ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.174

Effective Example Extrapolation for Long-Tail Relation Extraction

Ben Li [1], Jungang Han [1, 2] and Xiaoying Pan [1, 2]

[1] Xi'an University of Posts & Telecommunications

[2] Shaanxi Key Laboratory of Network Data Analysis and Intelligence Processing

Abstract. Relationship extraction (RE) aims to identify the relationship between two entities in a sentence and is an important step to complete the knowledge graph (KG). In the medical field, the distribution of data is often unbalanced, for example, it is indisputable that the incidence of common diseases is higher than that of rare diseases, and both the size of departments and the number of cases result in an unbalanced distribution of electronic medical record data. Relationship extraction is more challenging when the relationship categories are distributed in long tails. Data augmentation is a common approach used to address category imbalance. We propose an effective example extrapolation (3E) data augmentation method: 3E generates new synthetic examples by simulating the example generation process of data-rich head relationships and extrapolating to an insufficient number of tail relationships categories. Experiments were con-ducted on the publicly available medical relationship extraction dataset 2010 i2b2/VA and compared with the upsampling meth-od to further validate its advantages in handling long-tail relationships.

Keywords: relation extraction, electronic medical record, data augmentation, example extrapolation, long-tail relations.

1. Introduction

The electronic medical record contains all the clinical medical information of the paper medical record. The extraction of relationships between entities is a fundamental and essential task that can provide data support for the subsequent construction of clinical databases and the generation of medical knowledge graphs, as in [1]. Relationship extraction aims to extract relationships between two given entities based on their related contexts. It is often expressed as a triple <E1, R, E2>, where E1 and E2 denote entities and R represents the semantic relationship between entities. Take the 2010 i2b2/VA dataset as an example, as shown in Fig. 1: the sentence contains two types of three entities DVT (problem), PE (problem), warfarin (treatment), and three sets of relationships between the three entities, namely <DVT, TrAP, warfarin>, <DVT, PIP, PE >, <PE, TrAP, warfarin> (please refer to Table I for the definition of the relationships).

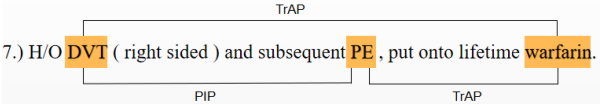

Fig. 1: An example of the 2010 i2b2/VA dataset.

Compared with traditional relationship extraction tasks, relationship extraction in electronic medical records is more usually suffers from the long-tail problem, e.g., there are not enough examples of certain relationship categories in the training data, which results in poor performance of the model in such "few" cases during testing. In the medical field, the distribution of data is often unbalanced. For example, it is undisputed that the incidence of common diseases is higher than that of rare diseases, both in terms of department size and the number of cases, resulting in an unbalanced distribution of electronic medical record data. As shown in Fig. 2, more than half of the tags in the 2010 i2b2/VA dataset, including TrWP, TrNAP, TrIP, etc., appear in less than 5% of the training data. Most recent studies on long-tail relation extraction have focused on datasets with hierarchical labels, such as the NYT [2] dataset with the relation label: /location/province/capital. Xu et al. [3] ex-ploited the interrelationships between relations to transfer knowledge from data-rich and semantically similar head relations to data-poor tail class relations. Zhang et al. [4] used GCNs to provide fine-grained relational knowledge between classes based on Xu. However, this class ap-proach cannot be applied well to other relational extrac-tion datasets not marked as hierarchical structures.

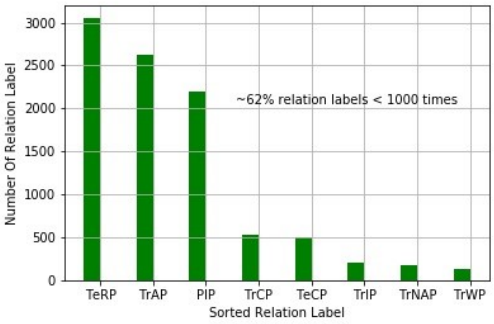

Fig. 2: Label distribution in the 2010 i2b2/VA dataset (except NA).

Data augmentation is a popular solution for category-imbalanced data, as in [5]-[7], either by replicating underrepresented examples or by synthesizing new examples using heuristics, as in [8]. However, these heuristics do not scale well and do not demonstrate the complexity of real examples well. In the field of vision, Eli Schwartz *et al*. [9] used an improved autoencoder to synthesize new examples from some examples in the classification after seeing them, thus improving the recognition of few-shot objects. In the field of NLP, Varun Kumar *et al*. [10] and Ateret Anaby-Tavor *et al*. [11] used pre-trained language models (LMs) for data expansion for text classification, where they synthesized new examples for a given label by fine-tuning the LMs.

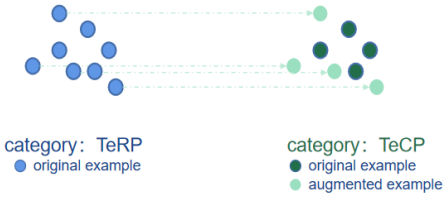

Fig. 3: Illustrates how 3E extrapolates the example from the TeRP category to the TeCP category.

An effective example extrapolation (3E) data expansion method is proposed to synthesize new examples for the tail relational category with insufficient data volume. As shown in Fig. 3, for a given relation extraction task, the inputs from the same category have some distribution in the hidden space and learn by 3E to infer new examples by simulating the generation process of example distribution in the data-rich head relation category and extrapolating to the insufficient number of tail relation categories.

We solve the problem of insufficient data volume for the tail relationship category by 3E data enhancement. And we conduct experiments on the publicly available 2010 i2b2/VA medical relationship extraction dataset to verify the feasibility and effectiveness of the proposed method and conduct comparison experiments with the baseline and upsampling methods to further validate its advantages in handling long-tail relationships.

2. Methodology

2.1. Problem Definition

In this paper, we apply 3E to the 2010 i2b2/VA medical relationship extraction dataset. First, we group the dataset according to the relationship labels and use the grouped data to train the example generative model. Then we use the generative model to generate new synthetic data for the under-represented relationships. Finally, we put the synthetic data together with the original data into the relationship extraction model for training.

We denote a piece of training data as u = (x, y), where x is the input sentences containing a pair of entity mentions and y is the category of the corresponding entity, e.g., u = ("she was treated with [steroids]e1 for [this swelling]e2 at the outside hospital...", TrIP).

2.2. Effective Example Extrapolation (3E)

We denote a piece of training data as u = (x, y), where x is the input sentences containing a pair of entity mentions and y is the category of the corresponding entity, e.g., u = ("she was treated with [steroids]e1 for [this swelling]e2 at the outside hospital...", TrIP).

3E learns by simulating the example generation process on data-rich relational groups and apply the knowlege learnd to relational groups with few data.

We consider the data with less than 5% of the data in the 2010 i2b2/VA dataset as tail relations denoted as r_{few}, and perform data expansion on these categories; the rest of the data are considered as head relations, denoted as r_{many}, which have sufficient data and will not be expanded.

We denote the true underlying distribution of an example as *p(e)*, and for group *s*, $p(e|s) \stackrel{\text{def}}{=} p(e|group_s(e) = true)$ is the true example distribution for that group, where $group_s(e)$ is a Boolean function indicating whether example e belongs to group *s*. To generalize to other groups, we characterize *s* with a random sample of K from that group, denoted as $e_{(1:k)}$. The example extrapolation task is to model the complete distribution of group *s* only when the following examples are given.

$$p(e|s) = p_{3E}(e|e_{1:k}) \tag{1}$$

Given a training set D, let $D_1, \ldots, D_s$ denote its S different groupings. Let $e_{1:k\sim D_s}$ denote the sample of K examples from D_s, drawn uniformly. The training function is :

$$\sum_{s\in M} p(s) \sum_{e^*\in D_s} E_{e_{1:k\sim D_{s\setminus e^*}}} [logp_{3E}(e^*|e_{1:k})] \tag{2}$$

where $p(s)$ is the defined prior probability for each subgroup, which we estimated empirically based on the training data from the experiments.

To optimize this function, we iterate over all training groupings ($s \in r_{\text{many}}$) and each example in each grouping (e^*), we take K examples from the same grouping ($e_{1:k}$), excluding e^*, and then we optimize the log-likelihood of e^* as the output, $e_{1:k}$ given as input.

2.3. Example Extrapolator

The generated model is designed to recover the full distribution of examples with only a small number of samples from that distribution. We implemented the example extrapolator as a neural sequence-to-sequence model. In particular, we used GPT-2 [12], a text-to-text Transformer model [13] that is pre-trained on a large text corpus. This provides the network with a large amount of world knowledge, which is crucial for the ability of the model to scale beyond a given exemplar. During the inference process (Fig. 5), the example generation model takes as input a cascade of examples from the same group in r_{few} and generates new examples belonging to the same segment.

To train the example generation model (Fig. 4), we randomly select N+1 samples from the data-rich r_{many} groups to simulate the process and optimize the log-likelihood of one of the samples to give the other N samples. The underrepresented subgroups are then inferred to generate new synthetic data, and the synthetic data are combined with the existing data and input to the relational extraction model, which is selected to be presented in Section III.

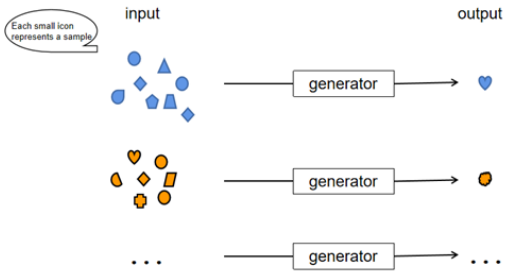

Fig. 4: Example extrapolator trained using data-rich head relations.

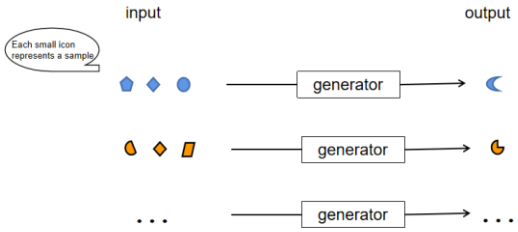

Fig. 5: Generating new synthetic data for tail relations with insufficient data volume.

3. Experiments

3.1. Dataset

We evaluate our model on the 2010 i2b2/VA [14] dataset, which is one of the recognized datasets for entity relation reviews in electronic medical records and has been widely used in recent works. The dataset, derived from three hospital discharge summaries contains eight relation types: treatment improve or cure medical problem(TrIP), treatment worsen medical problem (TrWP), treatment caused medical problems (TrCP), treatment administered medical problem (TrAP), treatment was not administered because of medical problem (TrNAP), test reveal medical problem (TeRP), test conducted to investigate medical problem (TeCP), and medical problem indicates medical problem (PIP); The three types of entities: problem, test, treatment. The problem is a phrase that describes the physical or mental abnormality of the patient as observed by the patient or the doctor, the test is a phrase that describes the further physical examination to obtain the patient's symptoms and health condition, and the treatment is a phrase that describes the treatment taken to treat the patient.

Table 1: The number of training and testing instances for each relation type in the i2b2 dataset

Class	Definition	Train size	Test size
TrIP	treatment improve or cure medical problem [a]	162	41
TrWP	treatment worsen medical problem	106	27
TrCP	treatment caused medical problems	420	106
TrAP	treatment administered medical problem	2093	524
TrNAP	treatment was not administered because of medical problem	139	35
TeRP	test reveal medical problem	2442	611
TeCP	test conducted to investigate medical problem	403	101
PIP	and medical problem indicates medical problem	1762	441
NA	none of the above	44637	11160
Total	/	52164	13046

If there are more than two entities in a sentence, one instance is created for each pair of entities. Since the part of 394 original training documents available for download contains only 170 training sets and 256 test sets, our preprocessing adopts the preprocessing steps used by Raj D. *et al*. [15]. all training and test instances were merged, and then the training and test sets were redistributed in the ratio of 8:2. Table I describes in detail the meaning of their relationship categories and the relevant statistical information after redistributing the training and test sets.

3.2. Experimental Setting

Evaluation metrics. As shown in Table I, there are eight positive relation types (predefined) and one negative relation type (in addition to the defined relationship). The performance of each positive relation type was evaluated using the precision, recall, and F1-measure. According to the official evaluation metrics [16], the performance of the model is defined based on the micro-averaged F1 scores of all positive relation types.

Relationship extraction Model. We use the CNN-Multi [17] model to achieve good relationship extraction performance even without any data expansion. To measure the contribution of the synthetic data generated by the generator, we use the exact configuration of the CNN-Multi model, the only difference being the input training data.

- Baseline: The model is trained using the original data without any expansion
- Upsampled: The model is trained based on the original data, but the examples of the tail relationship category are upsampled until their data amount reaches half of the head relationship
- 3E: The data synthesized for the tail relation category is sampled to half of the head relation, and the sampled synthesized data is put into the model for training along with the original data

Experimental environment. The experiments are based on Linux operating system, model coding is based on python 3.5, and Tensorflow deep learning framework.

3.3. Results

Table II shows the performance of the CNN-Multi model with different data expansion methods, and the proposed 3E method significantly outperforms the upsampling method with a 3% improvement in the F1 value.

Table 2: Model performance comparison

Methods	Precision/%	Recall/%	F1 score/%
Bsaeline	73.05	66.58	69.67
Upsampled	76.58	69.98	73.13
3E	78.23	75.35	76.79

Table III further shows the recognition performance of the CNN-Multi model using different data augmentation methods on the tail relationship categories. It can be found that: the classification ability of the synthetic data generated by 3E for small category samples is significantly improved compared to the baseline model, with the largest improvement of the TrWP class.

Table 3: Class-level performance of various models on i2b2 dataset (based on f1 score)

Methods	Bsaeline	Upsampled	3E
TrIP	3.82	51.02	71.43
TrWP	0	40.43	60.00
TrCP	48.14	59.57	64.97
TrAP	72.83	74.30	79.30
TrNAP	5.21	38.63	58.33
TeRP	83.02	82.61	86.91
TeCP	33.41	63.82	67.00
PIP	63.63	66.89	68.99

4. Conclusion and Future Work

In this paper, we propose a data expansion method, 3E, for alleviating the long-tail relationship extraction problem in electronic medical records. 3E learns the hidden distribution of classes from head relationships containing rich training examples and uses this knowledge to extrapolate to a smaller number of tail relationships to generate new examples. The experiments in this paper demonstrate that this is an effective method for data expansion.

For future work, we hope to apply this approach to other domains, such as images or speech, where we need to explore architectures other than pretrained seq2seq models.

5. Acknowledgment

This work was supported by the national key research and development project (No.2019YFC0121502).

6. References

[1] Wasserman, Richard C. "Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research." Academic pediatrics 11.4 (2011): 280-287.

[2] Riedel, Sebastian, Limin Yao, and Andrew McCallum. "Modeling relations and their mentions without labeled text." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2010.

[3] Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun and Peng Li. "Hierarchical relation extraction with coarse-to-fine grained attention. " Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2236-2245.

[4] Zhang, Ningyu, *et al*. "Long-tail relation extraction via knowledge graph embeddings and graph convolution networks." arXiv preprint arXiv:1903.01306 (2019).

[5] Feng, Steven Y., *et al*. "A survey of data augmentation approaches for nlp." arXiv preprint arXiv:2105.03075 (2021).

[6] Perez, Luis, and Jason Wang. "The effectiveness of data augmentation in image classification using deep learning." In Convolutional Neural Networks Vis. Recognit, 11: 1-8.

[7] Jia, Robin, and Percy Liang. "Data recombination for neural semantic parsing." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12-22.

[8] Akyürek, Ekin, Afra Feyza Akyürek, and Jacob Andreas. "Learning to recombine and resample data for compositional generalization." arXiv preprint arXiv:2010.03706 (2020).

[9] Schwartz, Eli, *et al*. "Delta-encoder: an effective sample synthesis method for few-shot object recognition." arXiv preprint arXiv:1806.04734 (2018).

[10] Varun Kumar, Hadrien Glaude, Cyprien de Lichy and William Campbell. "A closer look at feature space data augmentation for few-shot intent classification." In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pages 1 – 10.

[11] Anaby-Tavor, Ateret, *et al*. "Do not have enough data? Deep learning to the rescue!." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05. 2020.

[12] Radford, Alec, *et al*. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.

[13] Vaswani, Ashish, *et al*. "Attention is all you need." Advances in neural information processing systems. 2017.

[14] Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner. "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge." Journal of the American Medical Informatics Association 20.5 (2013): 806-813.

[15] Desh Raj, Sunil Kumar Sahu and Ashish Anand. "Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text. " In Proceedings of the 21st conference on computational natural language learning *(CoNLL 2017)*, pages 311-321.

[16] Özlem Uzuner, Brett R South, Shuying Shen and Scott L DuVall. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text." Journal of the American Medical Informatics Association 18.5 (2011): 552-556.

[17] He, Bin, Yi Guan, and Rui Dai. "Classifying medical relations in clinical text via convolutional neural networks." Artificial intelligence in text via convolutional neural networks." Artificial intelligence in medicine 93 (2019): 43-49.