Generating Clinical Text from Dialogue

Teppei Nakano
School of Medicine
Keio University
Shinjuku, Tokyo 1580081
keiohigh2nd@gmail.com

Abstract

Information technology changed the practice of medicine. Large volumes of clinical data are now digitized as part of routine patient care. Based on the clinical data, researchers mine and extract information to make various models such as prediction of disease onsets. Despite the increasing emphasis on collecting information in structured fields of EMRs, much of the key information needed for measuring and driving efficient medicine still resides in unstructured (free) text. The free text is noisy, sparse, and incomplete data. Thus analyzing such free text is one of the challenging research topics in computer science. However few people discuss on how such free text is generated. Here we propose a brand new method to generate free text. Generating text from conversations between a patient and a doctor can make clear rich free text. Thus analyzing such text is much easier. Our ultimate goal is to generate clear rich clinical notes from conversation and doctors don't need to take notes using keyboard.

1 Introduction

The rapid growth of information technology changed the practice of medicine. Large volumes of clinical data are now digitized. Researchers and doctors can use them easily. It is clear that Electronic Medical Record (EMR) systems changed the way we do medicine. Despite the increasing emphasis on collecting information in structured fields of EMRs, much of the key information needed for driving efficient medicine still resides in unstructured (free) text. The essence of clinical information is in the unstructured free text. Thus enthusiastic researchers have tried to mine and extract information from the text to make them structured. Once the free text becomes structured data, researchers can easily make various models including prediction of disease onsets and treatment effects.

However few people discuss on how such free text is generated. Clinical text is confusing because there are no unified formats, no standard writing styles, and no typical patients. So organizations such as Society of Internal Medicine and Journal of Internal Medicine lay out a standardized formats and styles but this approach is not working well so far because the approach completely depends on effort of human personnel.

Text has different interpretations. In order to record a state of patients, clinicians must take detailed rich notes. However in clinical practice, abbreviation and simple text is favorable because of its efficiency and usage. For instance, a patient with difficulty in breathing has a medical record such as "difficulty in breathing, SpO2 96%, FVC 50%, small shadows on chest x-p". This free text is interpretable only for primary care doctors of this patient. It is hard to imagine the patient from this text. So in the same way, analyzing this data is also difficult because the text is reduced information. We clinician should add "He has a difficulty in breathing when he carries his daughter in his arms." This information supports to show the state of this patient. Although it is desirable to make detailed text, the limited time and efficiency made it impossible. The information extracted from clinical notes is thus much diminished.

Here we point out two problems of current clinical text: 1) Not unified styles and formats, 2) Limitation of time to write notes which results in reduce debt information of patients. To these problems, there are two approaches to solve: 1) Improvement of algorithms to analyze, 2) Standardize and unify formats and generate rich text. In this paper, we propose the latter approach that generate free text from dialogue between patients and doctors shown in Fig 1.

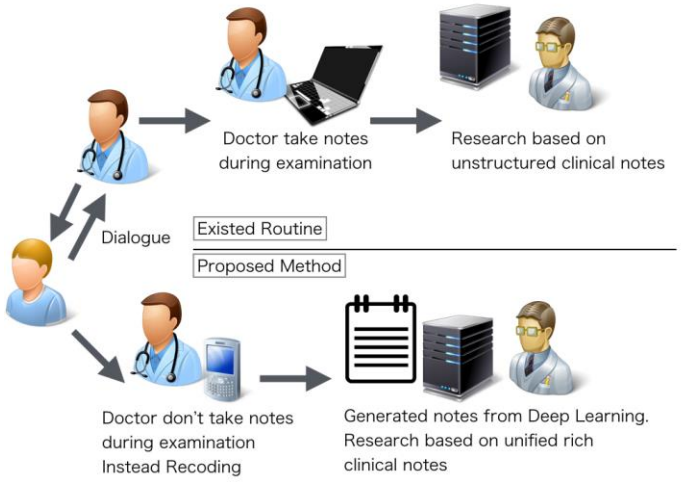

Fig 1. Overview of Our Approach

2 Related Research

There are many researches based on clinical free text. They propose a new method to convert noisy, unlabeled, abbreviated text into efficient representation such as bag of words or unified medical language system ID. Common research in this field is to extract information from free text such as medication and treatments (annotation problem) [1]. Contrary to extraction models, the research of [2] tried to predict the onset of rheumatoid arthritis disease activity from the EMR using bag of words. The other paper [3] predicts chief complaints at triage time in the emergency department. However none of the studies discuss how clinicians generate free text. If we had more rich clinical text, researchers could make more accurate and more efficient models. Thus we propose a new method of generating free text from dialogue between a patient and a doctor.

The topic we proposed here is close to meeting summarization which is not a popular task in computer science [4, 5]. There are two differences between meeting summarization and our dialogue summarization: 1) more than two people talk in a dialogue 2) clinical conversation is not diverse compared with meeting summarization.

There are many researches of dialogue response generation using deep recurrent neural networks (RNNs) with impressive results [6]. Our approach can be naturally cast as mapping an input sequence of words in a source dialogue between patients and doctors to a target sequence of words called clinical notes. Recently deep-learning based models that map an input sequence into another output sequence, called sequence-to-sequence models, have been successful in many fields such as machine translation [7], speech recognition [8], and so on. A similar model to our task is the attentional Recurrent Neural Network (RNN) encoder-decoder model proposed in [7], which produced state-of-the-art performance in machine translation (MT).

Despite the similarities, generating clinical note is a very different task from MT. Unlike in MT, the target (clinical note) is relatively short and does not depend on the length of the source (dialogue). In this point, our task is closer to generating abstracts than machine translation. In addition, a challenge in generating clinical note is to optimally compress the original dialogue in a unified manner.

3 Methods

Our model (Fig 2) is based on the attentional encoder-decoder RNN that was originally developed for machine translation [9]. We also pointed out the similarity between text summarization and generating clinical note from a conversation. Thus our model is implemented based on [10] which is the state-of-the-art model in text summarization. The encoder consists of a bidirectional GRU-RNN, while the decoder consists of a uni-directional GRU-RNN with the same hidden-state size as that of the encoder. Our model also includes switch generator. When the switch shows 'G', the existing old generator consisting of the softmax layer is used to generate a word, and when it shows 'P', the pointer network is activated to copy the word from one of the source document positions. However the large differences between [10] and ours are pre-processing of data and parameters of RNN. Since our input text is generated from human conversation, there are a lot of noises such as "Ah", "You know". Thus we removed such kind of spoken words. As for parameters of RNN, the number of encoding decoding timestep is much shorter than [10] because our data come from conversation not from a complicated newspaper story. Converting from voice of dialogue to text, we used Google Speech API [11]. The input of our model is dialogue text and the output is clinical notes.

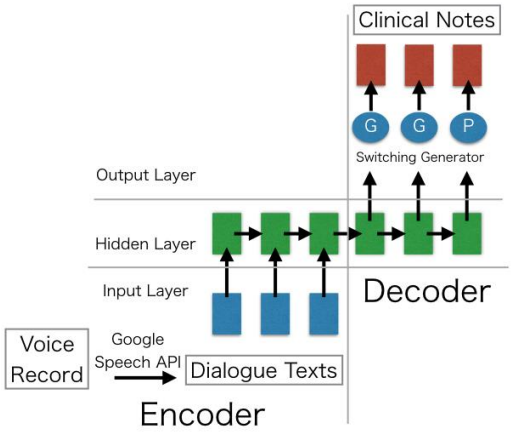

Fig 2. Our Model

4 Experiment and Result

The model in [10] is trained based on the annotated Gigaword corpus as described in [13]. They use about 3.8M training examples. Unfortunately conversation dataset in Japanese is about 800 examples which we collect as a routine practice of medicine. To reinforce the volume of training data, we also extract information from Japanese National Examination of Medical Doctor and clinical textbooks [14].

Training: For all the models we discuss below, we used 128 dimensional word2vec vectors [15] trained on the same corpus to initialize the model embeddings. The hidden state dimension of the encoder and decoder was fixed at 256 in all our experiments. The encoder vocabulary size was 42791. The encoding layer is 2, encode timestep is 60, decode time step is 15. We used Adadelta [11] for training, with an initial learning rate of 0.001. We used a batch-size of 30 and randomly shuffled the training data at every epoch, while sorting every10 batches according to their lengths to speed up training. We did not use any dropout or regularization, but applied gradient clipping. We used early stopping based on the validation set and used the best model on the validation set to report all test performance. Our training data is about 800 dialogues. We used a single CPU to train our models on this dataset. The model took under 3hours per epoch with Tensorflow implementation.

Evaluation and Results: We evaluated our models using Rouge-L metric [16]. The metric is Longest Common Subsequence based statistics which takes into account sentence level

structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically. ROUGE is a software package for automated evaluation of summaries.

The score of our model is 10.01 (Table 1). We have implemented other algorithms: rule based model and the best performance model in summarization task without preprocessing of data explained in Sec 3. The rule based model is to find sentences that include important clinical words. Results of our model are shown in Table 2. The generated notes is not accurate enough. Comparing three models, we found preprocessing of input data is effective.

Table 1. Experiment Results Based on ROUGE-L

Model	Our model	Rule Based	Best Model in Text Summarization [10] + No Preprocessing of data
Rouge-L score	**10.01**	9.58	8.14

Table 2. Experiment Results. Parenthesis is translation from Japanese to English by authors to show the results.

Question	Answer	Our Model Output	Ground Truth Notes
今日はどうされましたか (What brings you here today ?)	えっと、そうですね。頭が痛くて (Ah, you know, I have a headache)	頭痛 (Headache)	頭痛 (Headache)
いつ頃からですか (When did it start ?)	おそらく昨日の朝からだったと思います (Probably from the yesterday morning)	朝 (morning)	昨日の朝から (From the yesterday morning)
どんな痛みですか (What kind of headache? Is it both sides? One side? Squeezing?)	うーん、今日の朝からは右側なんですが、だんだん痛みが左側に動いている気がして、押さえつけられる感じもします。 (Yes, ah, but, this morning is the right side only and gradually moving to left and tighten.)	今朝から徐々に左 (From this morning gradually left)	今朝から続く頭痛は徐々に左から右へ動き締め付けられる痛み。 (Headache started this morning and it is moving from the right head to the left head and the patient felt like tighten)

5 Conclusion

In this work, we apply the attentional encoder decoder for the task of generating clinical text from dialogue between patients and doctors. Our model is not enough promising results with our datasets. This is due to small dataset and noisy conversation. Thus to increase the volume of data, we developed a simple application to collect data in routine clinical practice. Currently as for the privacy, recoding patients' voice is possible when we have an agreement from patients. As part of our future work, we plan to focus our efforts on this data and build more robust models.

References

[1] Farooq, F. Malgireddy, M, R. Yu, S. Krishinapuram, B. (2011) Extracting Medication Information from Clinical Text. ICML 2011 Workshop on Learning from Unstructured Clinical Text.

[2] Lin, C. (2013). Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PLoS ONE 8, e69932

[3] Jernite Y, Halpern Y, Horng S, Sontag D. (2013) Predicting chief complaints at triage time in the emergency department. NIPS 2013 Workshop on Machine Learning for Clinical Data Analysis and Healthcare.

[4.] Garg, N, Favre, B, Riedhammer, K. and Hakkani-Tür, D. ClusterRank: A Graph Based Method for Meeting Summarization. In Proc. of the 10th Annual Conference of the International Speech Communication (INTERSPEECH 2009), pages 1499–1502, 2009.

[5] Gillick, D. Riedhammer, K. Favre, B. and Hakkani-Tur, D. A Global Optimization Framework for Meeting Summarization. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), pages 4769–4772, 2009.

[6] Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2016). A hierarchical latent variable encoder-decoder model for generating dialogues. arXiv preprint arXiv:1605.06069.

[7]Bahdanau, D. Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

[8] Bahdanau, D. Chorowski, J. Serdyuk, D. Brakel, P. and Bengio, Y. (2015). End-to-end attention based large vocabulary speech recognition. CoRR, abs/1508.04395.

[9] Bahdanau, D. Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.

[10] Nallapati, R. Xiang, B. Bowen, Z. Noguiera, C. Gulcehre, C. (2016). Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. arXiv preprint arXiv:1602.06023.

[11] Matthew D. Zeiler. (2012). ADADELTA: an adaptive learning rate method. CoRR, abs/1212.5701.

[12] Google Cloud Speech API Documentation. https://cloud.google.com/speech/docs

[13] Alexander M. Rush, Sumit Chopra, and Jason Weston. (2015). A neural attention model for abstractive sentence summarization. CoRR, abs/1509.00685.

[14] Ministry of Health, Labour and Welfare, National Medical Practitioners Qualifying Examination Japan, http://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp140512-01.html

[15] Mikolov, T. Sutskever, I. Chen, K. Corrado, G, and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. CoRR, abs/1310.4546.

[16] ROUGE: Recall-Oriented Understudy of Gisting Evaluation A software package for automated evaluation of summaries. http://www.berouge.com/Pages/default. aspx