Using data augmentation for NER-RE joint learning tasks for clinical history information extraction

Xiaodong Zhu, Miao Chen, Daniel Slaughter, Elizabeth Lyon, Pallavi Misra, Michael Biorn

Background

In the clinical trial and drug development industry, patient clinical history information is critical in helping determine the eligibility of a patient to be enrolled in a clinical trial. Unfortunately, a great deal of patient history is in text format, which means we need to employ natural language processing (NLP) methods to extract needed structured information. We also face the challenge of lack of labeled data for training such NLP models. In this study we introduce a deep learning based method to augment NLP tasks including named entity recognition (NER) and relation extraction (RE) to extract key oncology information from patient records. We chose to focus on extracting oncology information while the same methodology can be easily tailored to other therapeutic areas with possibly different entity and relation types.

The text of patients' clinical history contains rich and heterogeneous information such as previous diagnosis, treatments, family history, and the possible new diagnosis which need to be confirmed. Extracting information from patients' history can facilitate cohort selection for clinical trials, as well as other medical tasks such as disease diagnosis, prevention and treatment. Currently, to the best of our knowledge there is no published study on extracting information from patient clinical history or augmenting clinical history records for health NER and RE tasks. Therefore, this study is motivated by these two gaps and contributes to closing the gaps in the literature.

We have previously constructed a transformer BERT based joint learning model for NER and RE tasks which successfully extracted information from clinical trial protocols (1). Besides the helpful model architecture, the model's success also relies on sufficient training data. It would be straightforward if we retrain the same model with patient clinical history, however, we face a new challenge that we do not possess sufficient labeled data. NLP model generally requires a large training data set, which is often difficult or even impossible to obtain in the clinical domain (2). In fact, lacking training data is one of major obstacles for applying deep learning techniques in the medical domain (3). Therefore, we need to establish some efficient methods to learn from the limited clinical text.

Data augmentation is one of the approaches to cope with the limited resources issue. This method has been widely and successfully used in computer vision (4). Its application was less successful in NLP though, as text are discrete and the labels are easy to be perturbed (5, 6). It is even harder for NER and RE tasks because tokens and their entity labels are tightly coupled and therefore popular text augmentation methods such as back translation would not apply easily since word order are not preserved. Recently, multiple methods for data augmentation have been investigated including both rule-based and neural network based methods (5-7). Dai and Adel's study applied four different ways of transforming text for text augmentation purposes: Label-wise Token Replacement (LwTR), Shuffle within Segments (SiS), Synonym Replacement (SR), and Mention Replacement (MR) (8), and yielding boosted performance. Kang et al used synonyms from UMLS for data augmentation (9). Inspired by these studies, we applied three transformations: LwTR, SiS, and MR to generate the augmented data. We also performed the entity replacement not only with the UMLS synonyms, but also used the broader concepts and narrower concepts. Our results demonstrated that such data augmentation can dramatically improve the generalization of the NER-RE joint-learning model.

Methods

In order to obtain high-quality labeled data, we sought help from subject matter experts (SME) to annotate a small set of clinical history text. The SMEs looked for and highlighted oncology related entities. For the current work, we used 13 named entity types and 6 types of relations between entities. In most cases one record contained multiple entities and relations. Figure 1 shows the entity and relation counts.

We obtained 432 clinical history records in total, which served as seed data for the data augmentation tasks. We split the data into two parts, with one part of 362 records as a training set and a seed set for augmentation, with the remaining 70 records as a test set. Table 1 shows the entity and relation types and counts (before augmentation).

Table 1A. Data Counts for NER Task

Entity Type	**train**	**test**
Cancer	348	82
Condition	262	35
Other_disease	111	17
Differential_consideration	104	22
Anatomic_location	43	13
biomolecule	35	6
Qualifier_modifier	30	6
Procedure	26	9
Temporal_contraint	22	2
Remission_condition	17	6
Negation_cue	13	5
Stage	8	2
ICD	5	2

Table 1B. Data Counts for RE Task

Relation Type	**train**	**test**
Has_consideration	143	31
Modified_by	95	20
Is_located	45	15
Has_history	31	6
Has_temMea	28	3
Is_negated	13	5

We applied LwTR (Figure 1B), SiS (Figure 1C) and MR (Figure 1D) transformations for data augmentation by following the methods described by Dai and Adel (8). Briefly, for each token/segment/entity, we first sampled from binomial distribution with p =0.5 to determine whether the transformation should be performed. For LwTR, tokens with the same entity labels were sampled randomly from the training set and used as replacements. For SiS, new sentences were generated by randomly shuffling tokens within a sentence segment. For MR, entities with the same type were sampled randomly from the training set and used as the replacement. Each of these transformations generated 362 new records.

To replace a named entity with a related concept from UMLS, we focused on four entity types: Condition, Cancer, Other disease, and Anatomic Location. For each entity within these four categories, we sampled from a binomial distribution with a fixed p value to determine whether it should be replaced. We obtained three sets of data with p =0.5,0.6, and 0.7 respectively. For each set, the UMLS ontology was used to retrieve the narrower concepts (Figure 1E), broader concepts (Figure 1F), and the

synonyms (Figure 1G). If no related concepts could be found from UMLS, the replacement was skipped. In total we obtained 1,319 records using UMLS concepts.

Figure 1. Data Augmentation with Different Methods

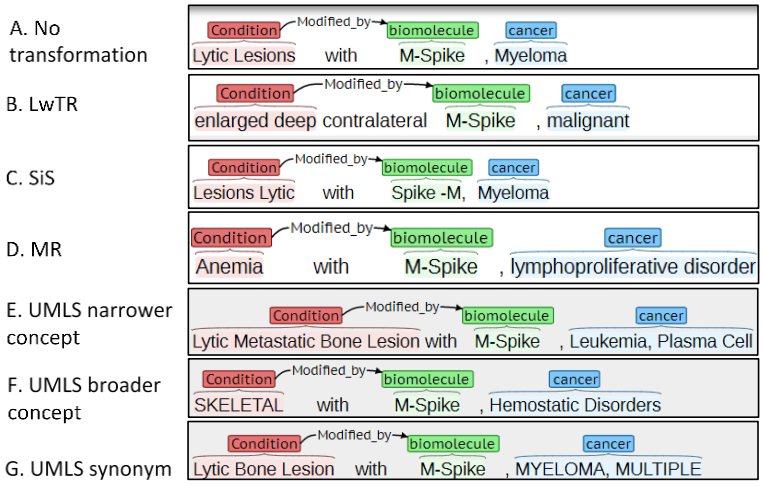

Figure 1. Data augmentation with different methods. A. original records, no transformation. B. LwTR (Label-wise Token Replacement). C. SiS (Shuffle within Segments). D. MR (Mention Replacement). Named entity were replaced. E-F. Entities were replaced by the randomly selected narrower concept in UMLS. Note that 'M-Spike' was not replaced as only entities of Condition, Cancer, Other Disease and Anatomic Location were processed

Results

As shown in table 2, training with the augmented data improved both the NER and RE tasks. NER performance was improved from f1 of 0.71 to 0.75. Interestingly, although no transformation was directly applied to the relation, we found RE performance was improved from f1 of 0.34 to 0.44. This was not surprising though, as in the joint-learning model, RE and NER tasks share the hidden representation and the loss is optimized towards both tasks. Thus, data augmentation designed for NER helps the RE task.

Table 2. NER & RE Task Performance from Models Trained with Different Data Sets

Data	Tasks	precision	Recall	F1
Original	NER performance	0.69	0.77	0.71
Original + Augmentation	NER performance	0.74	0.77	0.75
Original	RE performance	0.32	0.42	0.34
Original + Augmentation	RE performance	0.51	0.375	0.44

Conclusion

In real-world situations, it is very difficult and costly to obtain a large annotated clinical text data. This study demonstrated that data augmentation can improve both the NER and RE tasks for information

extracted from patient clinical history. The data augmentation methods used here were rule-based transformation of the original training data, and rule-based replacement of terms from the UMLS ontology. For future work, it would be interesting to experiment with neural network based data augmentation methods, as well as to evaluate how the extraction results impact downstream business tasks such as cohort selection.

References

1. Chen M, Lan G, Du F, Lobanov VS, editors. Joint Learning with Pre-trained Transformer on Named Entity Recognition and Relation Extraction Tasks for Clinical Analytics. CLINICALNLP; 2020.
2. Roberts K. Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP. Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP); Osaka, Japan2016.
3. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. npj Digital Medicine. 2019;2(1):43.
4. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019;6(1):60.
5. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, et al. A Survey of Data Augmentation Approaches for NLP. ArXiv. 2021;abs/2105.03075.
6. Chen J, Tam D, Raffel C, Bansal M, Yang D. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. ArXiv. 2021;abs/2106.07499.
7. Shorten C, Khoshgoftaar TM, Furht B. Text Data Augmentation for Deep Learning. Journal of Big Data. 2021;8(1):101.
8. Dai X, Adel H. An Analysis of Simple Data Augmentation for Named Entity Recognition. ArXiv. 2020;abs/2010.11683.
9. Kang T, Perotte A, Tang Y, Ta C, Weng C. UMLS-based data augmentation for natural language processing of clinical research literature. J Am Med Inform Assoc. 2021;28(4):812-23.