

分类号_____

密级_____

U D C_____

编号 10736_____

西北师范大学

硕士学位论文

(专业学位)

基于电子病历的临床医学事件

时序信息抽取技术研究

研究生姓名: 张敏钰

指导教师姓名、职称: 张志昌 教授

实践指导教师姓名、职称: 李小龙 工程师

专业学位类别: 工程硕士

专业学位领域: 计算机技术

专项计划: _____

二〇二一年五月

Research on Temporal Information Extraction of Clinical Events for EMRs

A Thesis Submitted to
Northwest Normal University
in partial fulfillment of the requirement
for the degree of
Master of Engineering
by
Zhang Minyu
Supervisor : Professor Zhang Zhichang

May, 2021

西北师范大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本学位论文引起的法律后果完全由本人承担。

学位论文作者签名：张敏钰

导师签名：张志昌

签字日期：2021年6月8日

西北师范大学学位论文版权使用授权书

本学位论文作者完全了解西北师范大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅。本人授权西北师范大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，可以公开学位论文的全部或部分内容。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：张敏红

签字日期：2021年6月8日

摘 要

基于电子病历的临床医学事件时序信息抽取是医疗领域与自然语言处理领域交叉形成的跨领域性研究任务。作为时序信息抽取技术在医疗领域的大胆尝试,基于电子病历的临床医学事件时序信息抽取受到了来自医疗领域和自然语言处理领域众多学者和专家的关注。随着深度学习技术的发展,基于电子病历的临床医学事件时序信息抽取任务的研究取得了很大的进展。尽管如此,基于电子病历的临床医学事件时序信息抽取任务的研究仍然面临着众多的挑战:(1)目前没有公开的用于临床医学事件时序信息抽取任务的语料库,研究人员需要自主构建语料库来进行实验;(2)电子病历中包含着大量晦涩难懂的医学术语,现有模型在进行临床医学事件抽取任务时忽略了上下文语境信息对于句子理解的重要性。此外,数据集类别分布不均衡也是研究中常常忽略的问题;(3)针对临床医学事件时序关系抽取任务,现有的模型仅使用句子语义来表征句子中的时序特征,忽略了句子中词语之间的依存关系对于时序关系抽取任务的重要性。因此,本文着眼于上述提到的三大挑战,对基于电子病历的临床医学事件时序信息抽取任务展开了研究。本文的主要贡献总结如下:

第一,针对临床医学事件时序信息语料库匮乏问题,本文自主构建了一个面向中文电子病历的临床医学事件信息语料库,并重构了一个面向英文电子病历的句内时序关系语料库。在制定临床医学事件信息语料库的标注规范时,本文参考了2012 I2B2 时序关系挑战赛发布的面向英文电子病历的临床医学事件时序信息语料库的标注规范,并结合了中文电子病历的行文特点对标注规范做了调整,形成了适合中文电子病历的临床医学事件信息语料库的标注规范。然后,本文根据面向中文电子病历的临床医学事件信息语料库的标注规范,开发了标注工具,利用标注工具完成了标注结果的规格化处理。最后,根据制定的语料库标注规范和开发的标注工具,在医疗领域相关人员指导下,标注人员对标注规范熟悉的情况下,完成了面向中文电子病历的临床医学事件信息语料库的构建。在面向英文电子病历的句内时序关系语料库的重构工作中,本文以2012 I2B2 时序关系挑战赛发布的面向英文电子病历的临床医学事件时序信息语料库为基础语料,并在此基础上进行了语料库的拆分和标注结果的修订。最后,以拆分和修订之后的结果构建了面向英文电子病历的句内时序关系语料库。

第二,针对现有模型对于句子上下文语境信息建模不充分、临床医学事件抽

取性能分布不均衡的问题，本文提出了一种融合预训练语言模型与对抗训练的 Encoder-Decoder 框架来提升模型性能。该方法将句子的上下文语境信息集成到汉字表征中来帮助模型理解特定语境下的词汇。此外，本文还引入了惩罚性权重来调整每个类别的损失贡献比例，以解决数据集类别分布不均衡引起的模型性能分布不均衡的问题。为了评估该模型的有效性，本文在自主构建的临床医学事件信息语料库上进行了一系列的实验。实验结果证明，本文提出的模型在整体性能和数据规模小的类别上的性能相较于本文的基线模型均有显著地提升。

第三，针对现有模型对于句子语义理解不完全、时序信息抽取不充分的问题，本文提出了一种融合图神经网络与预训练语言模型的新方法，以增强模型对于句子语义的理解能力和时序特征的抽取能力。该方法利用融合上下文语境信息的词嵌入来表征句子中的时序元素，利用融合多头注意力机制的图神经网络来捕获句子中词语之间的依存关系，利用密集连接层来更新依存关系图中的节点表征。最后，将获得的时序元素表征和句子中词语之间的依存关系表征拼接起来作为时序关系抽取的依据。为了评估该模型的有效性，本文在自主重构的临床医学事件时序关系语料库上进行了一系列的实验。实验结果证明，本文提出的模型在整体性能和子类性能上相较于本文的基线模型均有显著地提升。

关键词：电子病历；语料库构建；临床医学事件抽取；时序关系抽取；预训练语言模型；对抗训练；图神经网络

Abstract

Temporal information extraction of clinical events based on electronic medical records is a cross-disciplinary research task formed by the intersection of medical field and natural language processing field. As a bold attempt of temporal information extraction technology in the medical field, temporal information extraction of clinical events based on electronic medical records has attracted the attention of many scholars and experts from both the medical field and natural language processing field. With the development of deep learning technology, great progress has been made in the research of temporal information extraction task of clinical events based on electronic medical records. In spite of this, the research of temporal information extraction task of clinical events based on electronic medical records still faces numerous challenges: (1) currently, there is no publicly available corpus for the temporal information extraction task of clinical events based on electronic medical records, so researchers need to build a corpus independently to conduct a series of experiments; (2) electronic medical records contain a large number of obscure medical terms, and the existing models ignore the importance of contextual information for sentence understanding when performing clinical events extraction task. In addition, the unbalanced distribution of datasets is also a problem often ignored in research; (3) for the temporal relationship extraction task of clinical events, the existing models only use sentence semantic information to represent the temporal characteristics of the sentence, ignoring the importance of the dependency relationship between words in the sentence for the temporal relationship extraction task. Therefore, this thesis focuses on the three challenges mentioned above, and studies the temporal information extraction task of clinical events based on electronic medical records. The main contributions of this thesis are summarized as follows:

Firstly, aiming at the problem of lack of temporal information corpus of clinical events, this thesis constructs a clinical event information corpus for Chinese electronic medical records independently, and reconstructs a sentence-internal temporal relationship corpus for English electronic medical records. When developing a clinical event information corpus annotation specification, this thesis refers to the temporal information corpus annotation specifications for English electronic medical records released by the 2012 I2B2 temporal relations challenge, and combines with the characteristics of Chinese

electronic medical records to make adjustment for annotation specification, forms an annotation specification of clinical event information corpus which is suitable for Chinese electronic medical records. Then, according to the annotation specification of the clinical event information corpus for Chinese electronic medical records, this thesis develops an annotation tool, and uses the annotation tool to complete the normalization processing of the annotation results. Finally, according to the corpus annotation specification and the annotation tool, under the guidance of relevant medical personnel in the medical field, and under the condition that the annotation personnel are familiar with the annotation specification, the construction of the clinical event information corpus for Chinese electronic medical records is completed. When reconstructing the sentence-internal temporal relationship corpus for English electronic medical records, this thesis takes the temporal information corpus of clinical events for English electronic medical records published by the 2012 I2B2 temporal relations challenge as the basic corpus, and on this basis, we split the corpus and revise the annotated results. Finally, a sentence-internal temporal relationship corpus for English electronic medical records is constructed based on the decomposed and revised results.

Secondly, aiming at the problem of inadequate modeling of sentence's contextual information and unbalanced performance distribution of clinical event extraction for the existing models, this thesis proposes an Encoder-Decoder framework that integrates pre-trained language model and adversarial training to improve the performance of model. The method integrates the contextual information of the sentence into the Chinese character representations to help the model understand the words in a specific context. In addition, this thesis also introduces punitive weights to adjust the loss contribution ratio of each category to solve the problem of unbalanced performance distribution caused by unbalanced category distribution of datasets. In order to evaluate the effectiveness of this model, a series of experiments are carried out on the self-constructed clinical event information corpus. Experimental results show that the proposed model has a significant improvement on the overall performance and the performance of categories with small data scale compared with the baseline models in this thesis.

Thirdly, aiming at the problem of insufficient understanding of sentence semantic information and insufficient extraction of temporal information for the existing models, this thesis proposes a new method combining graph neural network and pre-trained language model to enhance the ability of the model to understand sentence semantic

information and extract temporal features. In this method, word embedding which integrates contextual information is used to represent the temporal elements in the sentence, graph neural network which integrates multi-head attention mechanism is used to capture the dependency relationships between words in the sentence, and dense connection layer is used to update the node representations in the dependency relationship graph. Finally, the representation of temporal elements and dependency relationship between words in the sentence are concatenated together as the basis for the temporal relationship extraction. In order to evaluate the effectiveness of this model, a series of experiments are conducted on the self-reconstructed temporal relationship corpus of clinical events. Experimental results show that the proposed model has significantly improved on the overall performance and the subclass performance compared with the baseline models in this paper.

Keywords: Electronic Medical Records; Corpus Construction; Clinical Event Extraction; Temporal Relationship Extraction; Pre-trained Language Model; Adversarial Training; Graph Neural Network

目 录

| | |
|------------------------------------|----|
| 第 1 章 绪论 | 1 |
| 1.1 研究背景及意义 | 1 |
| 1.2 研究现状 | 3 |
| 1.2.1 基于电子病历的临床医学事件抽取 | 3 |
| 1.2.2 基于电子病历的临床医学事件时序关系抽取 | 4 |
| 1.3 本文研究内容 | 5 |
| 1.4 论文结构安排 | 6 |
| 第 2 章 相关任务与知识介绍 | 7 |
| 2.1 相关任务介绍 | 7 |
| 2.1.1 基于电子病历的临床医学事件抽取 | 8 |
| 2.1.2 基于电子病历的临床医学事件时序关系抽取 | 9 |
| 2.2 相关技术介绍 | 10 |
| 2.2.1 Transformer | 10 |
| 2.2.2 BERT | 13 |
| 2.2.3 图神经网络 | 16 |
| 2.3 系统性能评估 | 17 |
| 2.3.1 评价指标 | 17 |
| 2.4 本章小结 | 18 |
| 第 3 章 基于电子病历的临床医学事件时序信息语料库构建 | 19 |
| 3.1 基于电子病历的临床医学事件信息语料库构建 | 19 |
| 3.1.1 语料选取 | 19 |
| 3.1.2 标注规范 | 19 |
| 3.1.3 标注示例 | 23 |
| 3.1.4 标注结果规格化 | 23 |
| 3.1.5 语料库标注 | 24 |
| 3.1.6 语料库数据统计与分析 | 25 |
| 3.2 基于电子病历的临床医学事件时序关系语料库重构 | 25 |
| 3.2.1 语料选取 | 25 |
| 3.2.2 时序关系分类体系 | 26 |
| 3.2.3 语料库重构 | 28 |

| | |
|---|-----------|
| 3.2.4 语料库数据统计与分析 | 29 |
| 3.3 本章小结 | 30 |
| 第4章 融合对抗训练与预训练语言模型的事件抽取 | 31 |
| 4.1 引言 | 31 |
| 4.2 任务描述 | 32 |
| 4.3 模型介绍 | 32 |
| 4.3.1 语义编码层 | 33 |
| 4.3.2 标签解码层 | 34 |
| 4.3.3 对抗训练的损失函数 | 34 |
| 4.4 实验结果与分析 | 34 |
| 4.4.1 数据集与评价指标 | 34 |
| 4.4.2 实验设置与基线模型 | 35 |
| 4.4.3 整体性能分析 | 37 |
| 4.4.4 子类性能分析 | 38 |
| 4.4.5 消融实验结果分析 | 40 |
| 4.5 本章小结 | 41 |
| 第5章 融合图神经网络与预训练语言模型的时序关系抽取 | 42 |
| 5.1 引言 | 42 |
| 5.2 任务描述 | 43 |
| 5.3 模型介绍 | 43 |
| 5.3.1 时序元素编码层 | 44 |
| 5.3.2 时序关系编码层 | 45 |
| 5.3.3 时序关系解码层 | 47 |
| 5.4 实验结果与分析 | 48 |
| 5.4.1 数据集与评价指标 | 48 |
| 5.4.2 实验设置与基线模型 | 49 |
| 5.4.3 整体性能分析 | 50 |
| 5.4.4 消融实验结果分析 | 50 |
| 5.5 本章小结 | 51 |
| 结 论 | 52 |
| 参考文献 | 54 |
| 致 谢 | 58 |
| 个人简历、在学期间发表的学术论文及研究成果 | 60 |

第1章 绪论

本章首先介绍基于电子病历的临床医学事件时序信息抽取任务的研究背景及意义；其次梳理该任务的研究现状，并介绍本文的研究内容；最后描述本文的结构安排。

1.1 研究背景及意义

随着大数据、云计算、自然语言处理等技术的普及，医疗健康服务行业开始融合互联网技术，朝着信息化、智能化的方向快速发展。电子病历是信息技术和网络技术在医疗领域的产物，内容包括出院小结和首次病程记录等。其中，出院小结是对患者治疗过程和治疗效果的总结，首次病程记录是对患者临床表现、入院检查和入院治疗过程的阶段性记录。出院小结和首次病程记录囊括了患者就医过程中的关键信息，属于电子病历中最重要的两类文本。如何从这两类文本中有效地提取人们关心的内容，成为了当前亟待解决的一个问题。基于电子病历的临床医学事件时序信息抽取任务就是在这样一个背景下产生的，目的就是高效地、自动化地提取电子病历中与患者临床时间线相关的所有内容，并以结构化的形式呈现出来。图 1-1 展示了临床医学事件时序信息抽取任务的整体框架，主要包含三个子任务，分别是时间表达式抽取、临床医学事件抽取和时序关系抽取。其中，时间表达式抽取和临床医学事件抽取属于基础性任务，时序关系抽取是建立在时间表达式抽取和临床医学事件抽取之上的具有一定难度的推理性任务。

时间表达式抽取任务旨在提取电子病历中包含的时间信息。电子病历中的时间信息是对临床医学事件时序性的直接表述，是推断时间表达式与临床医学事件之间或者临床医学事件与临床医学事件之间时序关系的依据。图 1-1 中包含的时间表达式有“2016-11-13 10:12”、“今晨”和“近日”，对应的时间表达式类型分别为“日期”、“一天中的部分”和“过去、现在、将来”。其中“日期”类型的时间表达式“2016-11-13 10:12”反映了患者入院检查的时间，“一天中部分”类型的时间表达式“今晨”反映了患者入院就诊的时间，“过去、现在、未来”类型的时间表达式“近日”反映了患者身体不适持续的时间。

临床医学事件抽取任务旨在提取与患者临床时间线相关的所有活动。电子病历中的临床医学事件是构成电子病历的核心元素，是时序信息的携带者，是构成时序关系的主体元素。图 1-1 中包含的临床医学事件有“就诊”、“自述”、“腹痛”、

“缓解”，“恶心”、“呕吐”、“腹泻”和“便秘”。其中“就诊”为特定类型的临床医学事件，“自述”为证据词类型的临床医学事件，“缓解”为体类型的临床医学事件，“腹痛”、“恶心”、“呕吐”、“腹泻”和“便秘”为问题类型的临床医学事件。其中，特定类型的临床医学事件“就诊”反映了患者身体不适时采取的措施。证据词类型的临床医学事件“自述”反映了信息的来源。体类型的临床医学事件“缓解”反映了上一个问题类型的临床医学事件“腹痛”的下一步发展方向。问题类型的临床医学事件“腹痛”、“恶心”、“呕吐”、“腹泻”和“便秘”反映了患者的身体状况。

时序关系抽取任务旨在提取电子病历中包含的时序关系，包括时间表达式与临床医学事件之间的时序关系和临床医学事件与临床医学事件之间的时序关系。电子病历中的时序关系反映了患者就诊期间相关事件的演变过程。图 1-1 左下角的部分为依据图中电子病历片段构建的患者临床时间线，该时间线清晰地展示了患者入院治疗期间相关临床医学事件随时间的变化过程。其中，时间表达式“近日”范围内发生的临床医学事件有“恶心”、“呕吐”、“腹泻”、“便秘”和“腹痛”，时间表达式“今晨”范围内发生的临床医学事件有“就诊”、“自述”和“缓解”。时间表达式“近日”范围内发生的临床医学事件全部发生在时间表达式“今晨”范围内发生的临床医学事件之前。

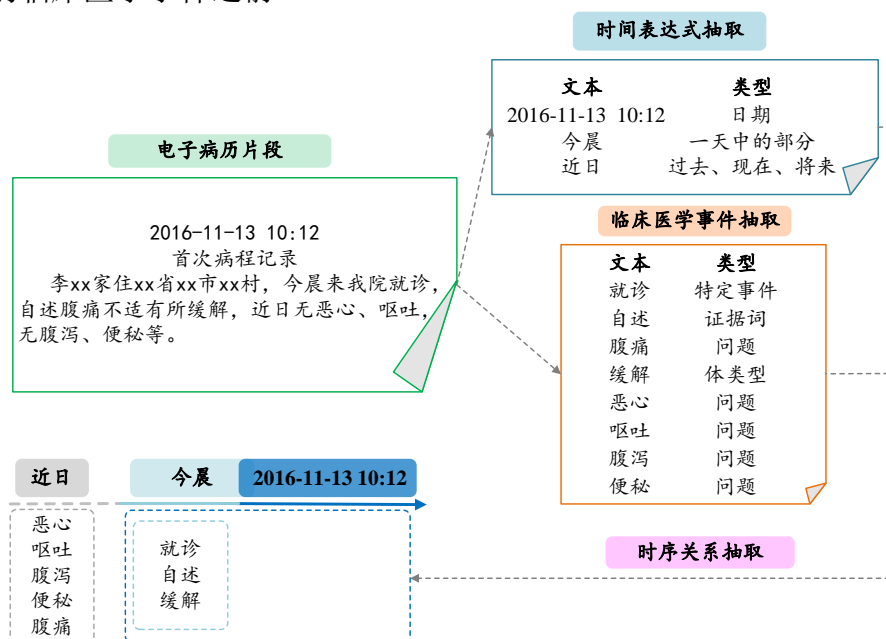


图 1-1 临床医学事件时序信息抽取任务的整体框架

通过上述例子，可以发现一篇电子病历往往是由多个临床医学事件构成，这些临床医学事件囊括了患者入院治疗期间的关键信息。临床医学事件往往不是孤立存在的，而是依据某个时间点展开，并与其他时间点和其他临床医学事件有着密切联系，这种相互关系反映了临床医学事件之间或者临床医学事件与时间表达式之

间的时序关系。抽取电子病历中包含的时序关系能够帮助医生梳理电子病历中的离散事件,实现临床医学事件的演变和发展方向的预测。因此,研究临床医学事件时序信息抽取的相关问题,不仅具有理论意义,还具有一定的实用价值。

1.2 研究现状

本小节将对基于电子病历的临床医学事件时序信息抽取任务的两个子任务临床医学事件抽取和时序关系抽取的研究现状进行梳理。

1.2.1 基于电子病历的临床医学事件抽取

临床医学事件抽取属于序列标注任务,旨在抽取电子病历中包含的临床医学事件片段,并为抽取出来的临床医学事件片段分配相应的边界标签和类别标签。目前,对于事件抽取任务的研究主要集中在开放域文本上,而在电子病历上的研究较少。考虑到面向开放域文本的命名实体识别和面向电子病历的医疗实体识别与临床医学事件抽取任务有相似之处,目前对于临床医学事件抽取任务的研究主要借鉴上述两个任务的研究方法。针对上述两个任务,现有的方法大致可以分为四类:基于规则的方法、基于字典的方法、传统机器学习的方法和深度学习的方法。

基于规则的方法依赖领域专家的经验 and 知识归纳总结出的相关规则,利用相应的规则模板作为分类器来抽取医疗实体。当规则足够详细时,系统性能较好,一旦出现规则之外的医疗实体,系统将无法识别。基于规则的方法缺乏灵活性,通常只针对特定任务,很难移植到其他任务,对于其他任务需要重新制定规则。由于方法的简单,早期的医疗实体识别系统主要采用该方法。例如,Kraus 等人针对大学医疗系统的临床记录,通过构建正则表达式,对其中提及的药品、剂量、服用方法等医疗实体进行了抽取^[1]。

基于字典的方法不需要标注好的训练数据,依赖于词汇资源和领域知识来识别医疗实体^[2]。利用该方法构建的系统需要特定领域的专家定期进行维护,系统维护的成本较高。基于字典的医疗实体识别系统有 IBM 的 MedKAT (Medical Knowledge Analysis Tool)^[3]和梅奥诊所的 CTAKES (Clinical Text Analysis and Knowledge Extraction System)^[4],这两个系统均基于 UIMA 框架^[5],对电子病历的预处理基本相同,包括句子边界识别、词性标注、浅层句法分析。MedKAT 利用癌症 ICD 抽取癌症病历中的疾病概念,并提出癌症疾病知识表示模型(Cancer Disease Knowledge Representation Model, CDKRM),表示抽取的疾病以及疾病关系的知识。CTAKES 利用 UMLS、SNOMED-CT 和 RXNORM 抽取电子病历中的概念。

传统机器学习的方法不需要依靠语言学知识构建规则，取而代之的是使用标注好的语料库进行训练，从语料中对特征进行统计，并建立模型。常用的方法有隐马尔科夫(Hidden Markov Models, HMMs)^[6]、朴素贝叶斯(Naive Bayes, NB)、支持向量机(Support Vector Machine, SVM)^[7]、条件随机场(Conditional Random Field, CRF)^[8]等。例如，Wang 等人分别采用条件随机场和支持向量机两种方法，并结合 SNOMED-CT 词典进行医疗实体识别，其中条件随机场模型达到了较好的效果^[9]。

随着深度学习的发展，深度学习的方法被引入到命名实体识别任务。例如，Tang 等人使用卷积神经网络(Convolutional Neural Networks, CNN)获得中文汉字的表征，并将其传入双向长短期记忆网络(Bidirectional Long Short Term Memory, BiLSTM)进行特征抽取，最后使用条件随机场进行解码，给句子中的每一个汉字预测一个边界标签和实体类型标签^[10]。Cao 等人引入对抗性迁移学习框架来联合训练中文命名实体识别任务和中文分词任务，旨在识别来自中文分词任务与中文命名实体识别任务共享的边界信息^[11]。上述方法虽然取得了不错的成绩，但是仍然存在一些缺陷。现有方法通过引入外部特征来增强词语的表征，这种策略虽然一定程度上给予了模型额外的知识支持，但是每次引入的特征并不总是会为模型带来好的影响，有时还会给模型带来错误信息。上下文语境信息属于模型根据自身信息产生的额外知识，这部分知识不受外部条件的限制，为句子中词汇的理解提供了强有力的支持，而这部分信息往往被现有模型忽略。此外，考虑到电子病历与其他文本之间是存在差异的，具有更强的领域性，对于电子病历片段的理解需要上下文语境信息的支持。因此，如何利用上下文语境信息来理解电子病历片段就变得极为重要。

1.2.2 基于电子病历的临床医学事件时序关系抽取

临床医学事件时序关系抽取一般被视为多分类任务，旨在判断电子病历片段中的两个临床医学事件或者一个临床医学事件和一个时间表达式之间是否存在时序关系，以及存在何种时序关系。目前，对于事件时序关系抽取任务的研究主要集中在开放域文本上。因此，对于临床医学事件时序关系抽取任务的研究主要借鉴面向开放域文本的事件时序关系抽取的相关方法。针对开放域文本的事件时序关系抽取任务的研究方法大致分为两类，分别是传统机器学习的方法和深度学习的方法。

早期的事件时序关系抽取任务的研究工作专注于提取文本中各种各样的特征。Daniel 等人提出了一种将单词配对作为有用特征的方法来提高实验性能^[12]。随着 TimeML(Time Markup Language)^[13]标注体系的建立和发展以及 TimeBank 时序信

息语料库和 Opinion 时序信息语料库的出现,大量研究人员将基于特征工程的统计机器学习方法应用在事件时序关系抽取任务上。Inderjeet 等人在 TimeBank 语料库和 Opinion 语料库的基础上利用事件属性构造特征向量,包括事件的类型、体态、形态、极性和时态等,并使用最大熵分类器进行时序关系抽取^[13-14]。

随着深度学习技术的发展,各种各样的神经网络模型被引入到自然语言处理领域的相关任务中。深度学习框架在实体关系抽取任务上取得了巨大进展^[15],研究人员受此启发,开始将其引入到事件时序关系抽取任务中。Cheng 等人^[16]根据 Xu 等人^[15]在实体关系抽取任务上的研究工作,将词语之间的最短依存路径作为模型的输入,构造了一种基于循环神经网络(Recurrent Neural Networks, RNN)的神经网络模型,在没有使用任何显性特征和外部资源的情况下,取得了非常出色的结果。上述方法虽然在事件时序关系抽取任务上取得了不错的成绩,但是仍然存在一些缺陷。该方法虽然已经将词语之间的依存关系引入到事件时序关系抽取任务中,但是对原始的依存关系图做了剪枝,导致模型对于依存关系捕获不完全,时序特征抽取不充分。

1.3 本文研究内容

目前,临床医学事件时序信息抽取任务的研究存在三大挑战,分别是:(1)目前没有公开的用于临床医学事件时序信息抽取任务的语料库,研究人员需要自主构建语料库来进行实验;(2)电子病历中包含着大量晦涩难懂的医学术语,现有模型在进行临床医学事件抽取任务时忽略了上下文语境信息对于句子理解的重要性。此外,数据集类别分布不均衡也是研究中常常忽略的问题;(3)针对临床医学事件时序关系抽取任务,现有的模型仅使用句子语义表征句子中的时序特征,忽略了句子中词语之间的依存关系对于时序关系抽取任务的重要性。针对上述的三大挑战,本文开展了以下三方面的工作:

针对临床医学事件时序信息语料库匮乏问题,本文参考了 2012 I2B2 时序关系挑战赛发布的面向英文电子病历的临床医学事件时序信息语料库的标注规范,并结合中文电子病历的行文特点,制定一份适合中文电子病历的临床医学事件信息语料库的标注规范,并开发了标注工具。在标注人员对标注规范熟悉的情况下,完成了面向中文电子病历的临床医学事件信息语料库的构建。此外,本文在 2012 I2B2 面向英文电子病历的临床医学事件时序信息语料库的基础上,重构了一个句内时序关系语料库。

针对现有模型对于句子上下文语境信息建模不充分、临床医学事件抽取性能分布不均衡的问题,本文提出了一种融合预训练语言模型与对抗训练的 Encoder-

Decoder 框架。该方法将句子的上下文语境信息集成到汉字表征中,以帮助模型理解特定语境下的词汇。此外,还引入惩罚性权重来调整每个类别的损失贡献比例,以解决数据集类别分布不均衡引起的模型性能分布不均衡的问题。

针对现有模型对于句子语义理解不完全、时序信息抽取不充分的问题,本文提出了一种融合图神经网络与预训练语言模型的方法,以增强模型对于句子语义的理解能力和时序特征的抽取能力。该方法利用融合上下文语境信息的词嵌入表征句子中的时序元素,利用融合多头注意力机制的图神经网络捕获句子中词语之间的依存关系,以此建立句子中词语之间的关联性。此外,还引入了密集连接层,并结合注意力指导层更新依存关系图中的节点嵌入,从而优化依存关系图的表征。融合上下文语境信息的单词表征从宏观角度为模型提供了时序特征,依存关系图从微观角度考虑句子的结构信息对于时序关系抽取的重要性。

1.4 论文结构安排

本文主要分为六个章节,内容如下:

第 1 章为绪论,该章介绍了基于电子病历的临床医学事件时序信息抽取任务的研究背景及意义,梳理了临床医学事件时序信息抽取的研究现状,提出了本文的主要研究内容和论文组织框架。

第 2 章为相关任务与知识介绍,该章对研究过程中涉及到的理论基础和相关技术进行了集中介绍,内容包括基于电子病历的临床医学事件抽取和基于电子病历的临床医学事件时序关系抽取任务的介绍,研究过程中使用到的深度学习技术和系统性能评价指标。

第 3 章为基于电子病历的临床医学事件时序信息语料库构建,该章完成了临床医学事件时序信息语料库构建的两个任务:临床医学事件信息语料库构建和临床医学事件时序关系语料库重构。

第 4 章为融合对抗训练与预训练语言模型的事件抽取,该章针对现有模型对于句子上下文语境信息建模不充分、临床医学事件抽取性能分布不均衡的问题,提出了一种融合对抗训练与预训练语言模型的 Encoder-Decoder 框架。为了评估本章提出的模型的有效性,在自主构建的临床医学事件信息语料库上进行了实验验证。

第 5 章为融合图神经网络与预训练语言模型的时序关系抽取,该章针对现有模型对于句子语义理解不完全、时序信息抽取不充分的问题,提出了一种融合图神经网络与预训练语言模型的方法。为了评估本章提出的模型的有效性,在自主重构的临床医学事件时序关系语料库上进行了一系列的实验。

结论,该章对本文工作进行了总结,并对后续工作进行了展望。

第2章 相关任务与知识介绍

本章将对研究过程中涉及到的理论基础和相关技术进行集中介绍，内容包括基于电子病历的临床医学事件抽取和基于电子病历的临床医学事件时序关系抽取任务的定义，研究过程中使用到的深度学习技术和模型性能评价准则。

2.1 相关任务介绍

目前，学术界和工业界将基于电子病历的临床医学事件时序信息抽取技术的研究划分为三个层次^[17]，如图 2-1 所示。第一层是基于电子病历的临床医学事件时序信息语料库构建，属于底层基础建设性研究，目的是为上层研究提供扎实的数据基础。第二层是基于电子病历的临床医学事件时序信息抽取，包括时间表达式抽取、事件抽取和时序关系抽取三个子任务。第三层是患者临床时间线构建，属于建立在第一层和第二层之上的应用性研究。目前，学术界的关注点主要集中在第一层和第二层的基础性研究，工业界则更多地关注第三层的应用性研究。

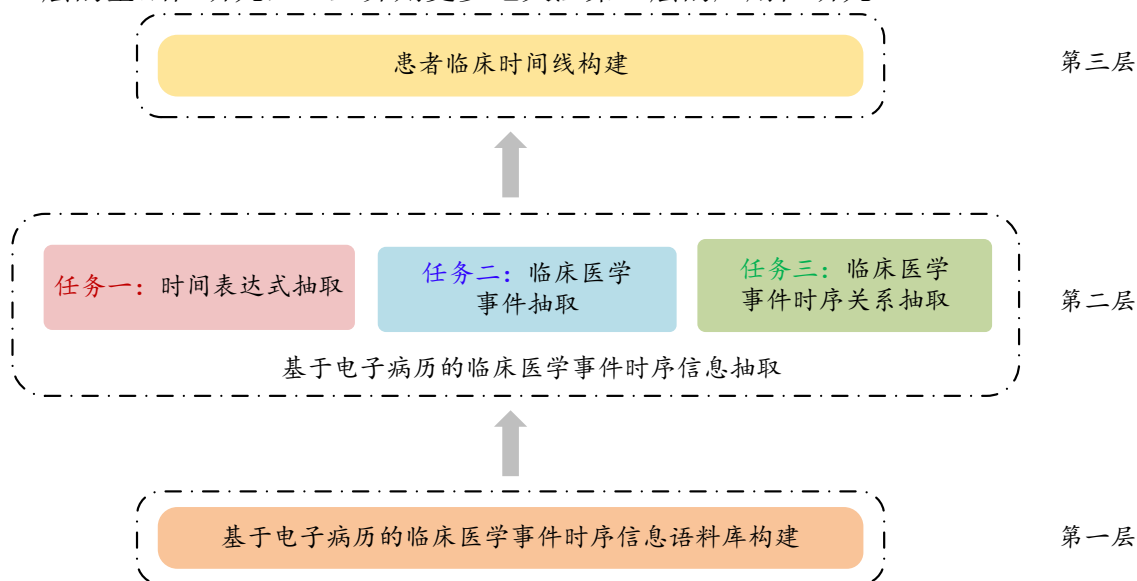


图 2-1 基于电子病历的临床医学事件时序信息抽取技术研究的三个层次

本文的研究工作主要集中在第一层和第二层，包括基于电子病历的临床医学事件时序信息语料库的构建和基于电子病历的临床医学事件时序信息抽取技术的相关内容。具体的研究工作会在后续章节详细介绍。接下来将简单介绍基于电子病历的临床医学事件抽取和基于电子病历的临床医学事件时序关系抽取任务。

2.1.1 基于电子病历的临床医学事件抽取

基于电子病历的临床医学事件抽取是基于电子病历的临床医学事件时序信息抽取的子任务，旨在提取电子病历中包含的临床医学事件。抽取出来的临床医学事件囊括了患者的既往病史、现病史、家族史、入院检查、医生诊断、手术、药物治疗等多个方面的信息。对这些资源进行加工处理可作为辅助信息帮助医生进行临床决策。本文主要研究七种类型的临床医学事件抽取，分别是问题类型、检查类型、治疗类型、临床科室类型、证据词类型、特定类型和体类型。

下面以一个简单的例子来说明基于电子病历的临床医学事件抽取任务的主要内容。图 2-2 是一份中文电子病历的首次病程记录的部分内容。

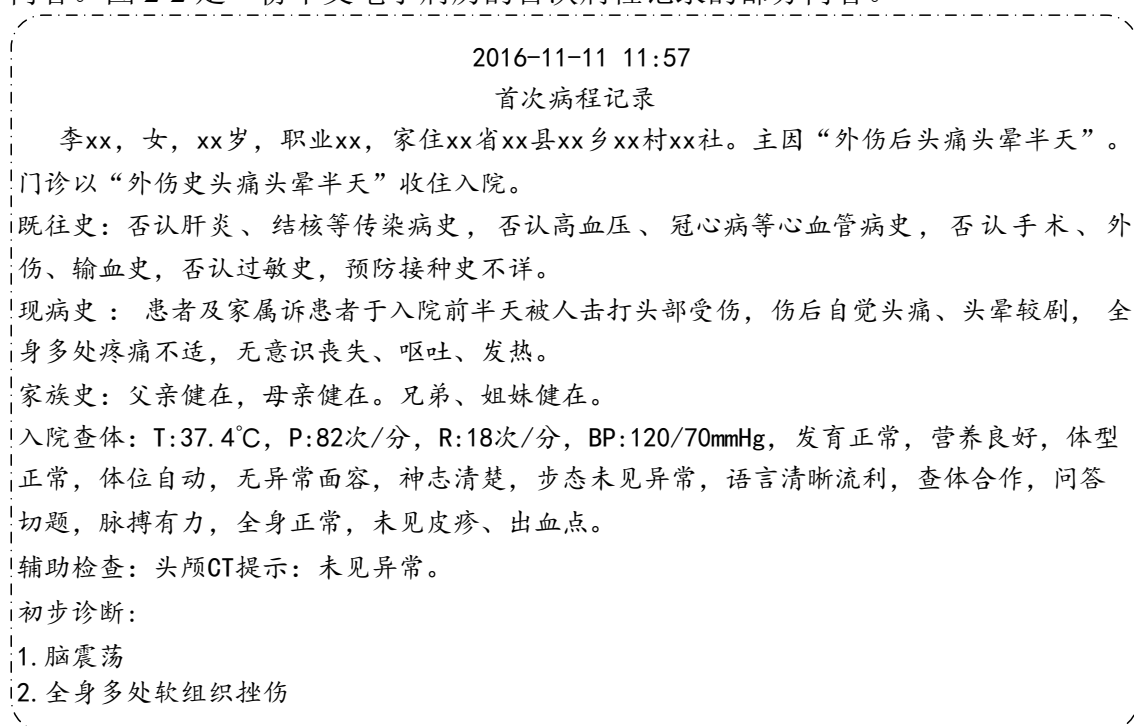


图 2-2 中文电子病历的首次病程记录的部分内容

如果考虑在上述电子病历上完成临床医学事件抽取任务，需要先抽取出电子病历中包含的临床医学事件片段，并为抽取出来的临床医学事件片段分配相应的类别标签。图 2-3 展示了该病历上临床医学事件的抽取结果。通过观察和统计，可以发现上述电子病历的首次病程记录部分共包含 53 个临床医学事件，其中问题类型 36 个，检查类型 8 个，治疗类型 1 个，临床科室类型 0 个，特定类型 3 个，证据词类型 5 个，体类型 0 个。

2016-11-11 11:57

首次病程记录

李xx, 女, xx岁, 职业xx, 家住xx省xx县xx乡xx村xx社。主因“**外伤后头痛头晕半天**”。门诊以“**外伤史头痛头晕半天**”收住**入院**。

既往史: 否认**肝炎、结核等传染病史**, 否认**高血压、冠心病等心血管病史**, 否认**手术、外伤、输血史**, 否认**过敏史**, **预防接种史**不详。

现病史: 患者及家属**诉**患者于**入院**前半天被人击打**头部受伤**, 伤后自觉**头痛、头晕较剧**, **全身多处疼痛**不适, 无**意识丧失、呕吐、发热**。

家族史: 父亲健在, 母亲健在。兄弟、姐妹健在。

入院查体: **T**:37.4℃, **P**:82次/分, **R**:18次/分, **BP**:120/70mmHg, **发育正常**, **营养不良**, **体型正常**, **体位自动**, 无**异常面容**, **神志清楚**, **步态未见异常**, **语言清晰流利**, **查体合作**, **问答切题**, **脉搏有力**, **全身正常**, 未见**皮疹、出血点**。

辅助检查: **头颅CT提示**: 未见异常。

初步诊断:

- 脑震荡**
- 全身多处软组织挫伤**

注:

问题 ■ 检查 ■ 治疗 ■ 临床科室 ■ 特定事件 ■ 证据词 ■ 体类型事件 ■

图 2-3 中文电子病历的首次病程记录部分临床医学事件的抽取结果

2.1.2 基于电子病历的临床医学事件时序关系抽取

基于电子病历的临床医学事件时序关系抽取旨在提取电子病历中包含的所有时序关系, 包括临床医学事件与临床医学事件之间的时序关系, 临床医学事件与时间表达式之间的时序关系和时间表达式与时间表达式之间的时序关系。从时序关系覆盖范围的角度划分, 又可以划分为句内时序关系, 句间时序关系以及文档创建时间与该文档范围内临床医学事件之间的时序关系。本文主要研究句内时序关系, 讨论句内时序关系中临床医学事件与临床医学事件之间的时序关系和临床医学事件与时间表达式之间的时序关系。

下面以两个简单的例子来说明基于电子病历的临床医学事件时序关系抽取任务的主要内容。

例 2.1 The patient had undergone **an operation for gastroenteritis (E1)** prior to **May 6 (T1)**.

译文: 患者在 5 月 6 日之前进行过一次肠胃炎手术。

例 2.1 展示的是临床医学事件与时间表达式之间的时序关系。句子中包含一个治疗类型的临床医学事件 an operation for gastroenteritis (E1) 和一个日期类型的时间表达式 May 6 (T1), 根据句子所表达的含义可以推测出两者之间的时序关系类型为 E1-BEFORE-T1。

例 2.2 The patient had completed **a preoperative workup (E1)** prior to the **kidney replacement surgery (E2)**.

译文：患者在换肾手术之前已完成术前检查。

例 2.2 展示的是临床医学事件与临床医学事件之间的时序关系。句子中包含一个检查类型的临床医学事件 a preoperative work up (E1) 和一个治疗类型临床医学事件 kidney replacement surgery (E2)，根据句子所表达的含义可以推测出两者之间的时序关系类型为 E1-BEFORE-E2。

2.2 相关技术介绍

目前，人工智能领域中最热的研究方向当属深度学习。深度学习的迅速发展受到了学术界和工业界的广泛关注，由于其拥有优秀的特征选择和提取能力，对包括机器翻译、目标识别、图像分割等在内的诸多任务产生了越来越重要的影响^[18]。本文将其在应用于基于电子病历的临床医学事件时序信息抽取任务上，涉及的技术包括：Transformer、BERT 和图神经网络。下面将对这些技术进行详细介绍。

2.2.1 Transformer

Google 研究团队在 2017 年 6 月发表了一篇名为“Attention is all you need”的论文^[19]，论文提出了一种基于自注意力机制的新型神经网络模型-Transformer。为证明 Transformer 的有效性，研究团队利用该模型在机器翻译任务上做了实验，实验结果显示 Transformer 在机器翻译任务上的表现远超其他模型，稳居排行榜第一名。虽然研究团队只在机器翻译任务上做了实验，但 Transformer 的提出也为其他自然语言处理任务提供了一种新的研究思路。

众所周知，大多数竞争性神经序列传导模型采用的都是 Encoder-Decoder 的结构^[20]。给定一个输入序列的符号表征结果 $(x_1, x_2, x_3, \dots, x_n)$ ，编码器对符号表征的结果进行编码，得到输入序列的上下文语义表示 $z = (z_1, z_2, z_3, \dots, z_n)$ 。对于生成的 z ，解码器进行解码，每一次生成一个元素，依次生成最终的解码序列 $(y_1, y_2, y_3, \dots, y_m)$ 。在解码过程中，模型的每一步都是自回归的，每一步的输出都依赖之前生成的符号表征^[21]。Transformer 的设计同样遵循 Encoder-Decoder 结构。图 2-4 为 Transformer 的整体框架。下面将对 Transformer 的整体框架和核心结构多头注意力机制进行详细介绍。

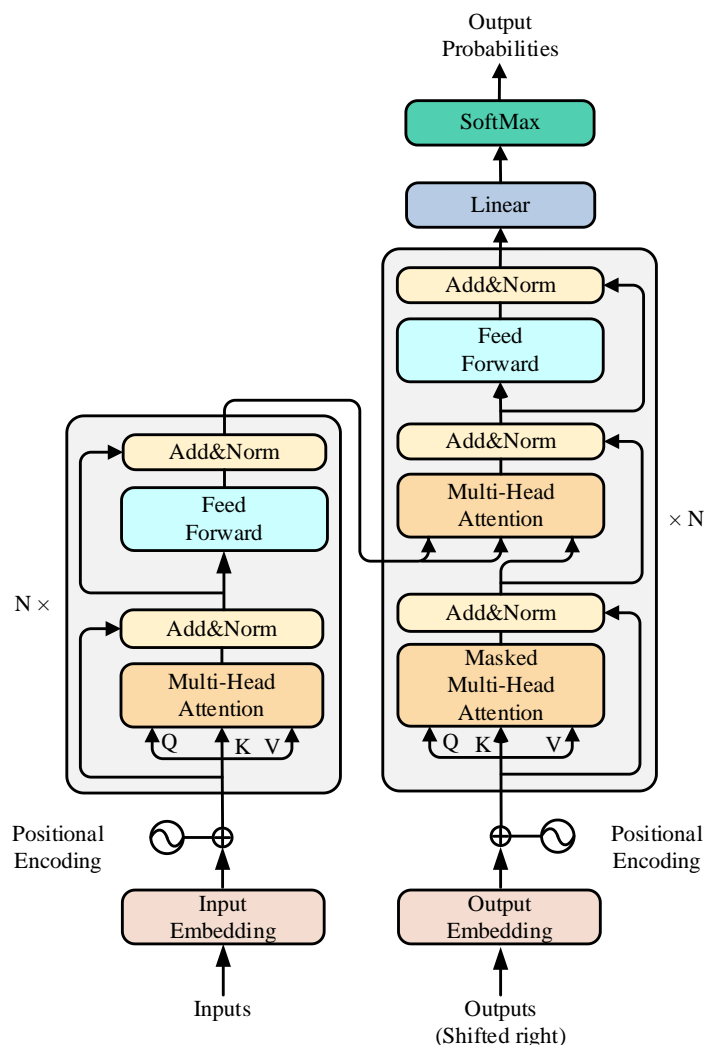


图 2-4 Transformer 的整体框架

2.2.1.1 编码器与解码器

Transformer 的编码器由 6 层完全相同的结构堆叠而成。每一层包含两个子层，第一层是一个多头注意力机制模块，第二层是一个全连接的前馈神经网络结构。每一个子层之后连接一个残差结构^[22]和标准化层^[23]进行归一化处理。第一层编码器的输入来自模型底层的原始输入，其余层编码器的输入来自于当前编码器上一层编码器的输出。

Transformer 的解码器同样由 6 层完全相同的结构堆叠而成。每一层包含三个子层，第一层是一个多头注意力机制模块，第二层是对编码器结构最顶层输出和上一层多头注意力机制模块的输出进行多头注意力机制处理的模块，第三层是一个全连接的前馈神经网络结构。每一个子层之后连接一个残差结构和标准化层进行归一化处理。值得注意的是，解码器的第一层多头注意力机制不同于编码器中第一

层的多头注意力机制和解码器中第二层的多头注意力机制，该注意力机制采用了掩码策略，保证位置 i 的预测只依赖位置 i 之前的已获得的输出结果。

2.2.1.2 多头注意力机制

注意力机制函数可以被理解是将 query 和一组 key - value 对通过一系列数学变换映射到输出的过程。注意力机制函数的输出 output 是通过在 value 上分配权值得到的。其中权值向量由 query 和 key 通过函数运算得到。图 2-5 为注意力机制的整体结构。

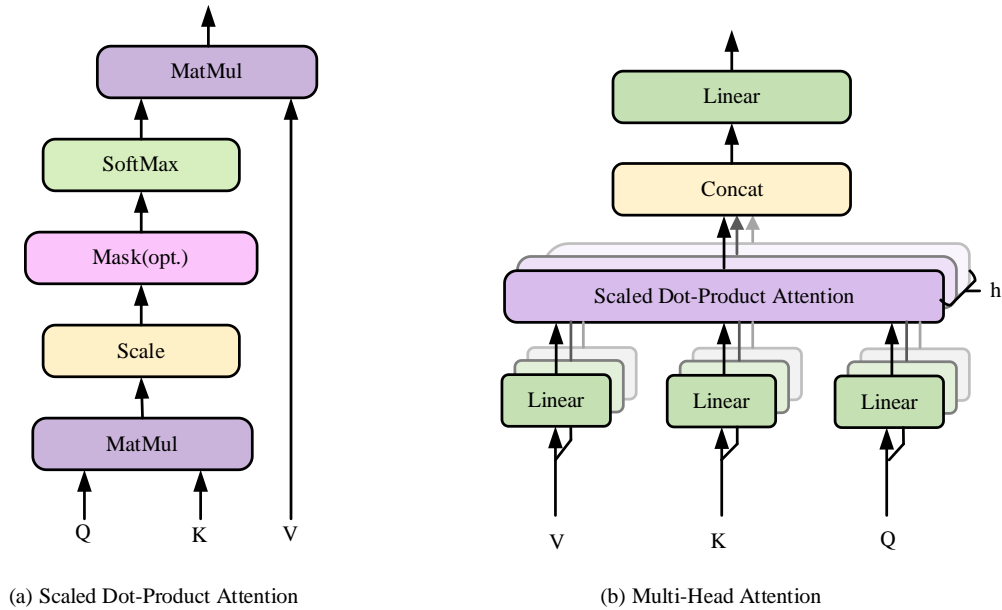


图 2-5 注意力机制的整体结构

图 2-5 左侧展示的是缩放点积注意力机制的整体结构，该结构的输入由 d_k 维的 query 和 key 以及 d_v 的 value 组成。缩放点积注意力机制输出结果的计算分为以下三步：首先计算 query 与所有 key 的点积，然后在第一步的结果上除以 $\sqrt{d_k}$ ，最后利用 softmax 函数处理第二步的结果，处理后的结果作用在 value 上。在实际运算过程中，通常是多个 query 构成一组 query 同时完成注意力机制输出结果的计算，即将多个 query 打包成 Q 矩阵，key 和 value 也打包成对应的 K 和 V 矩阵参与运算，计算公式如式(2-1)：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2-1)$$

图 2-5 右侧展示的是多头注意力机制的整体结构。多头注意力机制的思路是进行多组注意力机制的运算，将所有计算结果综合起来作为多头注意力机制处理

后的输出。为了能够进行多组注意力机制的运算，需要对 Q, K, V 矩阵进行 h 组线性变化，生成 h 组新的 Q, K, V 矩阵，对新的 h 组 Q, K, V 矩阵并行地执行注意力机制的运算，产生 h 组 d_v 维的输出，然后将其拼接起来得到多头注意力机制处理后的最终输出，计算公式如式(2-2)和(2-3)：

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O \quad (2-2)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2-3)$$

其中， $W_i^Q \in R^{d_{\text{model}} \times d_k}, W_i^K \in R^{d_{\text{model}} \times d_k}, W_i^V \in R^{d_{\text{model}} \times d_v}, W^O \in R^{hd_v \times d_{\text{model}}}$ 。

2.2.1.3 关于 Transformer 的思考

在 Transformer 出现之前，循环神经网络^[25]，特别是其变体长短期记忆网络(Long Short Term Memory, LSTM)^[24]、门控循环单元(Gate Recurrent Unit, GRU)和卷积神经网络被认为是特征提取上的主流方法。Transformer 出现后，这一局面发生了改变。RNN 模型采用的是递归式框架结构，当前位置的输入由两部分构成：前一时刻的输出和当前位置的初始输入，这导致模型无法并行操作。此外 RNN 模型也无法很好地学习全局结构信息，主要原因是 RNN 模型本质是马尔科夫决策过程。CNN 模型虽然可以并行操作，不存在运行时间缓慢的问题，但却只能获取局部信息，而每一次信息获取范围的扩大只能通过改变卷积核的大小实现。Transformer 很好的解决了上述问题。Transformer 优良的性能得益于多头注意力机制的使用，注意力机制的使用促使模型在编码每一个位置的信息时会考虑句子中其他位置的信息，这使得句子中每一个位置的信息编码结果都融合了句子的上下文语义信息，解决了 CNN 只能捕获局部语义信息的问题。而多头注意力机制的应用使得编码的信息更加丰富。此外，在对句子的每一个位置进行编码时，不需要等待之前的结果，所有位置可以并行编码，这解决了 RNN 模型无法并行操作的问题。

2.2.2 BERT

Google 研究团队在 2018 年 10 月发表了一篇名为“BERT: Pre-training of Deep Bidirectional Transformer for Language Understanding”^[26]的论文，论文提出了一种基于 Transformer 结构的深度双向预训练语言模型。为证明 BERT 模型的有效性，研究团队利用该模型在 11 项自然语言处理任务上做了实验，实验结果显示 BERT 在 11 项自然语言处理任务上都刷新了以往的记录。

BERT 模型的本质是预训练语言模型。预训练语言模型应用于下游任务能有效提升模型性能这一观点在多年之前就被证实。截止目前，已有多项自然语言处理

任务采用融合预训练语言模型的思路进行研究^[27-29]，包括句子级任务自然语言推理^[30-31]和词级任务命名实体识别^[32]。其中，自然语言推理任务通过理解句子语义完成对句子之间关系的预测，命名实体识别任务同样也需要理解句子语义并完成对句子中实体类别的判断。预训练语言模型能够为上述任务提供合理的语境信息，辅助模型进行句子语义理解。

目前，将预训练语言模型应用于下游任务有两种策略：基于特征的方法和基于微调的方法。基于特征的方法，以 ELMO(Embeddings from Language Models)^[33]为例，该模型使用预训练语言模型的输出作为额外特征应用于特定任务。基于微调的方法，以 GPT(Generative Pre-trained Transformer)^[29]为例，该模型引入最小的任务特定参数，并通过简单地微调预训练参数对下游任务进行训练。这两种方法在预训练过程中共享相同的目标函数，且都使用单向语言模型来学习语言表示。然而，这两种策略都具有一定的局限性，特别是微调的方法，局限性来源于标准语言模型是单向的，这限制了预训练语言模型架构的选择。例如 OpenAI GPT 采用从左到右的单向 Transformer 构建预训练语言模型，导致 Transformer 中自注意力机制层的每一个词只会对之前位置的词产生影响，无法捕捉完整的全局信息。这种局限性对于句子级任务不是很大，而对于词级任务来说，影响巨大，甚至是毁灭性的影响。

Google 研究团队提出的 BERT 模型，就是在基于微调的方法上做了改进。受 Cloze 任务的启发^[34]，BERT 模型使用“(MLM) masked Language Model”作为预训练的一个目标，减轻语言模型建模的单向局限性束缚。MLM 从输入中随机选取一些词汇进行遮盖，仅仅依靠遮盖后的语境完成对原始词汇 ID 的推测。与之前单向的预训练语言模型不同，该掩码遮盖策略允许词汇表征融合左右两侧的语境信息，这使得模型的建模方式从原来的单向建模转换为双向建模。此外，BERT 还引入了“Next Sentence Prediction”任务来联合训练语言模型。“Next Sentence Prediction”任务旨在预测给定句子的下一个句子。该任务的设定可以训练 BERT 理解句子之间关系的能力。在上述两种任务的帮助下，BERT 在大量句子级和词级任务上都表现出了很好的性能。

2.2.2.1 BERT 的基本框架

图 2-6 为 BERT 的基本框架，BERT 由多个双向 Transformer 编码器堆叠而成。每一个 Transformer 编码器包含两个子层，第一层是多头注意力机制模块，第二层是全连接的前馈神经网络结构。每一个子层之后连接残差结构和标准化层进行归一化处理。第一层 Transformer 编码器的输入来自模型底层的原始输入，其余层 Transformer 编码器的输入来自上一层编码器的输出。

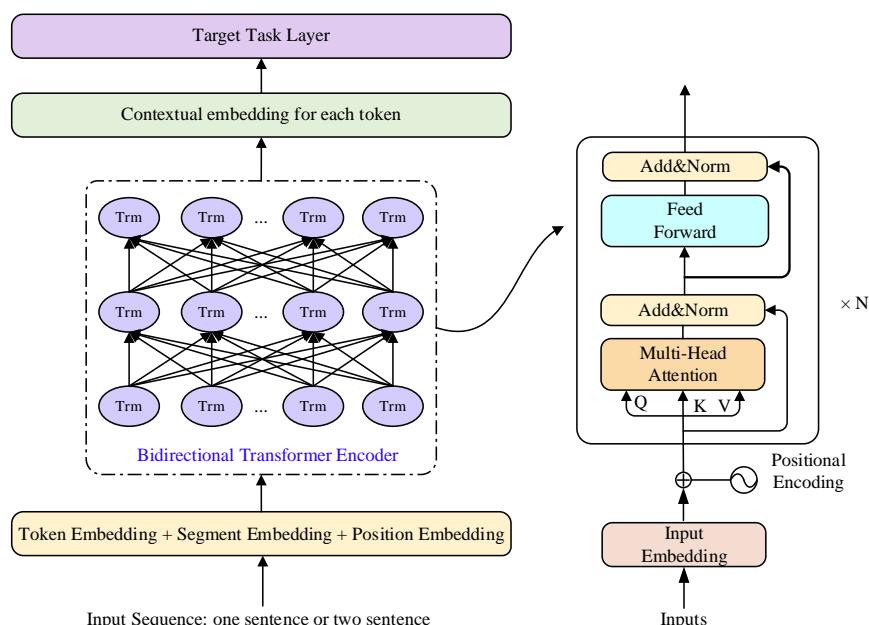


图 2-6 BERT 的整体框架

2.2.2.2 BERT 的输入和输出

BERT 为单句任务和多句任务设定了不同的输入和输出格式。图 2-7 展示了 BERT 的输入和输出格式。考虑到本文研究的任务都是单句任务，这里只对单句任务的输入和输出格式做简单地介绍。BERT 的输入包含三部分：词嵌入、句子分割嵌入和位置嵌入。其中，词嵌入是对当前位置符号的表征，句子分割嵌入表示当前位置符号属于哪一个句子，位置嵌入表示当前位置符号在句子中的位置信息。本文的两个任务都属于单句任务，句子分割嵌入可设置为全 1 的嵌入。本文的两个任务在输出格式上有一点的区别，临床医学事件抽取任务需要获得除句子切割符 [CLS] 和 [SEP] 之外其他位置的所有向量表征用于临床医学事件抽取。时序关系抽取任务只使用句子结束符 [SEP] 位置的向量表征作为最终分类结果的依据。

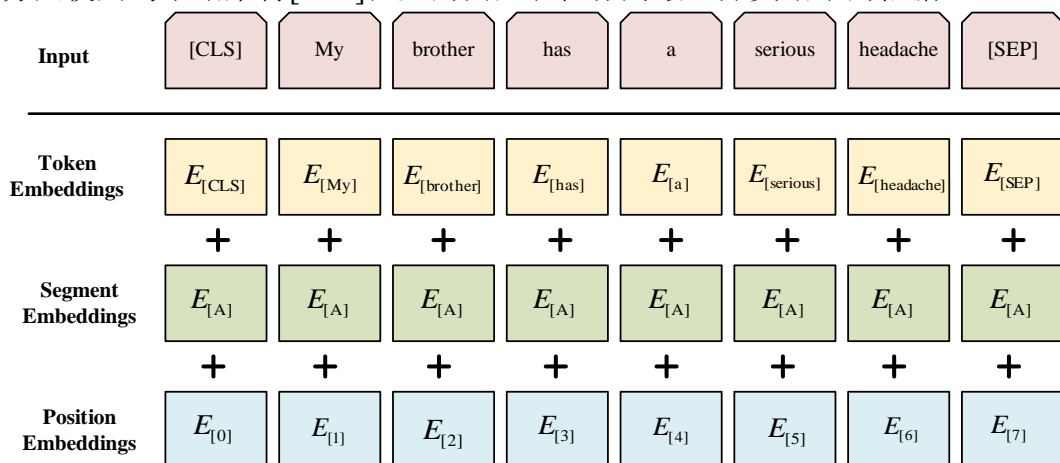


图 2-7 BERT 的输入和输出格式

2.2.3 图神经网络

在过去的几年里，深度学习方法凭借其在欧式空间数据上优秀的特征提取能力，在自然语言处理领域的众多任务上大放异彩，推动着自然语言处理领域众多任务的研究迈上新的台阶。然而实际应用场景中的许多数据是从非欧式空间生成的，传统的深度学习方法虽然也能处理这类数据，但实际的表现却难以使人满意^[35]。以保险和金融风险评估为例，一个完备的风险评估系统不仅需要基于个人的履历、行为习惯、健康程度等进行分析处理，还需通过其亲友、同事、同学之间的往来数据和相互评价进行进一步信用的评估和判断。基于图类型数据的学习系统能够利用用户之间、用户与商品之间的交互，做出非常准确的因果和关联测，但基于图类型数据的特征提取并不是传统深度学习方法所擅长的。推其原因，则是因为现有的深度学习方法认为数据样本之间是彼此独立的，无依赖关系。然而，对于图类型数据来说，情况并非如此，每一个图都是由若干无序且大小可变的节点构成，每一个节点又有若干个相邻节点围绕，节点之间存在一定的依赖关系，这就使得传统的深度学习方法无法很好地捕获图类型数据中包含的依赖关系。鉴于此背景下，研究人员开始探索如何将深度学习方法在图类型数据上进行扩展，受卷积神经网络和循环神经网络设计思路的启发，研究人员定义和设计了用于处理图类型数据的神经网络结构—图神经网络。

图神经网络处理的数据类型为图类型数据，图类型数据包含两类特征：节点特征和边特征，图神经网络的本质就是提取图类型数据中包含的节点特征和边特征。下面将结合图 2-8，对图神经网络处理数据的过程进行简单介绍。给定一个包含 9 个节点的图，图神经网络处理数据的第一步就是输入的初始化，即初始化节点和边的表征，这里使用维度为 d_v 的特征向量对 9 个节点进行表征，并集成起来构成矩阵 \mathbf{V} 。此外，图类型数据还包含了边特征，边特征反映的是节点之间的关联度，这样用一个 9×9 的矩阵 \mathbf{E} 表示图类型数据中节点的初始关联度。在获得了节点特征 \mathbf{V} 和边特征 \mathbf{E} 之后，下一步就是利用边特征对节点特征进行更新，这一步是图神经网络处理数据的关键步骤。计算过程如式(2-4)和(2-5)所示：

$$\mathbf{H}_i = f(\mathbf{H}_{i-1}, \mathbf{E}) \quad (2-4)$$

$$f(\mathbf{H}_i, \mathbf{E}) = \text{ReLu}(\mathbf{E}\mathbf{H}_i\mathbf{W}_i) \quad (2-5)$$

其中， $\mathbf{H}_0 = \mathbf{V}$ 。不同网络的区别体现在映射 f 的设计不同，式(2-5)中映射 f 为 ReLu 函数。为了能够从多角度对图类型数据进行特征提取，需要对包含节点特征和边特征的原始图表征进行多组线性映射形成多个图，图中的边特征采用注意力机制计算获得，整个计算过程和多头注意力机制的计算过程完全一致。

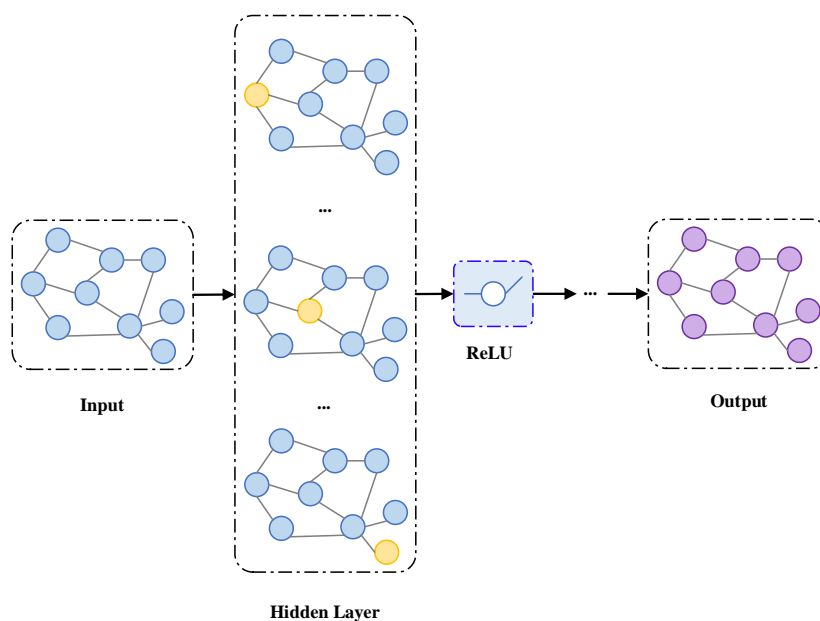


图 2-8 图神经网络处理数据的过程

2.3 系统性能评估

本小节将介绍基于电子病历的临床医学事件抽取和基于电子病历的临床医学事件时序关系抽取任务的评价指标。

2.3.1 评价指标

本文主要研究两个任务，分别为基于电子病历的临床医学事件抽取和基于电子病历的临床医学事件时序关系抽取。上述两个任务都属于分类任务，只是分类的粒度有差异，前者属于字粒度的分类任务，后者属于句子粒度的分类任务。自然语言处理领域中分类任务常用的评价指标：正确率(Accuracy)、准确率(Precision)、召回率(Recall)和 F1 值(F1-score)进行模型性能评估。表 2-1 为分类结果的混淆矩阵(Confusion matrix)，其中真正例 TP 表示真实结果为正例且预测结果也为正例的样本个数，假正例 FP 表示真实结果为负例且预测结果为正例的样本个数，假负例 FN 表示真实结果为正例且预测结果为负例的样本个数，真负例 TN 表示真实结果为负例且预测结果为负例的样本个数。

表 2-1 分类结果的混淆矩阵

| 真实情况 | 预测为正例 | 预测为负例 |
|------|---------|---------|
| 正例 | 真正例(TP) | 假负例(FN) |
| 负例 | 假正例(FP) | 真负例(TN) |

正确率是指测试集中被正确分类的样本在测试集所有样本中所占的比例，计算公式如式(2-6)：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (2-6)$$

正确率的计算公式简单且易于理解，但是正确率不能作为模型性能评价的唯一指标。在数据分布不均衡的情况下，正确率无法正确反映模型的真实性能。假设负例样本在总样本占比极小，当负例样本全部分类错误时，正确率计算结果却很高，这与模型对于负例样本分类的真实性能是不符的。因此，本文采用准确率、召回率和F1值进行模型性能评估。

准确率是指测试集中识别正确的正例样本占预测结果为正例样本的比例，它反映了系统的准确率，计算公式如式(2-7)：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2-7)$$

召回率是指测试集中识别正确的正例样本占有所有正例样本的比例，它反映了系统的完备性，计算公式如式(2-8)：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2-8)$$

F_β 是对召回率和准确率的综合考虑，计算公式如式(2-9)：

$$F_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \quad (2-9)$$

其中， β 是准确率和召回率的相对权重，用于衡量召回率对准确率的相对重要性。 $\beta > 1$ 时，召回率有更大的影响， $1 > \beta > 0$ 时，准确率有更大的影响， $\beta = 1$ ，称为F1值。

2.4 本章小结

本章主要阐述了基于电子病历的临床医学事件时序信息抽取技术研究中涉及到的理论基础和关键技术。首先，介绍了基于电子病历的临床医学事件时序信息抽取技术研究的三个层次，点明了本文主要关注第一层次和第二层次的研究，并介绍了临床医学事件抽取和临床医学事件时序关系抽取任务；然后，介绍了研究过程中使用到的深度学习技术；最后，介绍了模型性能的评价指标。

第3章 基于电子病历的临床医学事件时序信息语料库构建

基于电子病历的临床医学事件时序信息语料库的构建是基于电子病历的临床医学事件时序信息抽取的上游任务，目的是为基于电子病历的临床医学事件时序信息抽取提供底层的数据支持。目前用于临床医学事件时序信息抽取任务的语料库数量较少，且都未公开，主要原因是电子病历不同于一般的叙事文本，属于患者的隐私信息，不易公开。鉴于此背景下，本章考虑自主构建基于电子病历的临床医学事件时序信息语料库，具体工作包括基于电子病历的临床医学事件信息语料库构建和基于电子病历的临床医学事件时序关系语料库重构。

3.1 基于电子病历的临床医学事件信息语料库构建

本小节主要介绍基于电子病历的临床医学事件信息语料库构建的整体过程，并对构建过程中的核心步骤进行详细介绍，内容包括标注语料的选取、标注规范的制定、标注示例的说明、标注结果的规格化、语料库标注的阐述和语料库数据的统计与分析。

3.1.1 语料选取

本小节使用的中文电子病历来源于甘肃省某三甲医院，涵盖心血管内科、神经外科、儿科、妇产科、传染科等 15 个科室，共计 4000 份。考虑到不同科室电子病历特点和医生撰写病历的结构不同，我们从每个科室随机抽取共 2000 份电子病历用于临床医学事件信息语料库构建。在获得基础语料后，没有立即开展语料库的标注工作，而是先抽取了 15 份电子病历进行了分析。通过分析，发现电子病历中信息存在冗余。经过归纳和整理，发现电子病历的首次病程记录和出院小结部分涵盖了电子病历的大部分内容，且内容相对完整。因此，本小节以电子病历的首次病程记录和出院小结作为语料标注的核心内容。首次病程记录主要包含四部分内容，分别是患者的主诉、入院检查的结果、医生出具的诊断结果和诊疗计划。出院小结主要包含五部分内容，分别为入院记录、诊断记录、治疗记录、出院情况和出院医嘱。

3.1.2 标注规范

在制定临床医学事件信息语料库标注规范时，参考了 2012 I2B2 (Informatics for

Integrating Biology & the Bedside)时序信息语料库标注规范。考虑到 2012 I2B2 时序信息语料库的标注文本为英文电子病历,而本章使用的是中文电子病历,两种不同语言书写的电子病历存在着行文结构上的差异,未直接将 2012 I2B2 时序信息语料库的标注规范应用在中文电子病历上,而是结合中文电子病历的特点,对其进行了改造,形成了适合中文电子病历的临床医学事件信息语料库的标注规范。新的中文电子病历事件信息语料库的标注规范定义了七类临床医学事件类型,为问题类型、检查类型、治疗类型、临床科室类型、证据词类型、特定类型和体类型。下面将进行详细介绍,并举例说明。其中引号中的内容为电子病历片段,蓝色字体部分为临床医学事件。

3.1.2.1 问题

问题类型(Problem)是指电子病历中医生出具的疾病诊断结果或由疾病导致的不适症状或异常表现,包括三类疾病、症状和异常检查结果。

1. 疾病

电子病历中对于疾病的描述分为以下六种情况:

(1) 疾病名称

例如:“**糖尿病**”,“**癫痫**”,“**泌尿系统感染**”,“头 MRI 示:**腔隙性脑梗死**”,“**肺感染**”,“**肿瘤**”,“ECG:**窦性心动过速**”,“查血常规发现**贫血**”等。

(2) 包含修饰词的疾病名称

例如:“**症状性癫痫**”,“**胆汁反流性胃炎**”,“**大脑中动脉栓塞**易导致大面积**脑梗死**”,“**再生障碍性贫血**”,“**慢性胆囊炎**”等。

(3) 病毒和细菌

例如:“**结核分枝杆菌**”,“**巨细胞病毒**”,“**葡萄杆菌**”等。

(4) 身体畸形

例如:“**畸形**”。

(5) 身体受到的外部侵袭

例如:“**骨折**”,“**腋下胸部及右侧臀部多处擦皮伤**”等。

(6) 病史

例如:“无**外伤史**”,“**家族遗传病**”,“否认**药物及食物过敏史**”等。

2. 症状

电子病历中对于症状的描述分为以下五种情况:

(1) 症状

例如:“**发热**”,“**恶心**”,“**咳嗽**”,“**呕吐**”,“**寒战**”,“**双眼黑蒙**”,“**大小便失**

禁”、“呼吸困难”、“焦虑”、“胳膊脱臼”、“新生儿缺陷”和“妊娠”等。

(2) 包含部位信息的症状

例如：“右侧口角流涎”、“阵发性枕部麻木”、“左上肢及双下肢活动不灵”、“双侧大脑前中后动脉迂曲”、“体重无明显减轻”、“双肺未闻及明显啰音”和“口唇无发绀”等。

(3) 包含修饰词的症状

例如：“发作性头晕”、“压榨样闷痛”、“头痛呈发作性钝痛，具体位置不定，持续时间为十余分钟”，“右侧肢体麻木加重”。

(4) 患者自述的症状

例如：“左侧肢体无力”、“耳鸣”、“视物旋转”、“神志清楚”、“鼻通气良好”、“上肢可抬举”、“无肢体活动障碍及抽搐发作，疼痛发作”等。

(5) 异常体征

例如：“右耳廓可及长约1.0cm创口，创缘尚齐”、“于当地医院化验血糖高，给予口服降糖药物治疗”、“全麻未醒”，“3cm可触及肿物”、“双下肢足背动脉仍未闻及明显搏动”、“头面部及双手背皮疹呈轻度苔藓化部分结痂”、“发热者给予补液营养支持”、“双眼角膜透明，前房常深，瞳孔圆，D≈5.0mm，呈药物性散大”等。

3. 异常检查结果

例如：“脑实质内高密度灶”、“右侧脑室体旁片状低密度病灶”和“双肺听诊可闻及少量痰鸣音”等。

3.1.2.2 检查

检查类型(Exam)是指为帮助医生更精确地诊断病人，所施加给病人的检查过程，包括四类：检查手段、检查程序和检查设备，以测量值和指标值量化的检查项目，以“症”字和“征”字结尾的检查项目，以及包含关键词的检查过程。

1. 检查手段、检查程序和检查设备

例如：“胸X光”、“心电图”、“头CT显示脑实质内高密灶”、“密切监测血常规”、“就诊于我院经骨穿等相关检查”和“心肺听诊”等。

2. 以测量值和指标值量化的检查项目

例如：“CRE”、“WBC”、“CBC”、“血培养”、“生化系列”、“左侧上肢近端肌力3级”、“血压”、“脉搏”、“肠鸣音正常”、“查体合作，问答切题，脉搏有利”，“全身正常”和“肌张力无增高或减低”等。

3. 以“症”字或“征”字结尾的检查项目

例如：“Babinski's 症”、“墨菲氏症”和“Murphy 氏征”等。

4. 包含关键词的检查过程

例如：“前庭功能试验”、“甲状腺功能测定可辅助诊断”、“心肺听诊无著征”、“活检”、“妇科检查”、“盆腔检查”和“定期产检”等。

3.1.2.3 治疗

治疗类型(Treatment)是指电子病历中记录的施加给病人的治疗程序、干预措施和医疗手段等，包括三类治疗过程、药物名称和其他治疗信息。

1. 治疗过程

例如：“化痰”、“输血”、“抗凝”、“扩冠治疗”、“TC 化疗”、“支具固定”、“鼻中隔偏曲矫正手术”、“术后补液”和“按计划免疫接种各种疫苗”等

2. 药物名称

例如：“奥扎格雪、脑蛋白水解物等静点”，“注射胰岛素控制血糖、20mg 布洛芬、胰岛素强化治疗、应用丹红、苏安等药物”等。

3. 其他治疗信息

例如：“调整用药”、“适当功能锻炼”、“穿弹力袜”和“支持对症”等。

3.1.2.4 临床科室

临床科室类型(Clinical department)是指电子病历中记录的患者在住院治疗期间所接触的科室。例如：“未给予任何治疗，现为求进一步诊治，遂来我院就诊，门诊以新生儿脐炎收住儿科”。

3.1.2.5 证据词

证据词类型(Evidence)是指电子病历中描述信息来源的字或词，包含主诉、报告、显示、揭示、示、呈、诊断、见、闻及、查及、引出、扪及等。例如：“患者精神状况较前略好转，诉咳嗽及气短症状略缓解”、“双肺叩诊呈过清音”、“胸部 CT 回报”、“行头颅 CT 检查提示”和“心脏彩超示符合高血压性心脏病”。

3.1.2.6 特定类型

特定类型(Occurrence)是指在除上述类型之外的发生在患者住院治疗期间的事件，如入院、转院、出院、平躺、站立和击打等。例如：“主因外伤后头痛头晕半天”，门诊以“外伤史头痛头晕半天收住入院”。

3.1.2.7 体类型

体类型(Aspectual event)用于描述当前事件的后续状态，如：开始、启用、重现、

恢复、继续、巩固、完成、停止、结束、停用、解除、重启、好转、减轻、缓解、改善、加重、退出、保留等。例如：“今日**停止**止血药物，加用活血药物，余**继续**原方案治疗，观察患者病情变化”。

3.1.3 标注示例

本小节将依据临床医学事件信息语料库的标注规范，选取一篇电子病历的首次病程记录部分进行标注。图 3-1 是电子病历的首次病程记录部分。

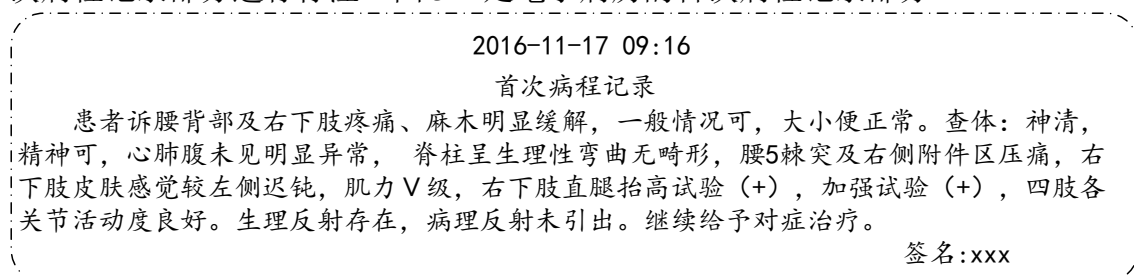


图 3-1 电子病历的首次病程记录部分

图 3-2 为上述电子病历的首次病程记录部分临床医学事件信息的标注结果。在进行临床医学事件信息语料库标注时，采用七种不同色块表示七种不同类型的临床医学事件，红色表示问题类型，绿色表示检查类型，橙色表示治疗类型，粉色表示临床科室类型，黄色表示证据词类型，蓝色表示特定类型事件，紫色表示体类型事件。经过观察和统计，上述电子病历共包含 24 个临床医学事件，其中问题类型 11 个，检查类型 7 个，治疗类型 1 个，临床科室类型 0 个，证据词类型 3 个，特定类型 0 个和体类型 2 个。



图 3-2 首次病程记录部分临床医学事件信息的标注结果

3.1.4 标注结果规格化

在进行临床医学事件信息语料库的标注时，采用七种不同色块表示七种不同类型的临床医学事件，这种方法形成的标注结果虽然看起来清晰明了且区分度高，但不利于后期在该语料库上开展的临床医学事件信息抽取任务的研究，鉴于此原

因,考虑自主设计一款标注工具用于标注结果的规格化输出。图 3-3 为标注结果的规格化输出,输出的文档类型为 xml 文档。

```
<Record>
  <Content Value="2016-11-17 09:16" 首次病程记录
    患者诉腰部及右下肢疼痛、麻木明显缓解,一般情况可,大小便正常。查体:神清,精神可,心肺腹未见明显异常,
    脊柱呈生理性弯曲无畸形,腰5棘突及右侧附件区压痛,右下肢皮肤感觉较左侧迟钝,肌力V级,右下肢直腿抬高试验(+),加强试验(
    +),四肢各关节活动度良好。生理反射存在,病理反射未引出。继续给予对症治疗。 签名:xxx " />
    <Event ID="1" StartLocation="31" EndLocation="31" EventText="诉" EventType="证据词" />
    <Event ID="2" StartLocation="32" EndLocation="40" EventText="腰部及右下肢疼痛" EventType="问题" />
    <Event ID="3" StartLocation="42" EndLocation="43" EventText="麻木" EventType="问题" />
    <Event ID="4" StartLocation="46" EndLocation="47" EventText="缓解" EventType="体类型事件" />
    <Event ID="5" StartLocation="49" EndLocation="53" EventText="一般情况可" EventType="问题" />
    <Event ID="6" StartLocation="55" EndLocation="59" EventText="大小便正常" EventType="问题" />
    <Event ID="7" StartLocation="61" EndLocation="62" EventText="查体" EventType="检查" />
    <Event ID="8" StartLocation="64" EndLocation="65" EventText="神清" EventType="问题" />
    <Event ID="9" StartLocation="67" EndLocation="69" EventText="精神可" EventType="问题" />
    <Event ID="10" StartLocation="71" EndLocation="73" EventText="心肺腹" EventType="检查" />
    <Event ID="11" StartLocation="75" EndLocation="75" EventText="见" EventType="证据词" />
    <Event ID="12" StartLocation="76" EndLocation="79" EventText="明显异常" EventType="问题" />
    <Event ID="13" StartLocation="82" EndLocation="92" EventText="脊柱呈生理性弯曲无畸形" EventType="问题" />
    <Event ID="14" StartLocation="94" EndLocation="105" EventText="腰5棘突及右侧附件区压痛" EventType="问题" />
    <Event ID="15" StartLocation="107" EndLocation="118" EventText="右下肢皮肤感觉较左侧迟钝" EventType="问题" />
    <Event ID="16" StartLocation="120" EndLocation="121" EventText="肌力" EventType="检查" />
    <Event ID="17" StartLocation="125" EndLocation="133" EventText="右下肢直腿抬高试验" EventType="检查" />
    <Event ID="18" StartLocation="138" EndLocation="141" EventText="加强试验" EventType="检查" />
    <Event ID="19" StartLocation="146" EndLocation="155" EventText="四肢各关节活动度良好" EventType="问题" />
    <Event ID="20" StartLocation="157" EndLocation="160" EventText="生理反射" EventType="检查" />
    <Event ID="21" StartLocation="164" EndLocation="167" EventText="病理反射" EventType="检查" />
    <Event ID="22" StartLocation="169" EndLocation="170" EventText="引出" EventType="证据词" />
    <Event ID="23" StartLocation="172" EndLocation="173" EventText="继续" EventType="体类型事件" />
    <Event ID="24" StartLocation="176" EndLocation="179" EventText="对症治疗" EventType="治疗" />
  </Content>
</Record>
```

图 3-3 首次病程记录部分临床医学事件信息标注结果的规格化输出

该文档包含两部分内容,Content 部分记录的是原始电子病历的内容,Event 部分记录的是电子病历中包含的临床医学事件信息。每一行为一个临床医学事件,记录的内容包括临床医学事件的编号(ID),临床医学事件在文档中的起始位置(StartLocation)和结束位置(EndLocation),临床医学事件所对应的电子病历片段(EventText)和临床医学事件类型(EventType)。其中临床医学事件在文档中的起始位置从 0 开始计算。以第一行为例,临床医学事件的编号为 1,临床医学事件在文档中的起始位置为第 31 个字符位置,结束位置为 31 个字符位置,临床医学事件所对应的电子病历片段为“诉”,临床医学事件类型为证据词类型。

3.1.5 语料库标注

在完成了上述的基础工作后,开始正式进入语料库的标注阶段。语料库的标注工作分为两个阶段预标注阶段和正式标注阶段。下面将对两个阶段的工作进行详细说明。预标注阶段需要完成两项任务标注规范的更新和标注人员的培训。在预标注阶段,将给每一个标注人员发放 56 份相同的电子病历,分四个阶段进行标注,每一个阶段标注人员标注相同的电子病历。每完成一个阶段的标注后,所有的标注人员一起校正标注结果。当遇到分歧时,所有标注人员对标注结果进行讨论,确定统一的标注形式,并对标注规范进行更新。预标注阶段结束后,进入正式的标注阶段,将给每一位标注人员分配 216 份电子病历进行标注。在标注过程中遇到不确定的标注项时,标注人员记录当前标注项。当所有语料标注结束后,汇总不确定的

标注项,所有标注人员进行讨论,确定最后的标注形式。至此,正式标注阶段结束。

3.1.6 语料库数据统计与分析

本小节将对基于电子病历的临床医学事件信息语料库的标注结果进行统计和分析。本次标注工作共标注电子病历 2000 份。表 3-1 为基于电子病历的临床医学事件信息语料库的数据分布。通过观察表 3-1,发现基于电子病历的临床医学事件信息语料库中非临床医学事件类型占比最大,高达 86.43%,临床医学事件占比 13.57%。其中,问题类型、检查类型、特定类型事件和体类型事件所占比例较多,临床科室类型所占比例最少。

表 3-1 基于电子病历的临床医学事件信息语料库的数据分布

| 类型 | 数量 | 占比 |
|---------------------------|--------|--------|
| 问题(Problem) | 66486 | 7.25% |
| 检查(Exam) | 26871 | 2.93% |
| 治疗(Treatment) | 13554 | 1.48% |
| 临床科室(Clinical department) | 167 | 0.02% |
| 证据词(Evidence) | 9426 | 1.03% |
| 特定类型(Occurrence) | 4440 | 0.48% |
| 体类型(Aspectual event) | 3489 | 0.38% |
| 其他类型(Other) | 792759 | 86.43% |

3.2 基于电子病历的临床医学事件时序关系语料库重构

本小节主要介绍基于电子病历的临床医学事件时序关系语料库重构的整体过程,并对语料库重构过程中的核心步骤进行详细介绍,内容包括语料的选取、时序关系分类体系的确立、语料库重构的阐述和语料库数据的统计与分析。

3.2.1 语料选取

本小节使用 2012 I2B2 时序关系挑战赛的评测语料来完成临床医学事件时序关系语料库的重构。2012 I2B2 时序关系挑战赛评测语料的原始数据由与 I2B2 合作的医疗机构和美国哈佛医学院教学医院贝斯以色列女执事医疗中心提供。原始数据由 310 份去隐私的出院小结构成,记录了患者的临床病史和住院治疗过程。2012 I2B2 时序关系挑战赛的评测语料标注了三类时序信息,分别是时间表达式、

临床医学事件和时序关系。对于时序关系部分，评测语料标注了三个层次的时序关系，分别是文档创建时间与该文档范围内临床医学事件之间的时序关系、句内时序关系和跨句子的时序关系。本文主要研究句内时序关系抽取，因此选择 2012 I2B2 时序关系挑战赛评测语料中句内时序关系部分进行重构。

3.2.2 时序关系分类体系

2012 I2B2 时序关系挑战赛的评测语料采用了两种时序关系分类体系，分别是粗粒度的时序关系分类体系和细粒度的时序关系分类体系。粗粒度的时序关系分类体系由三类时序关系构成，分别是 BEFORE、AFTER 和 OVERLAP。细粒度的时序关系体系是对粗粒度的时序关系体系的细化，包含八类时序关系，分别是 BEFORE、AFTER、OVERLAP、SIMULTANEOUS、BEGUN_BY、ENDED_BY、DURING 和 BEFORE_OVERLAP。粗粒度的时序关系分类体系是在细粒度的时序关系分类体系上进行了合并，即 BEFORE、BEFORE_OVERLAP 和 ENDED_BY 合并为 BEFORE，AFTER 和 BEGUN_BY 合并为 AFTER，OVERLAP、SIMULTANEOUS 和 DURING 合并为 OVERLAP。本小节在重构临床医学事件时序关系语料库时选择 2012 I2B2 时序关系挑战赛的评测语料中细粒度的时序关系分类体系。下面将结合例子对 2012 I2B2 时序关系挑战赛的评测语料中细粒度的时序关系分类体系进行介绍。

1. BEFORE

BEFORE 类型表示临床医学事件 E_1 发生在临床医学事件 E_2 之前或者临床医学事件 E_1 发生在时间表达式 T_1 之前或者时间表达式 T_1 发生在临床医学事件 E_1 之前。

例 3.1 The patient was given **stress dose steroids (E1)** prior to **his surgery (E2)**.

译文：病人手术前注射了压力剂量类固醇。

例 3.1 中包含两个 TREATMENT 类型的临床医学事件 stress dose steroids (E1) 和 his surgery (E2)，两者之间的时序关系类型为 E1-BEFORE-E2。

2. AFTER

AFTER 类型表示临床医学事件 E_1 发生在临床医学事件 E_2 之后或者临床医学事件 E_1 发生在时间表达式 T_1 之后或者时间表达式 T_1 发生在临床医学事件 E_1 之后，与 BEFORE 类型是相对应的。

例 3.2 He had **another serious concussion (E1)** prior to **August 17 (T1)**.

译文：他在 8 月 17 日之前又发生了一次严重的脑震荡。

例 3.2 中包含一个 PROBLEM 类型的临床医学事件 another serious concussion (E1) 和一个 DATE 类型的时间表达式 August 17 (T1)，两者之间的时序关系类型为

T1-AFTER-E1。

3. OVERLAP

OVERLAP 类型表示临床医学事件 E_1 和临床医学事件 E_2 大致发生在同一个时刻或者时间表达式 T_1 和临床医学事件 E_1 大致发生在同一个时刻，两者存在时间线的部分重合。

例 3.3 He recently developed **fatigue (E1)** and **dyspnea (E2)**.

译文：他最近感到疲劳和呼吸困难。

例 3.3 中包含两个 PROBLEM 类型的临床医学事件 fatigue (E1) 和 dyspnea (E2)，两者之间的时序关系类型为 E1-OVERLAP-E2。

4. SIMULTANEOUS

SIMULTANEOUS 类型表示临床医学事件 E_1 和临床医学事件 E_2 发生在同一个时刻或者时间表达式 T_1 和临床医学事件 E_1 发生在同一个时刻，两者存在时间线的完全重合。

例 3.4 The patient was taken to the **operating room (E1)** and underwent **an orthotopic liver transplant (E2)**.

译文：病人被送到手术室并接受了原位肝移植手术。

例 3.4 中包含一个 CLINICAL_DEPT 类型的临床医学事件 operating room (E1) 和一个 TREATMENT 类型的临床医学事件 an orthotopic liver transplant (E2)，两者之间的时序关系类型为 E1-SIMULTANEOUS-E2。

5. BEGUN_BY

BEGUN_BY 类型表示临床医学事件 E_1 发生在临床医学事件 E_2 开始时或者临床医学事件 E_1 在时间表达式 T_1 时开始。

例 3.5 On **postoperative day #1 (T1)**, he was started on **Percocet (E1)** two tabs po.q4h.

译文：术后第 1 天，他开始服用扑热啮两片。

例 3.5 中包含一个 TREATMENT 类型的临床医学事件 Percocet (E1) 和一个 DATE 类型的时间表达式 postoperative day #1 (T1)，两者之间的时序关系类型为 E1-BEGUN_BY-T1。

6. ENDED_BY

ENDED_BY 类型表示临床医学事件 E_1 发生在临床医学事件 E_2 结束时或者临床医学事件 E_1 在时间表达式 T_1 时结束。

例 3.6 His **nasogastric tube (E1)** was discontinued on **05-26-98 (T1)**.

译文：他的鼻胃管于 05-26-98 被停用。

例 3.6 中包含一个 TREATMENT 类型的临床医学事件 nasogastric tube (E1) 和

一个 DATE 类型的时间表达式 05-26-98 (T1), 两者之间的时序关系类型为 E1-ENDED_BY-E1。

7. DURING

DURING 类型表示临床医学事件 E_1 发生的时间跨度完全包含临床医学事件 E_2 发生的时间跨度或者临床医学事件 E_1 发生在时间表达式 T_1 的时间跨度中。

例 3.7 He remained in the **Intensive Care Unit (E1)** for **two days (T1)**.

译文: 他在加护病房呆了两天。

例 3.7 中包含一个 CLINICAL_DEPT 类型的临床医学事件 Intensive Care Unit (E1) 和一个 DURATION 类型的时间表达式 two days (T1), 两者之间的时序关系类型为 E1-DURING-T1。

8. BEFORE_OVERLAP

BEFORE_OVERLAP 类型表示临床医学事件 E_1 发生在临床医学事件 E_2 之前且在临床医学事件 E_2 发生时一直持续或者临床医学事件 E_1 发生在时间表达式 T_1 之前且一直持续。

例 3.8 The patient had an undocumented **history of possible atrial fibrillation (E1)** prior to **admission (E2)**.

译文: 患者入院前有可能患有房颤的未登记病史。

例 3.8 中包含一个 PROBLEM 类型的临床医学事件 history of possible atrial fibrillation (E1) 和一个 OCCURRENCE 类型的临床医学事件 admission (E2), 两者之间的时序关系类型为 E1-BEFORE_OVERLAP-E2。

3.2.3 语料库重构

临床医学事件时序关系语料库重构工作的第一步是完成语料的拆分, 即从 2012 I2B2 时序关系挑战赛的评测语料中剔除文档创建时间与该文档范围内临床医学事件之间的时序关系和跨句子的时序关系的标注结果, 保留句内时序关系的标注结果。完成之后的下一步就是修订句内时序关系的标注结果。下面将结合两个例子来解释标注结果的修订过程。

例 3.9 The patient was given stress dose steroids prior to his surgery.

译文: 病人手术前注射了压力剂量类固醇。

标注结果: <TLINK ID = "TL0"

Text = "The patient was given stress dose steroids prior to his surgery."

Event1 = "stress dose steroids"

Event1Start = "5"


```

Event1End = "7"
Event1Type = "TREATMENT"
Event2 = "his surgery"
Event2Start = "10"
Event2End = "11"
Event2Type = "TREATMENT"
RelationType = "E1-BEFORE-E2" />

```

例 3.10 He remained in the Intensive Care Unit for two days .

译文：他在加护病房呆了两天。

标注结果：<TLINK ID = "TL1"

```

Text = "He remained in the Intensive Care Unit for two days."
Event1 = "Intensive Care Unit"
Event1Start = "5"
Event1End = "7"
Event1Type = "CLINICAL_DEPT"
Time1 = "two days"
Time1Start = "9"
Time1End = "10"
Time1Type = "DURATION"
RelationType = "E1-DURING-T1" />

```

例 3.9 展示的是临床医学事件与临床医学事件之间的时序关系，例 3.10 展示的是临床医学事件与时间表达式之间的时序关系。两个例子中标注了时序关系的编号(ID)、原始电子病历的片段(Text)、临床医学事件片段(Event)、时间表达式片段(Time)、临床医学事件的开始位置(EventStart)和结束位置(EventEnd)、时间表达式的开始位置(TimeStart)和结束位置(TimeEnd)、临床医学事件类型(EventType)、时间表达式类型(TimeType)和时序关系类型(RelationType)。

3.2.4 语料库数据统计与分析

本小节将对临床医学事件时序关系语料库重构的结果进行统计与分析，本次重构工作共标注 4000 条电子病历片段，表 3-2 为基于电子病历的临床医学事件时序关系语料库的数据分布。通过观察表 3-2，可以发现基于电子病历的临床医学事件时序关系语料库中 BEFORE 类型和 AFTER 类型占比最大，高达 37.80%和 43.10%，DURING 类型和 BEFORE_OVERLAP 类型占比最少，分别为 0.50%和 0.60%。

表 3-2 基于电子病历的临床医学事件时序关系语料库的数据分布情况

| 类型 | 数量 | 占比 |
|----------------|------|--------|
| BEFORE | 1512 | 37.80% |
| AFTER | 1724 | 43.10% |
| OVERLAP | 324 | 8.10% |
| SIMULTANEOUS | 264 | 6.60% |
| DURING | 20 | 0.50% |
| BEFORE_OVERLAP | 24 | 0.60% |
| BEGUN_BY | 57 | 1.42% |
| ENDED_BY | 75 | 1.88% |

3.3 本章小结

本章主要介绍了基于电子病历的临床医学事件时序信息语料库的构建工作，包括临床医学事件信息语料库的构建和临床医学事件时序关系语料库的重构。针对临床医学事件信息语料库的构建工作，介绍了语料的选取、标注规范的制定、标注的示例、标注结果的规格化，并对语料库数据进行了统计和分析。针对临床医学事件时序关系语料库的重构工作，介绍了语料的选取、时序关系分类的体系、语料库重构的过程，并对语料库数据进行了统计和分析。

第4章 融合对抗训练与预训练语言模型的事件抽取

4.1 引言

随着智慧医疗产业的不断发展,构建一个智能的临床决策支持系统变得迫在眉睫。作为临床决策支持系统的重要组成部分,面向中文电子病历的临床医学事件抽取受到了来自学术界和工业界的持续关注。面向中文电子病历的临床医学事件抽取旨在抽取中文电子病历中与患者临床时间线相关所有信息,并为抽取出来的信息分配相应的类别标签。在过去的十几年里,众多方法被应用于去解决面向中文电子病历的临床医学事件抽取任务,例如:隐马尔科夫模型^[6]、支持向量机模型^[7]和条件随机场模型^[36]等。最近,随着深度学习技术的发展,研究人员开始考虑使用神经网络模型去解决面向中文电子病历的临床医学事件抽取任务^[10,37-39]。

虽然这些方法在面向中文电子病历的临床医学事件抽取任务上取得了不错的成绩,但仍有一些问题没有被很好的解决。其中最显著的问题是没有公开的、标注好的用于面向中文电子病历的临床医学事件抽取任务的语料库。研究人员需要自主构建一个面向中文电子病历的临床医学事件信息语料库来完成临床医学事件抽取任务的研究。受标注成本的限制,自主构建的临床医学事件信息语料库规模通常很小,且包含着大量的噪声。然而,模型性能和鲁棒性的提升很大程度上取决于标注好的训练数据的规模,小规模训练数据会限制模型的性能和鲁棒性的提升。为了解决这一问题,一些研究人员将外部特征集成到汉字表征中,以提高模型性能^[40-41]。然而,这种方法的性能好坏依赖于引入的外部特征的优劣,当引入的外部特征包含的信息越充分,模型性能越好。此外,这种方法忽略了上下文语境信息在句子语义理解中的重要性。BERT (Bidirectional Encoder Representations from Transformers)是在大量无标注文本上训练生成的,且BERT可以动态的生成上下文表征^[26]。BERT生成的上下文表征可以帮助模型正确地理解电子病历中晦涩难懂的医学术语。此外,BERT还可以提高模型在少量标注数据上的性能。因此,如何有效地利用BERT提供的上下文信息就成了一个至关重要的问题。

除了上述问题,面向中文电子病历的临床医学事件抽取任务还常面临着数据集类别分布不均衡的问题。数据集类别分布不均衡的问题会降低模型的整体性能,促使模型的决策偏向于数据规模大的类别,从而导致模型对于数据规模小的类别分类错误。因此,如何提高模型在数据规模小的类别上的性能且不降低在其他类别上的性能就成为了另一个值得关注的问题。

为了解决上述两个问题，本章提出了一种融合对抗训练与预训练语言模型的 Encoder-Decoder 框架。该方法将 BERT 生成的上下文语境信息融合到汉字表征中，以帮助模型理解特定语境下的词汇。此外，为了解决数据集类别分布不均衡的问题，本章引入了一种对抗训练的损失函数，以提高模型在数据规模小的类别上的性能。最后，在自主构建的语料库上完成了模型的训练和评估。实验结果证明，本章提出的模型在整体性能和数据规模小的类别上的性能相较于本章的基线模型均有显著地提升。

4.2 任务描述

本小节将从模型处理的角度对基于中文电子病历的临床医学事件抽取任务进行介绍。本章采用深度学习的方法研究基于中文电子病历的临床医学事件抽取，在深度学习中，该任务被视为序列标注任务，旨在给句子中的每一汉字分配一个标签，该标签包含了当前位置汉字的边界信息和临床医学事件类别信息。图 4-1 展示了一个具体例子，给定如下句子“病人先前因心脏病入院”，模型将会为句子中的“心脏病”一词中每一个汉字分配边界标签“B”（Begin，表示词语的开始位置）、“M”（Middle，表示词语的中间位置）和“E”（End，表示词语的结束位置），分配临床医学事件的类型标签“Pro”（Problem，问题类型），为“入院”一词中的每一汉字分配边界标签“B”和“E”，分配临床医学事件的类型标签“Occ”（Occurrence，特定类型）。其中边界标签的作用是将临床医学事件从给定的电子病历片段中分割出来，临床医学事件的类型标签则是用于标识分割出来的临床医学事件的类型。

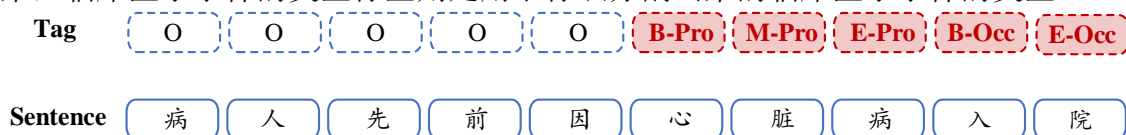


图 4-1 基于中文电子病历的事件抽取任务的示例

4.3 模型介绍

针对面向中文电子病历的临床医学事件抽取任务，本章提出了一种基于迁移学习的 Encoder-Decoder 框架，图 4-2 展示了模型的整体框架。该模型由语义编码层和标签解码层两部分构成^[42]。整个模型的处理过程如下：首先，将电子病历切割成固定长度的短句，以批处理的方式将短句打包；其次，将打包后的数据输入语义编码层，得到语义编码层的输出；然后，将语义编码层的输出输入两层 Transformer 构建的标签解码层，得到标签解码层的输出；最后，利用标签解码层的输出预测临床医学事件的类别标签。

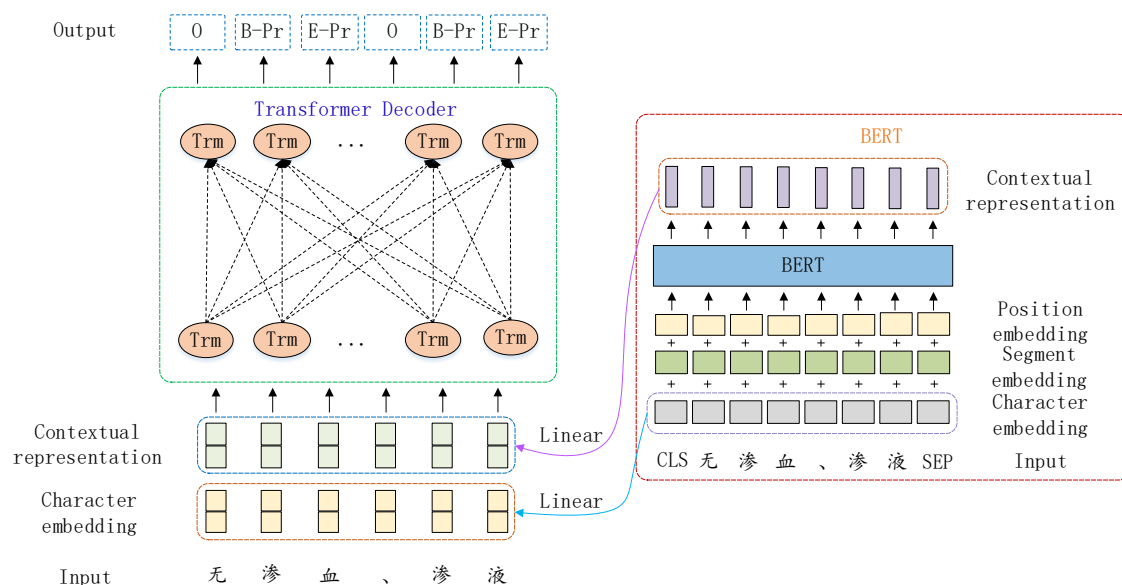


图 4-2 基于迁移学习的 Encoder-Decoder 框架

4.3.1 语义编码层

本章利用语义编码层完成两类信息的编码，分别是句子中每一个汉字信息的编码和句子的上下文语境信息的编码。其中汉字信息的编码来自 BERT 输入前的随机初始化学嵌入，上下文语境信息的编码来自 BERT 最后一层的输出。其中，上下文语境信息编码结果的生成借鉴了迁移学习的思路，属于参数迁移的迁移学习。本章使用迁移学习的目的是精进上下文语境信息的表征。本章使用的语料是电子病历文本，属于专业性文本，包含了大量晦涩难懂的医学术语。因此，对于电子病历文本的理解需要借助电子病历文本的上下文语境信息。目前，主要使用两种方式生成上下文语境信息，分别是：随机初始化和迁移学习。随机初始化的方式简单，但是在语义表征上存在缺陷。本章使用迁移学习的方式生成上下文语境信息，具体来说就是利用电子病历文本对 BERT 模型进行微调，使开放域文本上训练得到的参数向特定领域文本靠近。与随机初始化的方式相比，这种方式生成的上下文语境信息更贴近电子病历文本。在获得上述两类信息的编码后，需要对上述两类信息的编码分别进行线性变换，降低输出的维度，最后将处理后的上下文语境信息的表示和初始字嵌入拼接起来作为语义编码层最终的输出。具体的计算过程如式(4-1)至式(4-4)所示：

$$R = \text{BERT}(c_1, c_2, c_3, \dots, c_n) \quad (4-1)$$

$$R'_{c_i} = W_R R_{c_i} + b_R \quad (4-2)$$

$$E'_{c_i} = W_E E_{c_i} + b_E \quad (4-3)$$

$$T_{c_i} = [R'_{c_i}; E'_{c_i}] \quad (4-4)$$

其中 R 表示 BERT 最后一层的输出结果, R'_{c_i} 和 E'_{c_i} 分别表示 c_i 降维后的上下文语义表征和初始的字嵌入, R_{c_i} 和 E_{c_i} 分别表示 c_i 降维前的上下文语义表征和初始的字嵌入, W_R, W_E, b_R, b_E 为线性变换的参数, T_{c_i} 表示 c_i 降维后的上下文语义表征 R'_{c_i} 和初始的字嵌入 E'_{c_i} 拼接后的结果。

4.3.2 标签解码层

在标签解码过程中,本章利用两层 Transformer 构建标签解码层,用于预测句子中每一个汉字的类别标签。标签解码层的输入是语义解码层的输出,标签解码层解码的过程就是解码输出包含类型标签信息的表示。

4.3.3 对抗训练的损失函数

本章引入对抗训练的损失函数来解决由数据集类别分布不均衡导致的模型性能不均衡的问题。具体的过程如下:本章在标准的交叉熵损失函数中引入惩罚性权重,以调节不同类别的数据对于损失的贡献比例。惩罚性权重的引入间接改变了数据集中不同类别数据的比例,从而使数据分布均衡。惩罚性权重的大小与模型预测的类别标签的概率有关,数据规模大的类别,模型通常不会预测错误,惩罚性权重值小,数据规模小的类别,模型预测错的概率大,相应的惩罚性权重值大。具体的计算过程如式(4-5)所示:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K (1 - y_{i,k} P_{i,k})^\alpha y_{i,k} \log P_{i,k} \quad (4-5)$$

其中 N 表示汉字的个数, K 表示临床医学事件的种类数目, $y_{i,k}$ 表示第 i 个汉字是否属于 k 类别,如果是取值为 1,否取值为 0, $P_{i,k}$ 表示第 i 个汉字预测为 k 类别的概率, $1 - y_{i,k} P_{i,k}$ 为惩罚项, α 为惩罚因子。

4.4 实验结果与分析

为了评估本章提出的模型的有效性,在自主构建的临床医学事件信息语料库上进行了一系列实验。本小节将对实验中涉及的数据集、评价指标、实验设置和基线模型进行介绍,并对相应的实验结果进行讨论和分析。

4.4.1 数据集与评价指标

在实验阶段，对面向中文电子病历的临床医学事件信息语料库进行了如下划分。首先，将 2000 份标注好的电子病历划分为初始训练集和测试集，比例为 3:1。然后，从初始训练集和测试集中随机选择 400 份标注好的电子病历作为验证集。最后，选择初始训练集和测试集的剩余部分分别构成最终的训练集和测试集。最终的训练集由 1300 份标注好的电子病历构成，最终的验证集由 400 份标注好的电子病历构成，最终的测试集由 300 份标注好的电子病历构成。

此外，还对面向中文电子病历的临床医学事件信息语料库进行了详细统计，如表 4-1 所示。通过观察，可以发现问题类型、检查类型、治疗类型、证据词类型、特定类型以及体类型的数据在训练集、验证集和测试集中占比很大，而临床科室类别的数据占比很小。在模型性能评估环节，本章使用准确率、召回率和 F1 值进行模型性能评估。

表 4-1 面向中文电子病历的临床医学事件信息语料库的数据分布

| 类型 | 训练集 | 验证集 | 测试集 | 总计 |
|---------------------------|--------|-------|--------|--------|
| 问题(Problem) | 47669 | 4006 | 14811 | 66486 |
| 检查(Exam) | 18939 | 1949 | 5983 | 26871 |
| 治疗(Treatment) | 9622 | 989 | 2943 | 13554 |
| 临床科室(Clinical department) | 114 | 9 | 44 | 167 |
| 证据词(Evidence) | 6624 | 666 | 2136 | 9426 |
| 特定类型(Occurrence) | 3104 | 356 | 980 | 4440 |
| 体类型(Aspectual Event) | 2386 | 291 | 812 | 3489 |
| 其他(Other) | 568897 | 50763 | 173099 | 792759 |
| 总计 | 657355 | 59029 | 200808 | 917192 |

4.4.2 实验设置与基线模型

本章提出的模型利用中文预训练语言模型 BERT-Base-Chinese 完成句子中上下文语境信息的提取工作，表 4-2 为中文预训练语言模型 BERT-Base-Chinese 的超参数设置情况。

表 4-2 中文预训练语言模型 BERT-Base-Chinese 的超参数设置情况

| 参数描述 | 取值 |
|--------------------|------|
| Transformer 编码器的个数 | 12 |
| 隐状态的维度 | 768 |
| 多头注意力机制的头数 | 12 |
| 参数个数 | 110M |

本章提出的模型采用两层 Transformer 作为解码器，每一个 Transformer 由两个子层构成，分别为多头注意力机制模块和全连接的前馈神经网络结构。表 4-3 为解码器模块的超参数设置情况。

表 4-3 解码器模块的超参数设置情况

| 参数描述 | 取值 |
|--------------------|-------|
| Transformer 解码器的个数 | 2 |
| 中文字嵌入的维度 | 308 |
| 上下文语义表征的维度 | 204 |
| 多头注意力机制的头数 | 8 |
| 隐状态的维度 | 512 |
| 学习率的初始值 | 0.001 |
| Dropout 值 | 0.1 |
| 批处理的个数 | 50 |
| 惩罚因子 | 2 |

在实验部分，本章使用多组基线模型：

- (1) CRF：本章使用 CRF++ 完成 CRF 模型的相关实验。
- (2) CNN-Softmax：该模型利用 CNN 完成特征抽取，并将抽取到的特征送入多层感知机网络完成标签解码。
- (3) CNN-CRF：该模型使用 CNN 完成特征提取，并将抽取到的特征送去 CRF 结构完成标签解码。
- (4) BiGRU-Softmax：该模型利用 BiGRU (Bidirectional Gated Recurrent Unit) 完成特征抽取，并将抽取到的特征送入多层感知机网络完成标签解码。
- (5) BiLSTM-CRF：该模型利用 BiLSTM 完成特征抽取，并将抽取到的特征送去 CRF 结构完成标签解码。
- (6) Attention-based CNN-BiLSTM-CRF：汤等人利用基于注意力机制的 CNN 结构生成句子中每一个汉字的表征，然后利用 BiLSTM 完成特征抽取，最后使用

CRF 结构完成标签解码^[10]。

(7) BERT-Softmax: 该模型利用 BERT 获得句子的语义表征, 然后将获得的句子语义表征送入多层感知机网络完成标签解码。

(8) BERT-BiLSTM-Softmax: 该模型利用 BERT 获得句子的语义表征, 然后将获得句子语义表征送入 BiLSTM 完成标签解码。

(9) BERT-Transform-Softmax: 该模型利用 BERT 获得句子的语义表征, 然后将获得句子语义表征送入 Transformer 完成标签解码。

4.4.3 整体性能分析

表 4-4 记录了本章提出的模型 AdvBERT 与多组基线模型的整体性能。通过观察表格, 可以发现本章提出的模型 AdvBERT 获得了高达 83.73% 的准确率, 86.56% 的召回率, 以及 85.12% 的 F1 值。与 BiLSTM-CRF 模型相比, 本章提出的模型 AdvBERT 将 F1 值从 82.97% 提高至 85.12%, 获得了 2.15% 的提升。与 BERT-Softmax 模型相比, 本章提出的模型 AdvBERT 将 F1 值从 83.34% 提高至 85.12%, 获得了 1.78% 的提升。与上述所有模型相比, 本章提出的模型 AdvBERT 的整体性能优于所有基线模型。

表 4-4 整体性能实验结果

| 模型 | Precision (%) | Recall (%) | F1-score (%) |
|--------------------------|---------------|------------|--------------|
| CRF | 79.96 | 85.37 | 82.58 |
| CNN-Softmax | 75.20 | 83.40 | 79.09 |
| CNN-CRF | 72.93 | 74.38 | 73.65 |
| BiGRU-Softmax | 80.27 | 85.62 | 82.86 |
| BiLSTM-CRF | 81.32 | 84.68 | 82.97 |
| CNN-BiLSTM-CRF | 80.98 | 74.84 | 77.79 |
| BERT-Softmax | 81.43 | 85.34 | 83.34 |
| BERT-BiLSTM-Softmax | 82.62 | 84.77 | 83.68 |
| BERT-Transformer-Softmax | 83.54 | 85.23 | 84.38 |
| AdvBERT | 83.73 | 86.56 | 85.12 |

本章提出的模型 AdvBERT 优于其他基线模型的原因可归纳为以下几点: 首先, 本章提出的模型 AdvBERT 利用了 BERT 生成上下文语义表征, 并将句子中的每一个汉字的字嵌入与其上下文语义表征拼接起来, 使得汉字表征融入了上下文语境信息, 增强了句子中汉字的表征。其次, 本章提出的模型 AdvBERT 使用 Transformer 作为标签解码器, Transformer 中的多头注意力机制可以帮助模型关注句子中不同

位置的语义表征,可以自主选择关键信息以及关键信息的比例。最后,本章提出的模型 AdvBERT 使用对抗训练的损失函数作为模型训练的损失函数,调节了数据集中每一类别数据的损失贡献比例,间接地改变了数据集中数据的分布情况。

4.4.4 子类性能分析

表 4-5 记录了本章提出的模型 AdvBERT 在每一个子类上的性能。通过观察,可以发现本章提出的模型 AdvBERT 在数据规模大的类别上的性能表现不错。例如,本章提出的模型 AdvBERT 在抽取问题类型的临床医学事件和检查类型的临床医学事件时的 F1 值分别为 88.77%和 85.85%。

表 4-5 本章提出的模型 AdvBERT 子类性能的实验结果

| 类别 | Precision (%) | Recall (%) | F1-score (%) |
|---------------------------|---------------|------------|--------------|
| 问题(Problem) | 87.42 | 90.16 | 88.77 |
| 检查(Exam) | 82.81 | 89.12 | 85.85 |
| 治疗(Treatment) | 81.85 | 88.06 | 84.84 |
| 临床科室(Clinical department) | 74.72 | 62.44 | 68.03 |
| 证据词(Evidence) | 76.13 | 86.40 | 80.94 |
| 特定类型(Occurrence) | 89.13 | 92.44 | 90.75 |
| 体类型(Aspectual Event) | 75.96 | 86.87 | 81.05 |

在本文自主构建的语料库中,临床科室类型被视为数据规模小的类型,主要原因在于临床科室类别的数据规模远小于其他类别的数据。数据集类别分布不均衡会导致模型在每一个类别上的性能差异大,为解决此问题,本章引入了一种对抗训练的损失函数,以调节数据集中每一类别数据对于损失的贡献比例,这一策略可以迫使模型关注数据规模小的类别,间接改善数据集类别分布不均衡的问题。本章提出的模型 AdvBERT 在临床科室类别上获得了 62.44%的召回率,68.03%的 F1 值。

为了证明本章引入的对抗训练的损失函数的有效性,本小节执行了一次对比实验,对比实验的模型与本章提出的模型 AdvBERT 的差异在于损失函数的设计不同,对比实验使用标准的交叉熵损失函数,本章提出的模型使用对抗训练的损失函数。表 4-6 记录了对比实验的子类性能的实验结果。其中,对比实验在抽取问题类型的临床医学事件和检查类型的临床医学事件时的 F1 值分别为 86.49%和 84.99%。本章提出的模型 AdvBERT 与对比实验相比,在上述两个类别的临床医学事件的抽取上性能是有提升的。此外,本章提出的模型 AdvBERT 在抽取治疗类型的临床医学事件的 F1 值为 84.84%,对比实验的结果为 85.01%。本章提出的模型 AdvBERT

在治疗类型的临床医学事件上的抽取性能虽有下降,但降低程度不大。据此可以证明,本章提出的模型 AdvBERT 引入对抗训练的损失函数并不会很大程度上降低模型抽取数据规模大的类别的性能。另外,本章提出的模型 AdvBERT 在抽取临床科室类别的临床医学事件的 F1 值为 68.03%,对比实验为 67.19%。据此可以证明,本章提出的模型 AdvBERT 引入对抗训练的损失函数可以提升模型抽取数据规模小的类别的性能。

表 4-6 对比实验的子类性能的实验结果

| 类型 | Precision (%) | Recall (%) | F1-score (%) |
|---------------------------|---------------|------------|--------------|
| 问题(Problem) | 84.17 | 88.94 | 86.49 |
| 检查(Exam) | 80.42 | 90.12 | 84.99 |
| 治疗(Treatment) | 81.92 | 88.34 | 85.01 |
| 临床科室(Clinical department) | 80.52 | 57.65 | 67.19 |
| 证据词(Evidence) | 75.58 | 84.19 | 79.65 |
| 特定类型(Occurrence) | 90.21 | 90.39 | 90.30 |
| 体类型(Aspectual Event) | 76.34 | 85.71 | 80.75 |

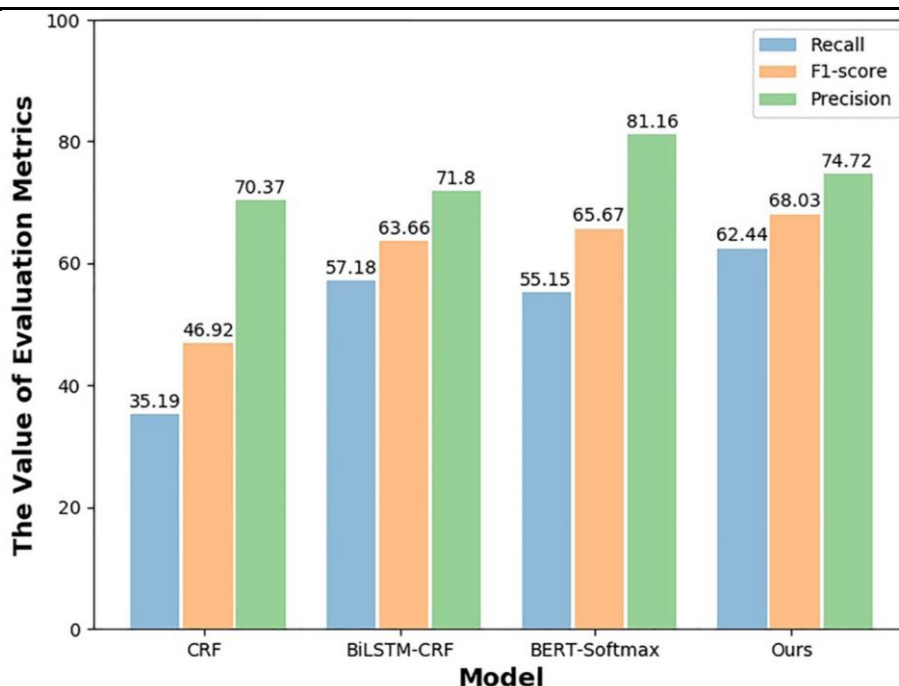


图 4-3 临床科室类别的实验结果

此外,本小节还记录了现有模型与本章提出的模型 AdvBERT 在抽取临床科室类别的临床医学事件的 F1 值,如图 4-3 所示。通过观察,可以发现本章提出的模型 AdvBERT 在抽取临床科室类别的临床医学事件的 F1 值为 68.03%,与 BERT-

Softmax 模型相比,本章提出的模型 AdvBERT 将 F1 值从 65.67%提高到 68.03%,增加了 2.36%。本章提出的模型 AdvBERT 在临床科室类别上性能的提升归功于对抗训练的损失函数,对抗训练的损失函数可以有效调节数据集中每一类别数据对于损失的贡献比例。

4.4.5 消融实验结果分析

本小节将对本章提出的模型中的每一部分的有效性进行分析,包括上下文语境信息的有效性分析,Transformer 解码器的有效性分析和对抗训练的损失函数的有效性分析。

为了证明上下文语境信息对于句子理解的重要性,本小节进行了以下的对比实验。对比实验的模型使用 BERT 中随机初始化的字嵌入表征汉字。表 4-7 记录了上下文语境信息有效性的实验结果。通过观察可以发现,本章提出的模型 AdvBERT 获得了 85.12%的 F1 值,对比实验获得了 83.20%的 F1 值,本章提出的模型 AdvBERT 相对于对比实验提升了 1.92%。这表明来自 BERT 的上下文语境信息能够提升基于中文电子病历的临床医学事件抽取任务的性能。主要原因在于上下文语境信息可以帮助模型理解特定语境下的句子语义,这一过程与人类思考的过程是一致的。

表 4-7 上下文语境信息有效性的实验结果

| 模型 | Precision (%) | Recall (%) | F1-score (%) |
|-------------------------------------|---------------|------------|--------------|
| AdvBERT(-contextual representation) | 82.34 | 84.07 | 83.20 |
| AdvBERT | 83.73 | 86.56 | 85.12 |

为了评估 Transformer 解码器的有效性,本小节进行了如下实验。对比实验使用 BiLSTM 作为解码器,损失函数采用本章提出的对抗训练的损失函数。表 4-8 记录了 Transformer 解码器有效性的实验结果。通过观察可以发现,本章提出的模型 AdvBERT 获得了最优的准确率 83.73%,召回率 86.56%以及 F1 值 85.12%。与对比实验所用模型相比,本章提出的模型 AdvBERT 将 F1 值从 84.51%提升至 85.12%,提高了 0.61%。这说明了 Transformer 比 BiLSTM 具有更好的解码能力。

表 4-8 Transformer 解码器有效性的实验结果

| 模型 | Precision (%) | Recall (%) | F1-score (%) |
|--|---------------|------------|--------------|
| BERT-BiLSTM-Softmax (adversarial loss) | 82.75 | 86.35 | 84.51 |
| AdvBERT | 83.73 | 86.56 | 85.12 |

为了证明对抗训练的损失函数对于临床科室类别的临床医学事件抽取的有效

性,本小节进行了以下实验。本小节选择 BiLSTM-CRF 模型和 BERT-Softmax 模型作为对比实验的模型,分别采用标准的交叉熵损失和对抗训练的损失函数计算损失。此外,还利用标准的交叉熵损失来代替对抗训练的损失函数在本章提出的模型 AdvBERT 上进行额外的实验。表 4-9 记录了详细的实验结果。通过观察可以发现,使用对抗训练的损失函数的 BERT-Softmax 模型优于标准的 BERT-Softmax 模型,F1 值提高了 1.35%。本章提出的模型 AdvBERT 优于相应的对比实验,F1 值提高了 0.84%。上述实验结果证明对抗训练的损失函数可以有效地缓解数据集类别不均衡的问题。此外,本小节还探讨了损失的分配策略,对比实验采用固定的损失分配比例。实验结果证明本章提出的模型 AdvBERT 优于对比实验,并获得了 4.89% 的 F1 值提升。主要原因在于本章提出的对抗训练的损失函数可以灵活地分配临床医学事件的损失贡献比例。

表 4-9 对抗训练的损失函数的有效性实验结果

| 模型 | Precision (%) | Recall (%) | F1-score (%) |
|--|---------------|------------|--------------|
| BiLSTM-CRF | 71.80 | 57.18 | 63.66 |
| BiLSTM-CRF(adversarial loss) | 73.14 | 59.28 | 65.48 |
| BERT-Softmax | 81.16 | 55.15 | 65.67 |
| BERT-Softmax(adversarial loss) | 73.28 | 61.74 | 67.02 |
| AdvBERT(standard cross-entropy loss) | 80.52 | 57.65 | 67.19 |
| AdvBERT(fixed rate of loss allocation) | 70.25 | 57.34 | 63.14 |
| AdvBERT | 74.72 | 62.44 | 68.03 |

4.5 本章小结

本章提出了一种用于面向中文电子病历的临床医学事件抽取任务的方法。首先,介绍了面向中文电子病历的临床医学事件抽取任务目前的研究难点,并从模型处理角度介绍了面向中文电子病历的临床医学事件抽取任务。其次,提出了一种基于迁移学习的 Encoder-Decoder 框架。最后,在自主构建的语料库上完成了模型的训练和评估。实验结果证明,本章提出的模型在整体性能和数据规模小的类别上的性能相较于本章的基线模型均有显著地提升。

第5章 融合图神经网络与预训练语言模型的时序关系抽取

5.1 引言

随着医学信息学的发展,电子病历在医院中的应用逐渐普及。电子病历中包含了大量的诊疗数据,这些数据反映着患者身体状况随时间的变化趋势,对这部分信息进行挖掘和分析,可辅助医生进行临床决策。电子病历囊括三类信息,分别为时间表达式、临床医学事件和时序关系。其中,时间表达式和临床医学事件属于电子病历中的显性信息,在电子病历有明确地描述,而时序关系属于电子病历中的非显性信息,在电子病历中可能被明确地描述出来,也有可能隐藏在句中,需要人为的推断。本文在上一章中探讨了临床医学事件的抽取,这一章将探讨时序关系的抽取。时序关系抽取与临床医学事件抽取相比,难度更高,属于建立在电子病历上的推理性任务,抽取出来的内容反映的是电子病历的动态信息。这部分动态信息将电子病历中的时间表达式和临床医学事件串联起来,形成了完整的链式结构,构建这样的链式结构是临床医学事件时序信息抽取任务的最终目标。

在深度学习中,时序关系抽取任务通常被视为句子分类任务,采用两种思路进行研究。一种思路认为时序关系抽取的核心是句子语义的表征,时序关系涵盖在句子语义表征中,良好的语义表征结果可以直接作为时序关系分类的依据。另一种思路认为句子中词语之间的依存关系对于时序关系抽取是至关重要的。两种思路孰好孰坏暂时没有绝对的定论,两种思路的本质都是提取能够反映句子时序性的信息,并通过包含句子时序性的特征推理句子的时序关系。因此,如何平衡句子语义和句子中词语之间依存关系对于时序关系抽取的影响是至关重要的。

除了上面提到的句子的时序信息,包含时序关系的句子元素表征对于句子时序关系的抽取也是至关重要的。句子中包含时序关系的元素表征通常有两种表示形式,分别是随机初始化的嵌入表示和引入外部特征的嵌入表示。随机初始化的嵌入表示是对元素的数字化描述,并不包含其他信息。包含外部特征的嵌入表示不仅包含元素的数字化描述,还包含反映元素特征的其他信息,表示层次更丰富。因此,选择何种元素表示方法对于句子中包含时序信息的元素表征是极为重要的。

为了解决上述两个问题,本章提出了一种融合图神经网络与预训练语言模型的时序关系抽取框架。具体来说就是利用图神经网络自主选择句子中的依存关系,而不进行人工干预,选择的自主权交给模型。对于句子中包含时序信息的元素表征部分,采用融合句子上下文语境信息的元素表征方式,考虑句子上下文信息对于元

素表征的影响。本章在构建模型时既考虑了句子语义对于时序关系抽取的影响，还考虑了句子中词语之间的依存关系对于时序关系抽取的影响，两种因素对于模型的影响在本模型中是相等的，没有高低之分，且相关信息的如何选择以及选择多少由模型自主判断。最后，本章在自主构建的时序关系语料库上完成了模型的训练和评估。实验结果证明，本章提出的模型优于所有基线模型，取得了具有竞争性的结果，并且模型整体设计思路也很好理解，模型的整体处理过程也符合人类推理时序关系的思路。

5.2 任务描述

本小节将从模型处理的角度对基于电子病历的临床医学事件时序关系抽取任务进行介绍。本章采用深度学习的方法研究基于电子病历的临床医学事件时序关系抽取任务，在深度学习中该任务被视为句子分类任务。图 5-1 展示了一个具体的例子，给定如下句子 “He had another serious concussion prior to August 17(他在 8 月 17 日之前又发生了一次脑震荡)” 以及句子中包含时序关系的 “Pro” 类型的临床医学事件 “serious concussion(脑震荡)” 和 “Date” 类型 (Date, 日期类型) 的时间表达式 “August 17(8 月 17 日)”，模型会为句子分配一个时序关系类型标签 “BEFORE”，表示 “Pro” 类型的临床医学事件 “serious concussion” 发生在 “Date” 类型的时间表达式 “August 17” 之前。图 5-1 给出的例子是时间表示式和临床医学事件之间的时序关系。除此之外，本章还讨论了临床医学事件与临床医学事件之间的时序关系。本章讨论的时序关系为句内的二元时序关系，即句子中只包含两个时序元素。

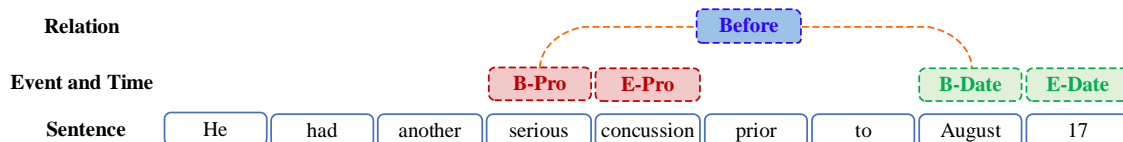


图 5-1 基于电子病历的临床医学事件时序关系抽取任务的示例

5.3 模型介绍

针对时序关系抽取任务，本章提出了一种多特征融合的时序关系抽取框架，图 5-2 展示了模型的整体框架。该模型由时序元素编码层、时序关系编码层和时序关系解码层三部分构成。整个模型的处理过程如下：首先，将电子病历切割成固定长度的短句，以批处理的方式将短句打包；其次，将打包后的数据输入 BERT 得到时序元素的词嵌入，将时序元素的词嵌入输入时序元素编码层，得到时序元素的最终编码结果；然后，将 BERT 输出的数据编码结果输入时序关系编码层，得到依存关

系的编码结果；最后，将时序元素的编码结果与依存关系的编码结果拼接起来作为时序关系解码层的输入，利用逻辑回归的分类器完成时序关系的预测。

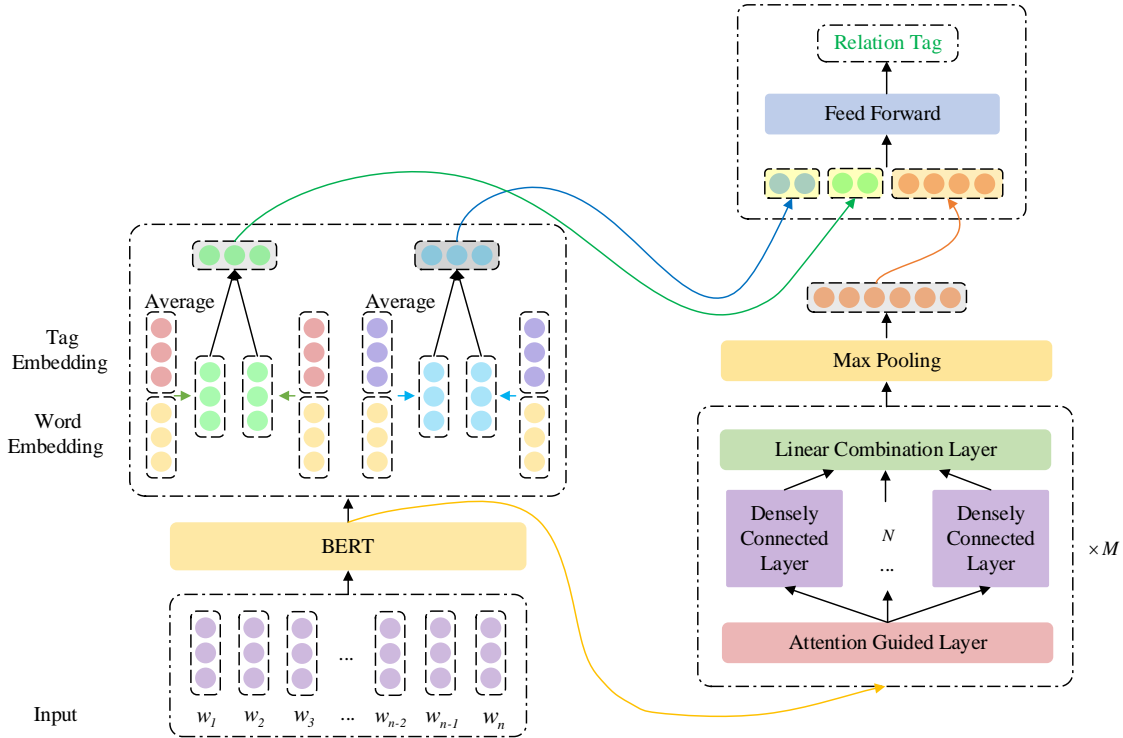


图 5-2 基于多特征融合的时序关系抽取框架

5.3.1 时序元素编码层

本章利用时序元素编码层完成句子中时间表达式和临床医学事件的编码。首先，选择 BERT 最后一层输出中每一个位置的嵌入作为句子中每一个单词融合上下文语境信息的嵌入表征。其次，考虑到每一个临床医学事件和每一个时间表达式可能由一个单词构成，也有可能由多个单词构成，将构成临床医学事件和时间表达式的子结构嵌入与对应的类型标签嵌入融合，融合的方式是将对应维度值相加，这样就得到了构成临床医学事件和时间表达式的嵌入表征序列，然后分别对得到的临床医学事件和时间表达式的嵌入表征序列执行平均化操作，平均化操作就是将构成临床医学事件和时间表达式的嵌入序列按其子结构排列的顺序，将对应维度值相加，再除以子结构的数目，最后还需要对得到的结果进行降维，即执行一次线性变换。具体的计算过程如式(5-1)至式(5-7)所示：

$$R = \text{BERT}(c_1, c_2, c_3, \dots, c_n) \quad (5-1)$$

$$E_I^j = R_{c_i} + T_{c_i} \quad (5-2)$$

$$T_I^k = R_{c_s} + T_{c_s} \quad (5-3)$$

$$E_l = \text{Average}(E_l^1, E_l^2, E_l^j, \dots, E_l^a) \quad (5-4)$$

$$T_l = \text{Average}(T_l^1, T_l^2, T_l^k, \dots, T_l^b) \quad (5-5)$$

$$E'_l = W_E E_l + b_E \quad (5-6)$$

$$T'_l = W_T T_l + b_T \quad (5-7)$$

其中 R 表示 BERT 最后一层的输出结果。 E_l^j 和 T_l^k 分别表示构成临床医学事件和时间表达式的子结构的嵌入表征,这一部分由字嵌入和标签嵌入构成, R_{c_i} 和 R_{c_s} 分别表示临床医学事件和时间表达式子结构的字嵌入, T_{c_i} 和 T_{c_s} 分别表示临床医学事件和时间表达式子结构的标签嵌入。 E'_l 和 T'_l 分别表示临床医学事件和时间表达式的最终表征, E_l 和 T_l 分别表示临床医学事件和时间表达式的初始表征, W_E, W_T, b_E, b_T 表示相应的线性变换的参数。

5.3.2 时序关系编码层

时序关系编码层的任务是基于注意力指导的图神经网络和密集连接层完成句子时序信息的提取。时序关系编码层由 M 层相同的结构堆叠而成,每一层由三类子层构成,分别为注意力指导层、密集连接层和线性融合层。其中,第一个时序关系编码层的输入是 BERT 最后一层的输出。其余时序关系编码层的输入为上一级的时序关系编码层的输出。下面将对上述三类子层进行详细介绍。

5.3.2.1 注意力指导层

注意力指导层的任务是完成句子中词语之间依存关系的编码。目前,词语之间依存关系最常用的编码方式为基于子树的依存关系编码^[43]。该方法借助句法分析工具完成依存关系分析^[44-46]。依存句法最早是由法国语言学家 Lucien Tesnière 在其著作《结构句法基础》中提出,作者将句子解析成一棵依存句法树,描述出各个词语之间的依存关系,即词语之间在句法上的搭配关系,这种搭配关系与句子语义密切相关。基于子树的依存关系编码的输入是原始依存关系树的子树,该子树来源于句法分析工具解析的结果,这种方法存在一个弊端,即句子中词语之间依存关系考虑不完整,通过句法分析工具对原始依存关系树进行了修剪,这种修剪方式可以认为是一种“硬剪枝”策略。本章采用融合多头注意力机制的图神经网络进行依存关系的提取,可以认为是一种“软剪枝”策略。图 5-3 展示了融合多头注意力机制的图神经网络进行依存关系编码的过程。句子中词语之间的依存关系属于图类型数据,句子中的词语为图中的节点,词语之间的依存关系为节点之间的边。模型的初始输入只包含节点嵌入,经过多头注意力机制的处理,依存关系图有原来的 1 个

变为多个，依存关系图中的边权值会发生变化。图 5-3 中， G^I 至 G^N 是在原始依存关系图 G^0 上更新边权值得到的新的依存关系图。

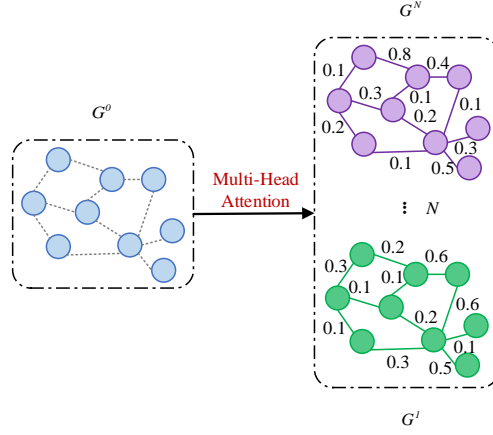


图 5-3 融合多头注意力机制的图神经网络进行依存关系编码的过程

依存关系图更新的过程实质上为依存关系图中边权值的更新过程，依存关系图中边权值的计算过程与多头注意力机制中注意力权重的计算过程是完全一致的。具体计算过程如式(5-8)所示：

$$A' = \text{softmax} \left(\frac{\mathbf{Q}W_i^Q \times (\mathbf{K}W_i^K)^T}{\sqrt{d}} \right) \quad (5-8)$$

其中， A' 表示依存关系图的边权值矩阵， \mathbf{Q} 和 \mathbf{K} 表示依存关系图中的节点表征，初始的 \mathbf{Q} 和 \mathbf{K} 为 BERT 最后一层输出中的节点表征，之后的 \mathbf{Q} 和 \mathbf{K} 为上一层时序关系编码层的输出中抽取的依存关系图的节点表征， W_i^Q 和 W_i^K 为线性变换的参数。

5.3.2.2 密集连接层

密集连接层的任务就是对依存关系图进行更深层的特征提取^[47-48]。密集连接层在进行更深层的特征提取时需要利用注意力指导层生成的依存关系图的边权值。密集连接层模块的个数与注意力指导层中多头注意力机制的头数一致。密集连接层处理的过程是并行操作的。图 5-4 展示了密集连接层进行深层特征提取的过程。

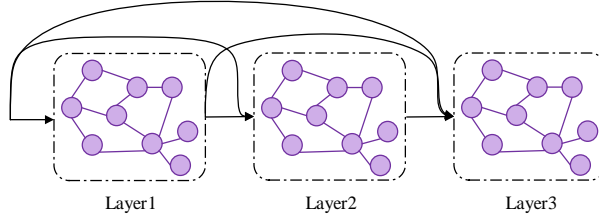


图 5-4 密集连接层进行深层特征提取的过程

密集连接层的每一层输入包含两部分信息：依存关系图的节点表征和之前层

的隐藏输出。密集连接层每一层输出的维度由依存关系图的节点表征的维度和密集连接层的层数决定，密集连接层最后一层的输出是密集连接层所有层输出集成的结果。密集连接层本质是对依存关系图的节点信息进行更新。具体计算过程如式(5-9)至式(5-12):

$$g_c^l = [x_c; H_c^l; H_c^2; \dots; H_c^{l-1}] \quad (5-9)$$

$$H_{t_c}^l = \rho\left(\sum_{d=1}^n A_{c,d}^t W_t^l g_d^l + b_t^l\right) \quad (5-10)$$

$$H_{out}^{t_c} = [H_{t_c}^1; H_{t_c}^2; H_{t_c}^3; \dots; H_{t_c}^l; \dots; H_{t_c}^L] \quad (5-11)$$

$$H_{out}^t = [H_{out}^{t_1}; H_{out}^{t_2}; H_{out}^{t_3}; \dots; H_{out}^{t_c}; \dots; H_{out}^{t_n}] \quad (5-12)$$

其中, g_c^l 表示密集连接层 l 层的输入, x_c 表示依存关系图的节点嵌入表征, H_c^l 表示节点 c 在密集连接层 l 层的输出, $H_{t_c}^l$ 表示第 t 个密集连接层中节点 c 在密集连接层 l 层的输出, 该值融合了与节点 c 有依存关系的所有节点的信息。 $A_{c,d}^t$ 表示第 t 个密集连接层中与节点 c 有依存关系的节点 d 之间的边权值。 g_d^l 表示与节点 c 有依存关系的节点 d 在密集连接层 l 层的输入, W_t^l 和 b_t^l 为线性变换的参数。 $H_{out}^{t_c}$ 表示节点 c 在密集连接层最后一层的输出。 H_{out}^t 表示第 t 个密集连接层最后一层的输出, 该值包含了更新后的依存关系图的所有节点信息。

5.3.2.3 线性融合层

线性融合层的任务是融合密集连接层输出的 N 层信息。当前线性融合层的输出会作为下一个时序关系编码层的输入。具体的计算过程如式(5-13)和式(5-14):

$$H_{out} = [H_{out}^1; H_{out}^2; H_{out}^3; \dots; H_{out}^t; \dots; H_{out}^N] \quad (5-13)$$

$$H_{comb} = W_{comb}^T H_{out}^T + b_{comb} \quad (5-14)$$

其中, H_{out} 表示密集连接层输出结果的拼接形式, H_{comb} 表示线性融合层输出的结果, W_{comb} 和 b_{comb} 为对应的线性变换的参数。

5.3.3 时序关系解码层

时序关系解码层的任务是解码时序关系编码层的信息, 输出时序关系的类型标签。时序关系解码层的输入融合了两部分信息: 时序元素的表征和时序关系的表征。在进行时序关系解码之前, 需要对时序关系编码层的输出结果执行最大池化和线性变换。然后, 将时序元素表征和时序关系表征进行拼接送入前馈神经网络完成信息的融合。最后, 将融合的信息送入逻辑回归的分类器, 输出最后的时序关系标

签^[49-51]。具体计算过程如式(5-15)至式(5-17):

$$H_{\text{relation}} = f(H_{\text{comb}}^T) \quad (5-15)$$

$$H'_{\text{relation}} = W_{\text{relation}} H_{\text{relation}} + b_{\text{relation}} \quad (5-16)$$

$$H_{\text{final}} = \text{FFNN}([E'_I; T'_I; H'_{\text{relation}}]) \quad (5-17)$$

其中, H_{relation} 表示对时序关系编码层的输出执行最大池化的结果。 H'_{relation} 表示对 H_{relation} 进行线性变换后的结果, W_{relation} 和 b_{relation} 是线性变换的参数。 H_{final} 表示时序元素表征与时序关系表征融合后的结果。

5.4 实验结果与分析

为了评估本章提出的模型的有效性, 本小节在自主重构的临床医学事件时序关系语料库上进行了一系列实验。本小节将对实验中涉及的数据集、评价指标、实验设置和基线模型进行介绍, 并对相应的实验结果进行讨论和分析。

5.4.1 数据集与评价指标

本小节在实验阶段使用的数据集来自自主重构的临床医学事件时序关系语料库。临床医学事件时序关系语料库包含 4000 条电子病历片段。本小节对临床医学事件时序关系语料库进行了如下划分, 选取 2500 条作为训练集数据, 500 条作为验证集数据, 1000 条作为测试集数据。此外, 还对临床医学事件时序关系语料库的数据分布情况进行了统计, 表 5-1 记录了临床医学事件时序关系语料库的数据分布情况。在模型性能评估环节, 使用准确率、召回率和 F1 值进行模型性能评估。

表 5-1 临床医学事件时序关系语料库的数据分布

| 类型 | 训练集 | 验证集 | 测试集 | 总计 |
|----------------|------|-----|------|------|
| BEFORE | 945 | 189 | 378 | 1512 |
| AFTER | 1077 | 215 | 432 | 1724 |
| OVERLAP | 202 | 41 | 81 | 324 |
| SIMULTANEOUS | 165 | 33 | 66 | 264 |
| DURING | 12 | 3 | 5 | 20 |
| BEFORE_OVERLAP | 15 | 3 | 6 | 24 |
| BEGUN_BY | 35 | 8 | 14 | 57 |
| ENDED_BY | 49 | 8 | 18 | 75 |
| 总计 | 2500 | 500 | 1000 | 4000 |

5.4.2 实验设置与基线模型

本章提出的模型利用英文预训练语言模型 BERT-Base-Cased 完成时序元素子结构的编码工作，表 5-2 为 BERT-Base-Cased 的超参数设置情况。

表 5-2 BERT-Base-Cased 的超参数设置情况

| 参数描述 | 取值 |
|--------------------|------|
| Transformer 编码器的个数 | 12 |
| 隐状态的维度 | 768 |
| 多头注意力机制的头数 | 12 |
| 参数个数 | 110M |

本章提出的模型采用时序关系编码层完成时序信息的提取，时序关系编码层由多个相同的结构堆叠而成，每一层包含三类子层，分别为注意力指导层、密集连接层和线性融合层。表 5-3 为时序关系编码层的超参数设置情况。

表 5-3 时序关系编码层的超参数设置情况

| 参数描述 | 取值 |
|-------------|-------|
| 时序关系编码层的个数 | 2 |
| 输入嵌入的维度 | 768 |
| 多头注意力机制的头数 | 4 |
| 密集连接层的个数 | 4 |
| 每一个密集连接层的层数 | 3 |
| 学习率的初始值 | 0.001 |
| 批处理的个数 | 50 |

在实验部分，本章使用多个基线模型与本章提出的方法进行比较：

(1) CNN：该模型利用 CNN 完成句子时序信息的提取，并将抽取到的句子的时序信息送入逻辑回归的分类器中完成时序关系的预测。

(2) BiLSTM：该模型利用 BiLSTM 完成句子时序信息的提取，并将抽取到的句子的时序信息送入逻辑回归的分类器中完成时序关系的预测。

(3) CNN-BiGRU：该模型利用 CNN 和 BiGRU 完成句子时序信息的提取，并将抽取到的句子的时序信息送入逻辑回归的分类器中完成时序关系的预测。

(4) BERT：该模型利用 BERT 完成句子时序信息的提取，并将抽取到的句子的时序信息送入逻辑回归的分类器中完成时序关系的预测。

5.4.3 整体性能分析

表 5-4 记录了本章提出的模型 BERT-GNN 与多组基线模型的整体性能。通过观察表格,可以发现本章提出的模型 BERT-GNN 获得了 72.25%的准确率, 69.18%的召回率和 70.68%的 F1 值。本章提出的模型 BERT-GNN 在准确率、召回率和 F1 值上远高于所有基线模型, 获得了最优的性能。与 BiLSTM 相比, 本章提出的模型 BERT-GNN 将 F1 值从 66.39%提升至 70.68%, 提高了 4.29%。与 BERT 模型相比, 本章提出的模型 BERT-GNN 将 F1 值从 68.85%提升至 70.68%, 提高了 1.83%。本章提出的模型 BERT-GNN 优于其他基线模型的原因包括以下几点: 本章提出的模型 BERT-GNN 采用 BERT 生成的上下文语境信息与单词嵌入拼接的结果表征时序元素, 这方式能够帮助模型更好的理解句子中的词汇。此外, 本章提出的模型 BERT-GNN 还引入依存关系编码层来获取词语之间的依存关系, 利用词语之间的相关性来建立时序元素与包含句子时序特征的词语之间的联系, 从而间接推理出句子中时序元素之间的时序关系类型。

表 5-4 整体性能实验结果

| 模型 | Precision (%) | Recall (%) | F1-score (%) |
|-----------|---------------|------------|--------------|
| CNN | 64.97 | 62.51 | 63.72 |
| BiLSTM | 67.31 | 65.49 | 66.39 |
| CNN-BiGRU | 62.14 | 60.23 | 61.17 |
| BERT | 69.54 | 68.17 | 68.85 |
| BERT-GNN | 72.25 | 69.18 | 70.68 |

5.4.4 消融实验结果分析

本小节将对本章提出的模型 BERT-GNN 中的每一部分的有效性进行分析, 包括语义编码层的有效性分析, 时序关系编码层的有效性分析。

为了验证语义编码层的有效性, 本小节进行了以下的对比实验。对比实验的模型使用随机初始化的单词嵌入和标签嵌入的拼接结果作为单词的表征, 其余部分与本章提出的模型 BERT-GNN 是一致的。表 5-5 记录了语义编码层有效性的实验结果。通过观察可以发现, 本章提出的模型 BERT-GNN 获得了 70.68%的 F1 值, 对比实验获得了 69.79%的 F1 值, 本章提出的模型 BERT-GNN 相比对比实验 F1 值提升了 0.89%。这表明融合上下文语境信息的单词表征能够表达单词在特定语境中含义。随机初始化的单词嵌入没有语境信息作为辅助特征, 单词的表征是不完整的, 不符合人类结合语境信息理解单词含义的思考方式。

表 5-5 语义编码层有效性的实验结果

| 模型 | Precision (%) | Recall (%) | F1-score (%) |
|--------------------------------------|---------------|------------|--------------|
| BERT-GNN(-contextual representation) | 70.13 | 69.45 | 69.79 |
| BERT-GNN | 72.25 | 69.18 | 70.68 |

为了评估时序关系编码层的有效性,进行了如下的对比实验。对比实验仅使用句子语义作为时序信息来推理句子中时序元素之间的时序关系类型。表 5-6 记录了时序关系编码层有效性的实验结果。通过观察可以发现,本章提出的模型 BERT-GNN 获得了最优的准确率 72.25%, 召回率 69.18%和 F1 值 70.68%。与仅使用句子语义进行时序关系推理的模型相比,本章提出的模型 BERT-GNN 在时序关系的推理中具有一定的优势,句子语义可以帮助模型从宏观角度理解句子,词语之间的依存关系可以帮助模型从微观角度理解句子的结构信息。

表 5-6 时序关系编码层有效性的实验结果

| 模型 | Precision (%) | Recall (%) | F1-score (%) |
|---|---------------|------------|--------------|
| BERT-GNN(-dependency relation encode layer) | 69.93 | 68.44 | 69.18 |
| BERT-GNN | 72.25 | 69.18 | 70.68 |

5.5 本章小结

本章提出了一种用于临床医学事件时序关系抽取任务的模型框架。在介绍模型框架之前,先介绍了临床医学事件时序关系抽取任务目前的研究难点,然后从模型处理的角度介绍了临床医学事件时序关系抽取任务的定义。之后介绍了本章提出的融合图神经网络与预训练语言模型的时序关系抽取框架。最后,本章进行了多组实验,并从模型整体性能分析和消融实验结果分析两个角度进行了模型性能的评估。实验结果证明,本章提出的模型优于本章所有的基线模型。本章提出的模型的优势包括以下两点:第一,本章提出的模型在时序元素的表征上考虑了上下文语境信息对于词语表征的重要性;第二,本章提出的模型考虑了句子中词语之间的依存关系对于时序关系抽取的重要性。

结 论

基于电子病历的临床医学事件时序信息抽取是医疗领域与自然语言处理领域交叉形成的跨领域性研究任务,旨在利用自然语言处理技术高效地抽取电子病历中的时序信息。抽取到的时序信息进行信息整合和逻辑推演后可作为关键信息用于构建患者临床时间线,辅助医生完成临床决策。作为时序信息抽取技术在医学领域的大胆尝试,基于电子病历的临床医学事件时序信息抽取受到了来自医疗领域和自然语言处理领域众多学者和专家的关注。目前,学术界和工业界将基于电子病历的临床医学事件时序信息抽取技术的研究划分为三个层次。第一层是基于电子病历的临床医学事件时序信息语料库构建。第二层是基于电子病历的临床医学事件时序信息抽取。第三层是患者临床时间线构建。本文的研究工作主要集中在第一层和第二层,包括基于电子病历的临床医学事件时序信息语料库的构建和基于电子病历的临床医学事件时序信息抽取技术的相关内容,主要的研究成果包括以下几个方面:

(1) 针对临床医学事件时序信息语料库匮乏问题,本文自主构建了一个面向中文电子病历的临床医学事件信息语料库,并重构了一个面向英文电子病历的句内时序关系语料库。在面向中文电子病历的临床医学事件信息语料库的构建工作中,完成了语料的选取、标注规范的制定、标注示例的说明、标注结果的规格化,并对语料库数据进行了统计和分析。在制定面向中文电子病历的临床医学事件信息语料库的标注规范时,参考了 2012 I2B2 时序关系挑战赛发布的面向英文电子病历的临床医学事件时序信息语料库的标注规范,并结合了中文电子病历的行文特点对标注规范做了调整,形成了适合中文电子病历的临床医学事件信息语料库的标注规范。然后,根据标注规范,开发了标注工具,利用标注工具完成了标注结果的规格化处理。最后,根据制定的语料库标注规范和开发的标注工具,在医疗领域相关人员指导下,标注人员对标注规范熟悉的情况下,完成了面向中文电子病历的临床医学事件信息语料库的构建。在面向英文电子病历的句内时序关系语料库的重构工作中,完成了语料的选取、时序关系分类体系的确立、语料库的重构,并对语料库数据进行了统计和分析。

(2) 针对现有模型对于句子上下文语境信息建模不充分、临床医学事件抽取性能分布不均衡的问题,本文提出了一种融合预训练语言模型与对抗训练的 Encoder-Decoder 框架。该方法将句子的上下文语境信息集成到汉字表征中来帮助模型理解特定语境下的词汇。此外,还引入惩罚性权重来调整每个类别的损失贡献

比例,以解决数据集类别分布不均衡引起的模型性能分布不均衡的问题。实验结果证明,本章提出的模型在整体性能和数据规模小的类别上的性能相较于本章的基线模型均有显著地提升。

(3) 针对现有模型对于句子语义理解不完全、时序信息抽取不充分的问题,本文提出了一种融合图神经网络与预训练语言模型的方法,该方法利用融合上下文语境信息的词嵌入表征句子中的时序元素,利用融合多头注意力机制的图神经网络捕获句子中词语之间的依存关系,利用词语之间的相关性来建立时序元素与包含句子时序特征的词语之间的联系,从而间接推理出句子中时序元素之间的时序关系类型。实验结果证明,本章提出的模型优于所有基线模型,取得了具有竞争性的结果。

经过上述的研究,本文在基于电子病历的临床医学事件时序信息抽取任务上取得了初步的成果。但是,仍有许多内容有待研究:

(1) 目前对于临床医学事件时序关系抽取任务的研究主要集中在英文电子病历上。本文计划在后期自主构建一个面向中文电子病历的临床医学事件时序关系语料库。

(2) 本文在临床医学事件时序信息抽取任务的研究中将临床医学事件抽取和时序关系抽取独立研究。然而,在实际应用场景中,临床医学事件抽取和时序关系抽取两个任务并非是独立的,而是具有一定联系的,本文计划在后期可以开展临床医学事件和时序关系的联合抽取研究。

(3) 本文在临床医学事件时序关系抽取任务中,仅研究了句内时序关系抽取。在实际应用场景中,临床医学事件时序关系抽取任务不仅包含句内时序关系抽取,还包括跨句子的时序关系抽取和文档创建时间与文档范围内事件的时序关系抽取。本文,计划在后期开展跨句子的时序关系抽取和文档创建时间与文档范围内事件的时序关系抽取任务的研究。

参考文献

- [1] Kraus S, Blake C, West S L. Information extraction from medical notes[J]. Medinformation, 2007, 1-2.
- [2] Gerner M, Nenadic G, Bergman C M. LINNAEUS: A species name identification system for biomedical literature[J]. BMC Bioinformatics, 2010, 11(1): 1-17.
- [3] Coden A, Savova G, Sominsky I, et al. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model[J]. Journal of biomedical informatics, 2009, 42(5): 937-949.
- [4] Savova G K, Masanz J J, Ogren P V, et al. Mayo clinical text analysis and knowledge extraction system (CTAKES): architecture, component evaluation and applications[J]. Journal of the American Medical Information Association, 2010, 17(5): 507-513.
- [5] Ferrucci D, Lally A. UIMA: An architectural approach to unstructured information processing in the corporate research environment[J]. Natural Language Engineering, 2004: 1-26.
- [6] Vlachos A. Evaluating and combining biomedical named entity recognition systems[C]. Proceedings of the Workshop on BioNLP: Biological, translation, and clinical language processing, Association for Computational Linguistics, 2007, 199-200.
- [7] Ju Z F, Wang J, Zhu F. Named entity recognition from biomedical text using SVM[C]. International Conference on Bioinformatics and Biomedical Engineering, Institute of Electrical and Electronics Engineers, 2011, 1-4.
- [8] Leaman R, Wei C H, Zou C, et al. Mining chemical patents with an ensemble of open systems[J]. Database, 2016 (2016).
- [9] Wang H, Zhang W, Zeng Q, et al. Extracting important from Chinese Operation Notes with natural language processing methods[J]. Journal of biomedical informatics, 2014, 48: 130-136.
- [10] Tang B Z, Wang X L, Yan J, et al. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF[C]. BMC Med, Inf. Decis. Making, 2019, 19(3): 74.
- [11] Cao P F, Chen Y B, Liu K, et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism[C]. Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, (2018), 182-192.
- [12] Daniel M, Abdessamad E. An unsupervised approach to recognizing discourse relations[C]. In Proceeding of the 40th Annual Meeting of the Association for Computational Linguistic, 2002: 368-375.
- [13] Inderjeet M, Marc V, Ben W, et al. Machine learning of temporal relations[C]. In Proceedings of the 21st Annual International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, 2006: 753-760.
- [14] Inderjeet M, Ben W, Marc V, et al. Three approaches to learning TLINKS in TimeML[J]. Technical Report CS-07-268, Computer Science Department, 2007.

- [15] Xu Y, Jia R, Mou L L, et al. Improved relation classification by deep recurrent neural networks with data augmentation[J]. arXiv preprint arXiv: 1601.03651, 2016.
- [16] Cheng F, Miyao Y. Classifying temporal relations by bidirectional LSTM over dependency paths[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017:1-6.
- [17] 付剑锋. 面向事件的知识处理研究[D]. 上海: 上海大学, 2010: 1-2.
- [18] 罗梟. 基于深度学习的自然语言研究综述[J]. 智能计算机与应用, 2020, 10(4): 133-134.
- [19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv: 1706.03762, 2017.
- [20] Ilya S, Oriol V, Quoc Le. Sequence to sequence learning with neural networks[J]. arXiv preprint arXiv: 1409.3215, 2014.
- [21] Alex G. Generating sequences with recurrent neural networks. arXiv preprint arXiv: 1308.0850, 2013.
- [22] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
- [23] Jimmy B, Jamie R K, Geoffrey H. Layer normalization. arXiv preprint arXiv: 1607.06450, 2016.
- [24] Sepp H, Jürgen S. Long short-term memory[J]. Neural computation, 9(8): 1735-1780, 1997.
- [25] Chung J Y, Gülçehre C, Kyunghyun C, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv: 1412.3555, 2014.
- [26] Jacob D, Chang M W, Kenton Lee, et al. BERT: Pre-training of deep bidirectional transformer for language understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.
- [27] Andrew M D, Quoc V Le. Semi-supervised sequence learning[J]. arXiv preprint arXiv: 1511.01432, 2015.
- [28] Jeremy H, Sebastian R. Universal language model fine-tuning for text classification[J]. arXiv preprint arXiv: 1801.06146, 2018.
- [29] Alec R, Karthik N, Tim S, et al. Improving language understanding with unsupervised learning[J]. 2018.
- [30] Samuel B, Gabor A, Christopher P, et al. A large annotated corpus for learning natural language inference[J]. arXiv preprint arXiv: 1508.05326, 2015.
- [31] Adina W, Nikita N, Samuel B. A broad-coverage challenge corpus for sentence understanding through inference[J]. arXiv preprint arXiv: 1704.05426, 2017.
- [32] Pranav R, Zhang J, Konstantin L, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv: 1606.05250, 2016.
- [33] Matthew P, Mark N, Mohit I, et al. Deep contextualized word Representations[J]. arXiv preprint arXiv: 1802.05365, 2018.
- [34] Wilson T. Cloze procedure: A new tool for measuring readability[J]. Journalism quarterly, 1953, 30(4): 415-433.

- [35] Wu Z H, Pan S R, Chen F W, et al. A Comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning system, 2020.
- [36] Mccallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[J]. 2003.
- [37] Li L S, Jin L K, Jiang Z C, et al. Biomedical named recognition based on extended recurrent neural networks[C]//2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Institute of Electrical and Electronic Engineers Computer Society, 2015, 649–652.
- [38] Li L S, Jiang Y X, Biomedical named entity recognition based on the two channels and sentencelevel reading control conditioned LSTM-CRF[C]//2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Institute of Electrical and Electronic Engineers Computer Society, 2017, 380–385.
- [39] Zhou X S, Xiong H Q, Zeng S H, et al. An approach for medical event detection in Chinese clinical notes of electronic health records[J]. BMC medical informatics and decision making, 2019, 19(2): 31-37.
- [40] Ouyang E, Li Y X, Jin L, et al. Exploring n-gram character presentation in bidirectional RNN-CRF for Chinese clinical named entity recognition[C]// CEUR Workshop Proceedings, Institute of Electrical and Electronic Engineers Computer Society, 2017, 1976: 37–42.
- [41] Wang Y F, Ananiadou S, Tsujii J I. Improve Chinese clinical named entity recognition performance by using the graphical and phonetic feature[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Institute of Electrical and Electronic Engineers Computer Society, 2018, 1582–1586.
- [42] Zhang Z C, Zhang M Y, Zhou T, et al. Pre-trained language model augmented adversarial training network for Chinese clinical event detection[J], Mathematical Biosciences and Engineering, 2020, 17(4): 2825-2841.
- [43] Kelvin X, Jimmy B, Ryan K, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// International conference on machine learning PMLR, 2015: 2048-2057.
- [44] Qiu, X P, Zhang Q, Huang J X. Fudannlp: A toolkit for chinese natural language processing[C]// Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations, 2013: 49-54.
- [45] Qi P, Timothy D, Zhang Y H, et al. Universal Dependency Parsing from Scratch In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2018, 160-170.
- [46] Che W X, Li Z H, Liu T. LTP: A Chinese Language Technology Platform. In Proceedings of the Colling 2010: Demonstrations. 2010: 13-16.
- [47] Guo Z J, Zhang Y, Teng Z Y, et al. Densely connected graph convolutional networks for graph-to-sequence learning[J]. Transactions of the Association of Computational Linguistics, 7(2019): 297-312.
- [48] Huang G, Liu Z, Laurens V M, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

- [49] Zhang Y H, Qi P, Christopher D. Graph convolution over pruned dependency trees improves relation extraction[J]. arXiv preprint arXiv: 1809. 10185, 2018.
- [50] Adam S, David R, David G T, et al. A simple neural network module for relational reasoning[J]. arXiv preprint arXiv: 1706. 01427, 2017.
- [51] Kenton L, He L H, Mike L, et al. End-to-end neural coreference resolution[J]. arXiv preprint arXiv: 1707. 07045, 2017.

致 谢

光阴似箭，岁月如歌。三年的研究生生活如白驹过隙，转瞬即逝。蓦然回首，思绪万千。三年的点点滴滴，感慨不已，欣慰之余而又无比庆幸。值得欣慰的三年的时间我是在汗水和拼搏中度过的，学到了许多受益无穷的东西，庆幸的是我来到了一个很好的环境，遇到了很多的良师益友。导师严谨的治学态度、渊博的学识和对工作的强烈责任心深深地感染了我，学业上他给了我精心的指导和帮助，使我学到了许多为人治学的道理，同时还在思想、生活上给我以无微不至的关怀和帮助，使我能够顺利地完成学业。在此谨向良师益友表示最衷心的感谢！

俗话说，“一日为师，终身为父”，首先我要感谢我的研究生导师张志昌教授，张老师渊博的专业知识，严肃的科学态度，严谨的治学精神和精益求精的工作作风深深地感染和激励着我，使我受用一生。张老师不仅仅是我专业上的引路人，更像一位慈父，在学习、生活中关心我，帮助我，从课题研究到论文选题及论文修改反复斟酌，反复修改，点拨迷津。他的言传身教是我一生的财富。

其次，我要感谢计算机学院的各位老师和并肩奋斗的兄弟姐妹们。虽历时三载，但学院老师严谨的科学态度，精益求精的工作作风，诲人不倦的高尚师德，严以律己、宽以待人的崇高风范，朴实无华、平易近人的人格魅力深深地感染和激励着我，鞭策着我。还有三年来一直陪伴在我身边的同窗好友，实验室互帮互助的点滴让我的研究生生活变得绚丽多彩；寝室内彼此之间姐妹般的感情，将成为我永远不能磨灭的美好回忆。感谢自然语言处理研究组的张珍文师兄和周侗师兄对我科研上的帮助。感谢自然语言处理研究组的张治满师兄、赵娟师姐和张瑞芳师姐对于我生活上的帮助。感谢实验室同级的于沛霖同学、曾扬扬同学、邱炎龙同学、朱林同学和杨晓莉同学三年内对我学业和生活上帮助。感谢实验室的刘丹师妹、秦小会师妹、严龙师弟、白炎青师弟和陈浩元师弟对我学业的帮助，他们的出现让我的生活变得丰富多彩。感谢研究生期间同寝的三位室友张宁宁同学、张芬同学和张梦炜同学。每当我迷茫和失落时，她们都会陪在我身边，对我进行开导和支持。

需要特别感谢的是我的父母。二十多年无怨无悔的付出，父母的养育之恩无以为报，他们是我多年求学路上的坚强后盾，在我面临人生选择的迷茫之际，为我排忧解难，对我无私的爱与照顾是我不断前进的动力。

最后向百忙之中参与本次论文评审的各位老师和参与本人毕业答辩的各位老师由衷感谢。千言万语汇成一句话，感谢西北师范大学给我提供了学习机会；老师在我人生路上给予启迪；同学们共同进步，共同拼搏，互相鼓励，结下了深厚的

友谊。谨以此文献给我尊敬的师长、亲人和朋友们，你们的出现让我的生活变得丰富多彩。

个人简历、在学期间发表的学术论文及研究成果

1. 个人简历

- [1] 1995 年 07 月 01 日出生于甘肃省金昌市金川区。
- [2] 2013 年 09 月——2017 年 06 月,在杭州师范大学杭州国际服务工程学院(系)软件工程专业本科毕业并获得工学学士学位。
- [3] 2018 年 09 月——至今,在西北师范大学计算机科学与工程学院(系)计算机技术专业攻读工程硕士专业学位。

2. 在学期间发表的学术论文

- [1] Zhang Z C, Zhang M Y, Zhou T, Qiu Y L. Pre-trained language model augmented adversarial training network for Chinese clinical event detection[J]. Mathematical Biosciences and Engineering, 2020, 17(4): 2825-2841. (SCI)
- [2] 张志昌, 周侗, 张瑞芳, 张敏钰. 融合双向 GRU 与注意力机制的医疗实体关系识别[J]. 计算机工程, 2020, 46(06): 296-302. (EI)
- [3] Zhang Z C, Qiu Y L, Yang X L, Zhang M Y. Enhanced character-level deep convolutional neural networks for cardiovascular disease prediction[J]. BMC Medical Informatics and Decision Making, 2020, 20(3): 1-10. (SCI)
- [4] Zhang Z C, Liu D, Zhang M Y, Qin X H. Combining Data Augmentation and Domain Information with TENER Model for Clinical Event Detection[J]. BMC Medical Informatics and Decision Making, 2021. (在投)

3. 在学期间取得的研究成果

- [1] 西北师范大学, 张志昌, 张敏钰等. 中考招生录取系统 V1.0[CP/CD]. 著作权登记号: 2019SR0510120, 2019.5.
- [2] 张志昌(主持), 张敏钰(参与). 企业在线交互式智能客户服务系统研发, 甘肃省重点研发计划项目. 课题编号: 17YF1GA016.

- [3] 张志昌（主持），张敏钰（参与）. 基于深度学习的中文文本蕴含关系识别技术研究，国家自然科学基金地区科学基金项目. 基金编号：61762081.