Utilizing BERT for entity relationship extraction in Chinese medical texts

Yaqian Ren

School of computer science and engineering, Central South University, ChangSha, 410083, China

renyaqian@stu.hebmu.edu.cn

Abstract. The Chinese medical sector has been somewhat lacking in knowledge graphs, a deficiency this study aims to address. By leveraging the prowess of the BERT pre-training model, a two-tier approach has been innovated that utilizes separate pre-trained encoders for both entity and relational models. These models are intricately linked: the output from the entity model seamlessly flows into the relational one, making it possible to adeptly extract entity relationships from Chinese medical texts. This research is anchored in the CMeIE dataset, sourced from the esteemed CHIP (China Health Information Processing) conference. This dataset stands as a recognized benchmark in evaluating Chinese medical texts. By harnessing this data, the methods have been rigorously tested and validated. The promising experimental results underscore the effectiveness of the approach in distilling relationships from Chinese medical literature. The implications of this research are profound. Beyond just enriching the Chinese medical domain, the boundaries of NLP technology are also being pushed. Potential applications are manifold: from constructing comprehensive Chinese medical knowledge graphs to assisting in early-stage medical diagnoses. This innovative approach not only addresses an existing gap but also sets the stage for future advancements in medical NLP.

Keywords: pre-trained model, BERT, chinese medical text, entity relationship extraction.

1. Introduction

Entity and relationship extraction remains pivotal in the realm of Natural Language Processing (NLP). Over the years, numerous researchers have delved deep into the complexities of this subject. Particularly in the medical sector, gleaning relationships from both unstructured and semi-structured data stands as the linchpin of knowledge-based question-answer systems crafted for medical inquiries.

This research domain gains significant prominence when analyzing Chinese medical texts. These documents encompass a wide spectrum, from scholarly medical literature to electronic health records and other associated materials. Conventionally, deciphering these texts primarily involved two tasks: named entity recognition (NER) and relationship classification [1]. Pioneers in this field historically addressed these tasks independently. Certain models were precisely crafted for recognizing named entities, while others honed in on pinpointing the relationships between them [2]. Yet, as the discipline matured, contemporary strategies began to embrace the Entity Relational Joint Model [3], a holistic approach.

In this research, a bifurcated tactic is employed. While the merits of holistic models are recognized, there's a preference to train the named entity model and the relationship classification model individually. The reason? Segmenting them allows for intricate fine-tuning during their respective training processes, resulting in enhanced accuracy and precision. To further augment the efficacy of the models, the formidable capabilities of the BERT model are harnessed. This cutting-edge language representation model has earned accolades for its ability to discern intricate contextual embeddings of words. Courtesy of BERT, these models can craft contextually rich portrayals of both entities and their interrelations. Furthermore, details from recognized entities are adeptly incorporated into the relational model's input in subsequent phases, amplifying its capabilities. Yet, the caliber of any model hinges on the quality of its training data. Considering the limited availability of labeled data for Chinese medical texts, the CMeIE dataset becomes the cornerstone of the training regimen. Esteemed for its depth and comprehensiveness, this dataset lays a solid foundation for effective learning.

By implementing this bespoke model to the task of entity-relation extraction in Chinese medical texts, the results are commendable. The garnered performance metrics not only vouch for this approach's robustness but also highlight potential avenues for enhancement. In a realm replete with intricacies, the world of entity and relationship extraction from Chinese medical texts offers a goldmine of opportunities for NLP enthusiasts. By leveraging insights from previous works [1-3], this research seeks to fill prevailing voids and augment the ever-expanding reservoir of knowledge in this domain. The employed methodology, which fuses conventional and modern techniques powered by BERT, points towards a bright horizon for subsequent endeavors.

2. Related Work

Although the entity and relationship extraction models proposed by these authors operate independently [4, 5], there has been a significant trend towards integrated entity and relationship extraction models in contemporary research [6, 7]. For example, Shang et al. presented OneRel, an integrated entity-relationship extraction model [8]. Sui et al. leveraged a transformer with non-autoregressive parallel decoding, utilizing a set for predictions which facilitates prompt and straightforward extraction of entity relationships [9]. As research progressed, the foundational neural network architectures for these models transitioned from CNN, RNN, and GNN to the transformer. Bai et al., for instance, introduced a segmented attention mechanism to distill local semantic features from word embeddings. These features were then used for relation classification via the merging of various embedded components [10]. In a similar vein, Zheng et al. explored associating relationships. Kui et al. adopted a BERT-centric approach for combined entity and relationship classification endeavors, incorporating a dynamic range attention feature. Pushing the envelope, Wadhwa et al. harnessed expansive language models such as GPT-3 and Flan-T5 Large for their relation extraction projects. The model under discussion, rooted in BERT, excels in its simplicity and demonstrates superior efficacy on the CMeIE dataset in comparison to the previously mentioned models.

3. Research methods

As illustrated in Figure 1, the method employs BERT encoders and is composed of two primary components: an entity recognition model and a relationship classification model. Firstly, the entity recognition model identifies the type of entity for each span in the input sentence. Following this, the relationship model evaluates pairs of entities in isolation. By linking the initial positions of the two entity tags, a unique span pair representation is generated. This paired representation is then fed into a feedforward network to determine the relationship category. The foundation of this technique is rooted in the studies conducted by Zhong et al.

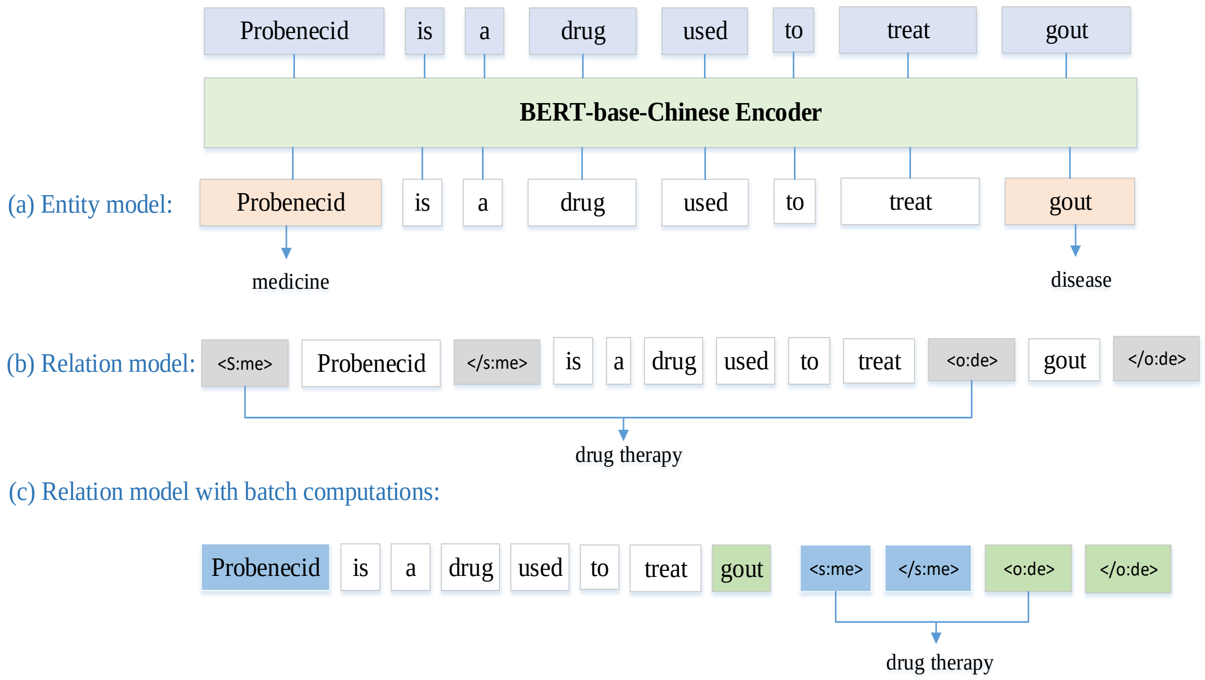

Figure 1. Structural diagram (photo/picture credit: original).

The manuscript harnesses the capabilities of the BERT-base-Chinese pre-trained model, with nuanced alterations to select hyperparameters and loss functions. This refined model is tailored for entity-relation extraction within Chinese medical literature. The evaluation process is rigorously executed using the CMeIE dataset, a distinguished academic benchmark specific to Chinese medical writings, presented by the China Conference on Health Information Processing.

4. Experiment

4.1. Dataset

This paper uses the Chinese medical text academic evaluation data set CMeIE to train the model, preprocesses and divide the data set, and performs word segmentation on the Chinese sentences. As shown in Table 1.

Table 1. The specifics of the dataset.

Dataset	sentences		
	Train	Dev	Test
CMeIE	14339	3585	3585

Table 2. Data style example of datasets.

Sentences	**Ner**						**Relations**				
	s_ sub	**e_ sub**	**ent _ type**	**s_ obj**	**e_ obj**	**ent_ type**	**s_ sub**	**e_ sub**	**s_ obj**	**e_ obj**	**relation**
Hemorrhoids and blood clots often appear after intense activity	0	2	disease	7	10	sociology	0	2	7	10	high risk factors

Table 2. (continued).

Tetanus often uses diazepam	0	2	disease	9	11	drug	0	2	9	11 medication

The data sample of the dataset is shown in Table 2.

4.2. Evaluation metrics

This article follows the standard evaluation protocol and uses the strict Micro-F1 value as the main evaluation index.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Precision is a key metric in the realm of machine learning and statistics. It gauges the proportion of positive identifications that were actually correct. To put it simply, when a model claims something is true (or positive), precision assesses how often it's right about that claim. For clarity, consider the formula for precision:

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

However, it's essential to differentiate between precision and recall. While precision is concerned with the accuracy of positive predictions, recall focuses on the proportion of actual positives that were correctly classified. Properly understanding and leveraging both these metrics can significantly enhance a model's efficacy and reliability. The F1 value can be calculated, while calculating Micro-F1 requires first calculating the total Precision and recall for all categories. The formula is as follows:

$$Precision_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{3}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{4}$$

$$F1_{micro} = 2 \times \frac{Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{5}$$

4.3. Experimental results

Utilizing the PyTorch framework, an experimental method has been developed for entity relationship extraction. When tested on the CMeIE dataset, this innovative approach demonstrates a marked improvement over prevailing techniques like BTCAMS-Syn and CasRelRoBERTa-wwm. Such advancements emphasize the evolving nature of machine learning tools and highlight the potential for more accurate and efficient extraction processes. Notably, these results underscore the importance of continuous research and development in the realm of entity relationship extraction. Through rigorous testing and comparison, it is evident that the new methodology outperforms established benchmarks, setting a new precedent for future work in the domain. As shown in Table 3.

Table 3. The result of the experiment.

Model		CMeIE	
	Prec. (%)	Rec. (%)	F1(%)
CasRel$_{\text{RoBERTa-wwm}}$	68.6	59.9	64.0
BTCAMS-Syn	64.51	57.08	60.57
OURS	79.92	76.43	78.04

5. Conclusion

The manuscript harnesses the power of the BERT-base-Chinese model, meticulously fine-tuning its hyperparameters and revamping its loss function. Rigorous experimentation with the CMeIE data reveals that this methodology yields remarkable outcomes on the CMeIE test set. Indeed, there's palpable potential for this approach in the realm of entity-relationship extraction in Chinese medical literature. Such advancements could significantly propel forward several NLP pursuits, including the construction of Chinese medical knowledge graphs, the refinement of Chinese medical Q&A systems, and the sophistication of information extraction mechanisms. Nevertheless, like any pioneering methodology, this one isn't without its constraints. Its primary limitation lies in its training data. Being solely trained on the CMeIE dataset means that it might not encapsulate the vast and intricate landscape of Chinese medical literature in its entirety. The CMeIE dataset, while invaluable, might not offer a panoramic view of all possible linguistic nuances and terminologies found in diverse medical sources. To truly serve as a lynchpin for transformative NLP applications in the Chinese medical domain, a more encompassing training approach is necessitated. To harness the full potential and ensure optimal performance across a variety of texts, future iterations of this method would be well-advised to integrate data from a broader spectrum of Chinese medical literature. Such a holistic approach would not only mitigate the risks associated with over-reliance on a singular dataset but also pave the way for an NLP tool that is robust, adaptable, and more resonant with the vast expanse of Chinese medical texts.

References

[1] Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. Computational Natural Language Learning (CoNLL), 147–155.

[2] Chan, Y. S., & Roth, D. (2011). Exploiting syntactico-semantic structures for relation extraction. Association for Computational Linguistics: Human Language Technologies (ACL-HLT), 551–560.

[3] Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., & Hajishirzi, H. (2019). A general framework for information extraction using dynamic span graphs. North American Chapter of the Association for Computational Linguistics (NAACL), 3036–3046.

[4] Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using LSTMs on sequences and tree structures. Association for Computational Linguistics (ACL), 1105–1116.

[5] Li, Q., & Ji, H. (2014). Incremental joint extraction of entity mentions and relations. Association for Computational Linguistics (ACL), 402–412.

[6] Wang, J., & Lu, W. (2020). Two are better than one: Joint entity and relation extraction with tablesequence encoders. Empirical Methods in Natural Language Processing (EMNLP).

[7] Sun, C., Gong, Y., Wu, Y., Gong, M., Jiang, D., Lan, M., Sun, S., & Duan, N. (2019). Joint type inference on entities and relations via graph convolutional networks. Association for Computational Linguistics (ACL), 1361–1370.

[8] Shang, Y., Huang, H., & Mao, X. -L. (2022). Onerel: Joint entity and relation extraction with one module in one step. CoRR, abs /2203.05412.

[9] Sui, D., Chen, Y., Liu, K., Zhao, J., & Zeng, X. (2023). Joint Entity and Relation Extraction With Set Prediction Networks. IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2023.3264735.

[10] Bai, T., Guan, H., Wang, S., Wang, Y., & Huang, L. (2021). Traditional Chinese medicine entity relation extraction based on CNN with segment attention. Neural Computing and Applications, 34(4), 2739–2748.