Clinical Text Generation through Leveraging Medical Concept and Relations

Anonymous EMNLP-IJCNLP submission

Abstract

With a neural sequence generation model, this study aims to develop a method writing patient clinical texts given brief medical history. As a proof-of-a-concept, we have demonstrated that it can be workable to use medical concept embedding in clinical text generation. Our model was based on the Sequence-to-Sequence architecture and trained with a large set of de-identified clinical text data. The quantitative result shows that our concept embedding method decreased the perplexity of the baseline architecture. Also, we discuss the analyzed results from human evaluation performed by medical doctors.

1 Introduction

Clinical texts have the potential to contain private information that can be used for identifying the patient of the document. Also, it is required social consensus and stakeholder's agreement to open and share healthcare data to the public (Kostkova et al., 2016). These barriers hinder natural language processing researchers from accessing a large-scale clinical document set.

Our aim is to utilize neural language models in producing a large amount of virtual clinical texts that are not only seemed to be written by medical doctors but also free to the mandatory of personal privacy protection as well as ownership issue. Also, the current large-scale medical corpus is based on English in terms of language, we would improve our model to apply for other languages such as Asian and European in the future.

In the medical domain, clinical text generation is in the initial stage to the best of our knowledge. A recent study demonstrated the Chinese medical record generation (Guan et al., 2019) based on generative adversarial nets (GAN) (Goodfellow et al., 2014). Though the GAN is known to generate more readable texts than other models, it is known to be unstable at the training.

Sequence-to-Sequence (Seq2Seq) (Sutskever et al., 2014) with Attention mechanism (Bahdanau et al., 2015) can be an alternative choice in terms of stability and the model's simplicity. We tested the Seq2Seq for the generation of clinical texts in this study. Seq2Seq is based on the encoder-decoder structure consisting of two recurrent neural networks (RNNs); the encoder provides the decoder a context vector consisting of summarized representations of the input. Given several sentences in the beginning part of a clinical note, our model was trained with the objective to generate the rest part of the document in this study.

Because clinicians tend to freely use medical terms in various forms for presenting an individual medical concept, we assumed medical knowledge would be required to augment the context of sentences beyond the word itself. Thus, we leveraged the medical concept unique identifiers (CUIs) from the UMLS (unified medical language system) (Lindberg et al., 1993) for the domain knowledge.

The medical domain knowledge would be provided into the model as an auxiliary. According to what (Dušek and Jurčíček, 2016) demonstrated, the auxiliary context may be either encoded by appended to the original word sequence in the form of a token sequence or separately encoded in another encoder. Thus, we tested both Seq2Seq encoder structures in order to provide the context from the CUIs appropriately for the decoder RNN. For the proof of concept, we demonstrate an approach of the clinical text generation handling clinical concepts in a medical thesaurus and embedding the concepts from the hierarchy tree.

2 Clinical text and medical thesaurus

Two databases were used in this study: MIMIC-III (Medical Information Mart for Intensive Care) and UMLS. The MIMIC-III is a large-scale health-related database (Johnson et al., 2016). It contains clinical notes having narrative descriptions of patients' previous history and current progress.

The UMLS is a thesaurus consisting of multiple medical terminology systems. The CUI is a concept identifier in the system. For the naming convention, a CUI consists of eight characters: the letter 'C' and seven numbers following the letter. The CUI in the UMLS represents the highly specific concept of a medical entity. Because a concept has multiple different names, a single CUI can be representative of multiple medical names.

We embedded the CUIs by utilizing the semantic relationship between medical concepts in the UMLS. In a simplified abstraction, we may say the relationships can be written in the form of a triple having elements as two CUIs and one relation label. For instance, 'Breast carcinoma' (C0678222) has child-parent (*PAR*) relationship with both 'Malignant neoplasms' (C0006142) and 'Breast diseases' (C0006145), then, these relationships can be written as (C0678222, C0006142, *PAR*) and (C0678222, C0006145, *PAR*). We mainly used this triple for our CUI embedding.

2.1 Data preprocessing

The de-identified clinical notes are preprocessed in multiple steps. SpaCy (Honnibal and Montani, 2017) was used in order to extract sentences from the documents. The sentence boundary detection module of the SpaCy was applied onto the text. With the syntactic parsing module, we selected narrative sentences satisfying sentence structure that contains either subject or object element.

After that, we tried to winnow sentences to have informative contents. At the first step, CUIs were extracted from sentences for mapping medical knowledge that is latent throughout a sentence into codes. MetaMap (Aronson, 2001) was utilized for identifying UMLS concepts from the text. At this process, sentences having at least one UMLS CUI were selected. Then, we extracted informative sentences in a document based on the Shannon entropy.

3 Hierarchical concept embedding

Our clinical text generation pipeline was based on the Seq2Seq. In order to make context beyond the word itself at the encoding phase, we used medical concept information that is a set of CUIs mapped to the texts. To use the concept, it was necessary to make embedding of the concepts. Figure 1 shows the abstraction of our concept embedding approach.

3.1 Concept embedding

Our neural architecture making an embedding for individual concepts consists of one input, one hidden, and one output layers. Given one concept for the input, it is trained to bring another CUI having any relationship with the input, and, jointly, to bring the corresponding relation label. The weight vector between the input and the hidden layer is used as the embedding of the CUIs.

Figure 1-a shows the concept embedding architecture. The k^{th} triple can be a form ($C_1^{(k)}$, $C_2^{(k)}$, $R^{(k)}$) and each notation is for the first CUI, the second CUI, and the relation label respectively. The set of the C_1 is obtained from the training data and the C_2s are ones having relationships with the C_1 in UMLS. We would note that our procedure is motivated by the Skip-gram (Mikolov et al., 2013): instead of recalling neighbor words in the Skip-gram, our approach pursues the objective of jointly recalling the related CUI and the relation label. Because our architecture cooperatively produces two outputs, the final cost is the average of the intermediate costs calculated with cross-entropy using one-hot encoding.

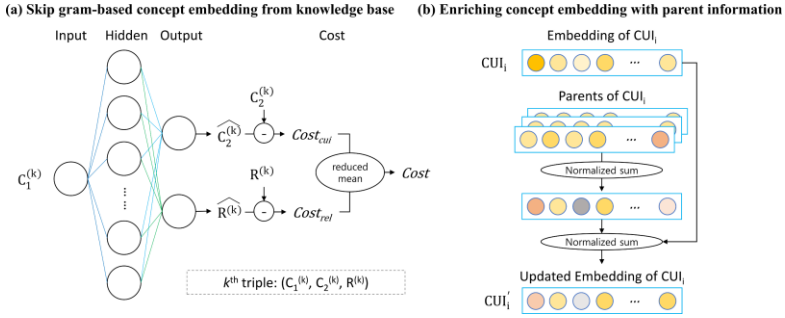

Figure 1. Graphical abstraction of the concept embedding.

3.2 Enriching the concept embedding

The concept embedding approach in the §3.1 is based on the idea that the vector representation of an individual concept should be closer to others if they share more entities in terms of a relationship. Our idea enriching the concept embedding is to make the vector of a specific concept be close with the vectors of its parents in a hierarchy tree.

This idea is motivated by FastText (Bojanowski et al., 2017). Because the skip-gram model ignores the morphology of words, the FastText makes a word vector as the sum of vectors of sub-words. Instead of summing vectors of sub-words, our hierarchical concept embedding algorithm sums up embedding vectors of parent concepts and the vector of the concept itself. We initially have a vector $\boldsymbol{c}_i$ for the i^{th} CUI from the skip-gram based concept embedding. Then, $\boldsymbol{c'}_i$, the updated vector of the $\boldsymbol{c}_i$ is calculated as follows (Eq. 1).

$$\boldsymbol{c'}_i = \frac{1}{2}\left(\boldsymbol{c}_i + \frac{1}{|Par_i|}\sum_{p \in Par_i} \boldsymbol{c_p}\right) \qquad (1),$$

where Par_i is the parent set of $\boldsymbol{c}_i$ (Figure 2-b). Embedding of unknown concepts are dynamically calculated by *l2*-normalization of the vector sum of concepts having a common prefix (length=7) with the unknown concept in addition to the vectors of the concepts that occur once.

4 Model design

The basis of our model is Seq2Seq with Attention mechanism. We noticed some sentences in the beginning part of a clinical text gives a more generalized description on a patient than the latter part. Thus, we set the encoder input as the first five sentences of a clinical note, and this is expected to provide the decoder a summarized representation of the patient's general description as a context vector. The decoder output was the rest part of the clinical note during the training time.

4.1 Incorporating medical concepts

We fed word sequence as well as a CUI sequence extracted from the source sentences with MetaMap into the encoder. The CUI is expected to lead the decoder in word selection to be close to the input in terms of semantics from the domain knowledge.

We tested two types of encoder structures demonstrated in (Dušek and Jurčíček, 2016) to provide auxiliary information. The first one is to concatenate the CUI sequence at the end of the word sequence in a single encoder (CS; Concept+Seq2Seq). The second structure is a Seq2Seq having dual encoders (CSD; Concept+SeqSeq with dual encoders). One encoder of the CSD makes a context vector from the first five sentences and another one makes a context vector from the CUI sequence. The encoded results from each encoder in the CSD are concatenated before going inward the decoder.

5 Evaluation settings and results

Our task was to generate clinical descriptions given the first introductory part of the full description. Thus, the source was the first five sentences and the target was the rest sentences. The embedding vectors were trained with 100,000 training set. FastText was used for the word embedding to cover out-of-vocabulary tokens prevalent in clinical notes. The vocabulary sizes were 46,975 and 49,758 for source and target respectively. For the concept embedding, the CUIs were from the training data, and the number of relationships extracted from the UMLS was 61,299,702 consisting of 50,942 CUIs for the C_1, 1,005,865 for the C_2, and 672 relation labels.

The language models were trained with 35,000 notes selected from the training set and validated with 8,603 notes from another set. The number of test set was 8,578. We evaluated five settings: they were named according to the method of the concept embedding. The baseline was the Seq2Seq with Attention mechanism. The baseline was trained without concepts. The others were given the sentence as well as the CUI information with different model structures and different CUI embedding methods. The second and the third models (CS, CSD) were tested in order to compare the encoder structure (single vs. dual). The CUI embedding method for the models was the skip-gram based concept embedding (§3.1). The last two models (HCSD and HCSD_T) used the hierarchical concept embedding (§3.2) in the dual encoder structure, and the last one simultaneously utilized CUI information from both of the source and the target texts. The RNN unit was three-layer Bi-LSTM and the RNN size was 400.

Table 1 shows the models' perplexity. Because the text generation task is an open-ended problem, a common evaluation method for this task is the perplexity. We observed some models using the concept showed lower perplexity than the Seq2Seq trained without the domain knowledge. The model CS using concepts in single encoder reduced the

test perplexity by 0.432 than the baseline. The HCSD using the hierarchical concept embedding in dual encoder showed subtle improvement.

Table 1. Perplexities of the clinical text generation models on MIMIC-III.

Model	Valid perplexity	Test perplexity
Seq2Seq	3.423	3.800
CS	**3.360**	**3.368**
CSD	3.822	4.195
HCSD	3.702	3.764
HCSD_T	3.830	4.197

Human evaluation with medical experts: Four human experts in Medicine evaluated the clinical notes in terms of quality. The expert group consists of four medical doctors having 10-years working experience on average. Duplicates in the generated texts were removed for this human evaluation. The questionnaire consists of two chapters: the first part asks evaluators to rate how much the generated texts are appropriate given the first five sentences in terms of clinical commonsense, and the second part asks them to identify a paragraph written by a human in a set of texts. Each part consists of ten questions (the full document is in the Appendix.)

For the first chapter, the evaluators were given separate paragraphs produced by the five models as well as human writers and independently rated texts in 5-point scale (1: very awkward, 5: strongly likely.) To prevent bias, we did not provide them the models' information as well as the fact that the paragraphs include humans' writings. Figure 2 shows boxplots summarizing the ratings from the evaluators. the plots show that the models using the concepts were rated higher than the basic model, and the best model (HCSD) achieved a median value of the rate that is equal to the median value for the human's writings (median rate=4). The models using hierarchical concept embedding also seem to achieve better performance than the models not using the hierarchy. The shape of the box for the HCSD indicates the rating scores were more closely distributed to the median than other models. An interesting point was that the humans' writings were sometimes rated low (outlier circles under the bar): we interpret this phenomenon may show the human's clinical writings sometimes do not seem to be formal and contain unaccustomed expressions even for experts.

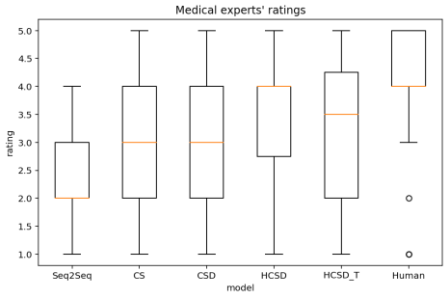

Figure 2. Box plots of the experts' ratings in terms of logical appropriateness of the clinical texts.

For the second part, the evaluators were given a set of three separate texts produced by either the generation models or human writer, and they were asked to identify a paragraph which was written by a human. We measured error rates of the evaluators for the identification of human's writing and the error rate was 65% on average (Table 2). Also, the evaluators reported they struggled to perform this task (2.63 is close to the highest level of difficulty.) For the purpose of comparison, the error rate from four non-experts are presented (who are graduate students in Biomedical Engineering.) This result may indicate the clinical text generation models can produce virtual texts seeming analogous with real ones for humans.

Table 2. Error rates (%) for the identification of human's writing and the difficulty level reported by evaluators (1: easy, 3: confused.)

	Expert	Non-expert
Error rate (average)	65	75
Level of difficulty (1-3)	2.63	2.58

Some evaluators reported the reason for their answers. Significant evidence for recognizing artificial writers were that repetitive sentence structure and ridiculous expression (e.g., '*the left ventricle is not clearly seen, but the left ventricle is not clearly seen.*'). Also, they thought a text was written by a human writer when the text contains causal relationships or when the description was in chronological order. This observation may provide clues for planning the direction of further study.

6 Conclusion

In this paper, we demonstrate a clinical text generation method based on the Seq2Seq model. Because this is a preliminary study, the current model seems to have more rooms for improvement, though, our method using concept embedding in the generation would be considered to lead the model to produce clinical texts looking the one existing in real-world. We plan to study the concept embedding method more in-depth with cutting edge models on the same task in the future.

Acknowledgments

(hidden for anonymized review)

References

Alan R Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus : The MetaMap Program. :17–21.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*, pages 1–15.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. In *Transactions of the Association for Computational Linguistics*, pages 135–146.

Ondřej Dušek and Filip Jurčíček. 2016. A Context-aware Natural Language Generator for Dialogue Systems. In *Proceedings of the SIGDIAL 2016 Conference*, pages 185–190.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NIPS2*.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2019. Generation of Synthetic Electronic Medical Record Text. In *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pages 374–380.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Alistair E W 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*, 3:160035.

Patty Kostkova, Helen Brewer, Simon de Lusignan, Edward Fottrell, Ben Goldacre, Graham Hart, Phil Koczan, Peter Knight, Corinne Marsolier, Rachel A. McKendry, Emma Ross, Angela Sasse, Ralph Sullivan, Sarah Chaytor, Olivia Stevenson, Raquel Velho, and John Tooke. 2016. Who Owns the Data? Open Data for Healthcare. *Frontiers in Public Health*, 4(February).

Donald A Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The Unified Medical Language System. *Methods of information in medicine*, 32(4):281–291.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *arXiv preprint arXiv:1301.3781*, pages 1–12.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS 2014*, pages 1–9.

7 Supplementary Material

7.1 Model training setting and model designs

7.2 Human evaluation document