III dataset. First, we removed unnecessary information inside the texts including dates, repetitive titles, hospital ids, new lines, numbering, uninformative symbols and punctuation. This along with replacing meaningful words make the text easier and more readable for the model to not get deviated from capturing targeted insight from input text. The text is collection of many notes assigned to patients so we extracted sentences from them.

We combined two dictionaries of medical terminology abbreviations, collected from madisonmemorial.org and nhs.uk, with their meaning to prepare labeled input data by replacing abbreviations in the sentences. Whole dictionary is used to replace abbreviations for non-medical users, while 60% and 36% of it used for nurses and and doctors, respectively which the latter is a subset of the former. Each row of prepared input data consists of around 500 words long sequence of sentences in two columns, without and with replacement of abbreviations. Each of these records have three replicates to show replacements for the three personalized categories by including the name of category at the begining of sequence in double quotation marks (e.g. "doctor"). So the final preprocessed input file is consist of 59520 records.

## 4 ENCODER-DECODER ARCHITECTURE

We have adopted ClinicalBERT[13] to build our encoder-decoder model Clinical-Text-Norm-BERT (see Fig.4). ClinicalBERT is pretrained as an encoder on clinical notes to predict whether the patient will be readmitted to the hospital within 30 days of discharge from intensive care. This pretrained ClinicalBERT is used as both the encoder and decoder,

decoder weights are copied from encoder and finetuned for our target task. GPT3 is reported to be performed better than bert2bert encoder-decoder in clinical sense disambiguation task, in which acronyms are replaced with expanded text[1]. However, we have chosen the bert2bert encoder-decoder instead of bert2gpt to avoid pretraining the GPT decoder from the scratch.

We have used most of the default settings of the BERT model configuration unchanged. Only a few of the settings have been changed. Although we have changed the maximum length of the sentence from 512 to 150 to suite the needs of our data preprocessing, we have left the default learning rate 5x10-5 unchanged.

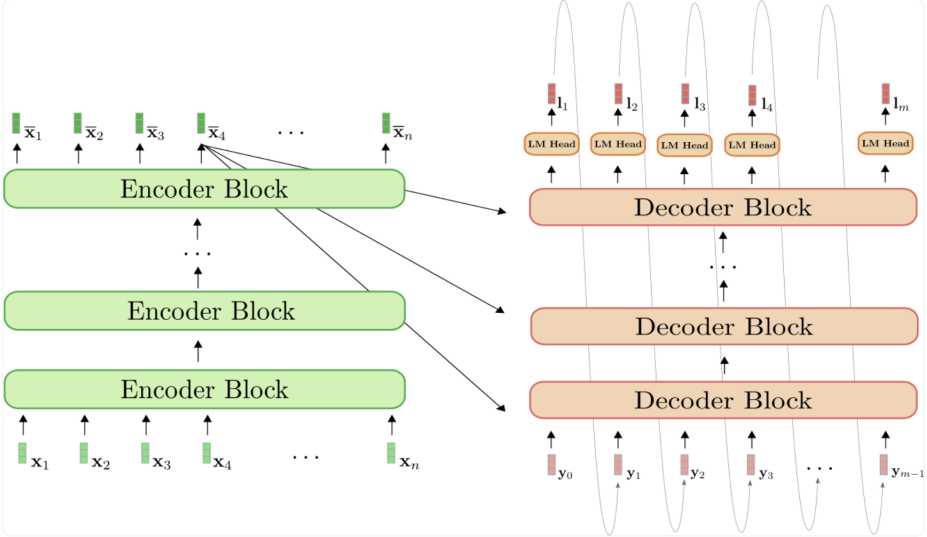


Fig. 4. Clinical-Text-Norm-BERT encoder-decoder architecture

## 5 EVALUATION

The input sentences to our encoder-decoder model are the sentences that contain abbreviations (acronyms). Labels (outputs) are the sentences with the expanded abbreviations. Similar to machine translation, we set up the model to predict the entire sentence with abbreviations expanded instead of predicting just expanded abbreviations. Some more code needs to be developed to set up the problem to predict only for the terms to be expanded so that learning will be faster than current set up.

Most of the bert model configuration default settings are unchanged except max length of input and output sentences, length penalty and 'no repeat ngram size'. After preprocessing most of the sentences are shorter i.e 95% of them are shorter than 150 words. While preparing the data for the model training, max length of sentence is set to 150 for both the input and output sentences. Later experiments are conducted with increased sentence length to around 500 words to increase the context information, however we have not seen any improvement in the model performance.

We have chosen the ROUGE metrics to evaluate the model performance. ROUGE is a set of metrics to calculate unigram, or bigram or multi-gram metrics. ROUGe-1 caluculates metrics on unigram (1-gram), ROUGE-2 measures bigram(2-gram) metrics. ROUGE-1 metrics precission, recall and F1 score on small set of the data can be seen in the

[126/126 1:59:03, Epoch 2/2]

| Step | Training Loss | Validation Loss | Rouge1 Precision | Rouge1 Recall | Rouge1 Fmeasure |
|---|---|---|---|---|---|
| 10 | 6.432100 | 23.376423 | 0.043600 | 0.021200 | 0.028300 |
| 20 | 6.309600 | 24.290266 | 0.040600 | 0.019900 | 0.026400 |
| 30 | 6.202500 | 25.589096 | 0.038400 | 0.018800 | 0.025000 |
| 40 | 6.206100 | 27.121540 | 0.035300 | 0.017400 | 0.023100 |
| 50 | 6.050000 | 26.727184 | 0.034400 | 0.016900 | 0.022400 |
| 60 | 6.132600 | 28.546398 | 0.019100 | 0.009100 | 0.012300 |
| 70 | 6.028400 | 28.022633 | 0.030100 | 0.014200 | 0.019000 |
| 80 | 5.953800 | 25.116844 | 0.045500 | 0.022200 | 0.029600 |

Fig. 5. Training metrics on small dataset

Figure 5. The model has also been trained on large dataset that contains approximately 1 million notes using gpu resources on the cluster.

As can be seen in the figure 5, training loss on small data seems to be decreasing, however on large dataset (one million notes) even after 10 epochs loss seems to be stuck around 5 and not decreasing at all. That means model learning is very slow or stuck at local minima and failing to come out of it. The maximum ROUGE-1 (unigram) precision achieved is 30% and recall and F1 score are hovering around just 10%. With ROUGE-2 (bigram) and ROUGUE-3 (trigram) metrics model completely failed and did not learn anything at all.

```
Inputs ..............................: ['no associated fever, no respiratory compromise.',
tory of "getting shaky" if etoh withdrawal.', 'no known seizure history.', 'given total of
idine (on this at baseline), thiamine/folate, b, magnesium (mg ), calcium (ionized ca ).',
spiratory compromise.', 'he was initially admitted to for dts, observation.', 'on arrival t
ory and physical exam.', 'given mg valium x and then mg.', 'labs recheckedica up to , repea
use with etoh and benzo use.', 'he drinks regularly and has intermittent binges of several
ad or happy.', "has been at several hospitals including for inpt detox from benzo's includi
Predictions~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~: ['the is : : to..', 'the : : to, and..', 'the was :
p.', 'thesp : : * : the, *, :,, of, -, mg, and, (, to, p,., ).', 'he was : : mg, the,, :, a
and : and, (, p, to,. and and..', 'no : :..', 'the was was : : the,, and,..', 'the patient
d and..', 'was was : : mg, :,, and, (,. and..']
```

Fig. 6. Screenshot of input sentences to the model and predicted sentences generated by the model

Experimented with different hyper-parameters (bert configuration settings) including length penalty and 'no repeat ngram size' showed some improvement in learning. Length penalty set to less than 0 encourages model to generate small sentences, while if it is set to greater than 1 encourages longer sentences. Length penalty default setting is zero. When we set length penalty to 1, model is paying more attention to punctuation and special characters like backslash, parenthesis, star symbols (see figure 6 for details). Setting length penalty to 3 forces model to learn to generate longer sentences, and it is learning to generate meaningful words but still performance is inadequate.

## 6 DISCUSSION

We have setup the personalized text normalization of clinical notes (EHRs) along the lines of machine translation. We give the model an input of clinical note sentences or a small paragraph that contains clinical terminology abbreviations,

and the model has to generate output sentences in which the abbreviations are expanded. Due to limited resources, we could not run extensive pretraining with clinical notes. We have preprocessed only approximately 10% of the data we collected from MIMIC-III clinical research data. After preprocessing this resulted in approximately 889,000 sentences to train the model.

There could be many reasons why model is not able learn quickly. First, due to time and resource constrains model is pretrained only on fraction of the data available to us. Second, predicting whole sentence instead of just abbreviation expansion maybe a slow process as the gradient is dispersed towards learning all the words in output sentence. Third, data preprocessing was a very challenging task and input text still has some special characters like backslash and some punctuation symbols such as quotation, colon and semicolon.

Future work will include better preprocessing of input sentences and reformulating the problem to predict only expanded abbreviation terms, and all other words in the input sentence serve as context information but will not be predicted. At present we are only expanding the abbreviations selectively to doctors, nurses, and non-medical persons, and not standardizing the numbers and fractions into text form. In the future work, we intend to normalize all the terms including fractions, numbers, and medical codes, and acronyms.

## 7 CONCLUSION

We attempted to implement personalized text normalization of clinical notes. EHR datasets MIMIC-III and n2c2 have been collected and preprocessed to cleanup the text and prepare labels for doctors, nurses, and non-medical persons by selectively expanding the abbreviations in the label sentences. We have built the ClinicalBERT encoder-decoder model architecture and fine-tuned on discharge summary notes. ROUGE-1 (unigram) precission score of 30% is achieved by the model trained on the small amount of the dataset. Further cleaning up of the data and finetuning on the larger dataset with 2.5 million notes is in progress, and the report will be updated with new results in the future.

## REFERENCES


[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Zero-Shot Clinical Information Extractors. *arXiv preprint arXiv:2205.12689* (2022).

[2] Jonathan Allen, M Sharon Hunnicutt, Dennis H Klatt, Robert C Armstrong, and David B Pisoni. 1987. *From text to speech: The MITalk system*. Cambridge University Press.

[3] Émilien Arnaud, Mahmoud Elbattah, Maxime Gignon, and Gilles Dequen. 2022. Learning Embeddings from Free-text Triage Notes using Pretrained Transformer Models.. In *HEALTHINF*. 835–841.

[4] Aiti Aw and Lianhau Lee. 2012. Personalized normalization for a multilingual chat system. In *Proceedings of the ACL 2012 System Demonstrations*. 31–36.

[5] Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédrick Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 770–779.

[6] Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. *arXiv preprint arXiv:1904.02036* (2019).

[7] Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)* 10, 3 (2007), 157–174.

[8] Grzegorz Chrupała. 2014. Normalizing tweets with edit scripts and recurrent neural embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 680–686.

[9] Peter Ebden and Richard Sproat. 2015. The Kestrel TTS text normalization system. *Natural Language Engineering* 21, 3 (2015), 333–353.

[10] Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. SemEval-2015 task 14: Analysis of clinical text. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 303–310.

[11] Hany Hassan and Arul Menezes. 2013. Social text normalization using contextual graph random walks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1577–1586.

[12] Na Hong, Andrew Wen, Feichen Shen, Sunghwan Sohn, Chen Wang, Hongfang Liu, and Guoqian Jiang. 2019. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *JAMIA open* 2, 4 (2019), 570–579.

[13] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).

[14] Gaurav Jetley and He Zhang. 2019. Electronic health records in IS research: Quality issues, essential thresholds and remedial actions. *Decision Support Systems* 126 (2019), 113137.

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

[16] Max Kaufmann and Jugal Kalita. 2010. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*, Vol. 16.

[17] Catherine Kobus, François Yvon, and Géraldine Damnati. 2008. Normalizing SMS: are two metaphors better than one?. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*. 441–448.

[18] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BEHRT: transformer for electronic health records. *Scientific reports* 10, 1 (2020), 1–12.

[19] Fei Liu, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1035–1044.

[20] Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 71–76.

[21] Xiaohua Liu, Ming Zhou, Xiangyang Zhou, Zhongyang Fu, and Furu Wei. 2012. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 526–535.

[22] Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. MCN: a comprehensive corpus for medical concept normalization. *Journal of biomedical informatics* 92 (2019), 103132.

[23] Wookhee Min and Bradford Mott. 2015. Ncsu_sas_wookhee: A deep contextual long-short term memory model for text normalization. In *Proceedings of the Workshop on Noisy User-generated Text*. 111–119.

[24] Denis Newman-Griffis, Guy Divita, Bart Desmet, Ayah Zirikly, Carolyn P Rosé, and Eric Fosler-Lussier. 2021. Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. *Journal of the American Medical Informatics Association* 28, 3 (2021), 516–532.

[25] Thomas A Oniki, Ning Zhuo, Calvin E Beebe, Hongfang Liu, Joseph F Coyle, Craig G Parker, Harold R Solbrig, Kyle Marchant, Vinod C Kaggal, Christopher G Chute, et al. 2016. Clinical element models in the SHARPn consortium. *Journal of the American Medical Informatics Association* 23, 2 (2016), 248–256.

[26] Jyotishman Pathak, Kent R Bailey, Calvin E Beebe, Steven Bethard, David S Carrell, Pei J Chen, Dmitriy Dligach, Cory M Endle, Lacey A Hart, Peter J Haug, et al. 2013. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association* 20, e2 (2013), e341–e348.

[27] Deana Pennell and Yang Liu. 2011. A character-level machine translation approach for normalization of sms abbreviations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. 974–982.

[28] Brian Roark, Richard Sproat, Cyril Allauzen, Michael Riley, Jeffrey Sorensen, and Terry Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 System Demonstrations*. 61–66.

[29] Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor, and Lawrence Cavedon. 2022. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–40.

[30] Guergana K Savova, Eugene Tseytlin, Sean Finan, Melissa Castine, Timothy Miller, Olga Medvedeva, David Harris, Harry Hochheiser, Chen Lin, Girish Chavan, et al. 2017. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer research* 77, 21 (2017), e115–e118.

[31] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics* 22, 5 (2017), 1589–1604.

[32] Michael Simmons, Ayush Singhal, and Zhiyong Lu. 2016. Text mining for precision medicine: bringing structure to EHRs and biomedical literature to understand genes and health. *Translational Biomedical Informatics* (2016), 139–166.

[33] Richard Sproat. 1996. Multilingual text analysis for text-to-speech synthesis. *Natural Language Engineering* 2, 4 (1996), 369–380.

[34] Richard Sproat. 2010. Lightly supervised learning of text normalization: Russian number names. In *2010 IEEE Spoken Language Technology Workshop*. IEEE, 436–441.

[35] Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer speech & language* 15, 3 (2001), 287–333.

[36] Richard Sproat and Keith Hall. 2014. Applications of maximum entropy rankers to problems in spoken language processing. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[37] Richard Sproat and Navdeep Jaitly. 2016. RNN approaches to text normalization: A challenge. *arXiv preprint arXiv:1611.00068* (2016).

[38] Yunqing Xia, Kam-Fai Wong, and Wenjie Li. 2006. A phonetic-based approach to Chinese chat text normalization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. 993–1000.

[39] Yi Yang and Jacob Eisenstein. 2013. A log-linear model for unsupervised text normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 61–72.

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

[41] Hao Zhang, Richard Sproat, Axel H Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization for speech applications. *Computational Linguistics* 45, 2 (2019), 293–337.

[42] Boyang Zhao. 2019. Clinical data extraction and normalization of cyrillic electronic health records via deep-learning natural language processing. *JCO Clinical Cancer Informatics* 3 (2019), 1–9.