中医临床切诊信息抽取与词法分析语料构建及联合建模方法

王亚强[1,2,3†]，蒋文[1,2,3]，蒋永光[4]，舒红平[1,3]
[1]成都信息工程大学软件工程学院
[2]成都信息工程大学数据科学与工程研究所
[3]软件自动生成与智能服务四川省重点实验室
[4]成都中医药大学基础医学院
[†]通讯作者：yaqwang@cuit.edu.cn

摘要

切诊是中医临床四诊方法中极具中医特色的疾病诊察方法，为中医临床辨证论治提供重要的依据，中医临床切诊信息抽取与词法分析研究具有重要的临床应用价值。本文首次开展了中医临床切诊信息抽取与词法分析语料构建及联合建模方法研究，以万余条中医临床记录为研究对象，提出了一种语料构建框架，分别制定了中医临床切诊信息抽取、中文分词和词性标注语料标注规范，形成了可支撑多任务联合建模的语料，语料最终的标注一致性达到0.94以上。基于同级多任务共享编码参数模型，探索了中医临床切诊信息抽取与词法分析联合建模方法，并验证了该方法的有效性。

关键词： 中医临床切诊信息；信息抽取；词法分析；语料构建方法；多任务学习

On Corpus Construction and Joint Modeling Method for Clinical Pulse Feeling and Palpation Information Extraction and Lexical Analysis of Traditional Chinese Medicine

Yaqiang Wang[1,2,3†], Wen Jiang[1,2,3], Yongguang Jiang[4], Hongping Shu[1,3]
[1]College of Software Engineering, Chengdu University of Information Technology
[2]Institute for Data Science and Engineering, Chengdu University of Information Technology
[3]Sichuan Key Laboratory of Software Automatic Generation and Intelligent Service
[4]Department of Preclinical Medicine, Chengdu University of Traditional Chinese Medicine
[†]Corresponding author：yaqwang@cuit.edu.cn

Abstract

Pulse feeling and palpation (PFP) are the most distinctive and representative clinical diagnostic methods of traditional Chinese medicine (TCM). They provide important evidence for TCM clinical practitioners to differentiate syndromes. Extracting PFP information and analyzing its lexical features have important clinical value. In this paper, we carried out research on corpus construction and joint modeling method for PFP information extraction (IE) and lexical analysis (LA) of TCM for the first time. Based on more than ten thousand TCM clinical records, we proposed a corpus construction framework, built annotation guidelines and constructed a labeled PFP IE and LA corpus to support multi-task learning. Labeling consistency evaluated by inter-annotator agreement value for the corpora achieve more than 0.94. Moreover, we attempted to jointly model PFP IE and LA tasks of TCM with a same-level-shared multi-task model, and experimental results verified effectiveness of the joint model.

Keywords: Clinical pulse feeling and palpation information of traditional Chinese medicine , Information extraction , Lexical analysis , Corpus construction method , Multi-task learning

1 引言

辨证论治是中医学认知和治疗疾病的基本原则，四诊合参是中医辨证的基本需要(李灿东, 2021)。切诊是中医临床四诊方法[1]中极具中医特色的疾病诊察方法，是中医专家经过长期的临床实践，逐步形成并不断补充完善建立起的诊察技术，为中医临床辨证提供重要的诊断信息(谭同来, 2010)。

切诊包括按诊和脉诊两种方法。其中，按诊是中医专家用手对患者体表特定部位进行触、摸、按、叩，通过观察患者的反应，探明疾病的部位、性质和程度的方法；脉诊是中医专家运用手指切按患者的脉搏动处，体验脉动应指的形象，以确定全身脏腑功能、气血、阴阳的协调能力状况综合信息的方法(谭同来, 2010)。它们的诊察结果通常记录在中医临床记录中。

抽取中医临床记录中的切诊描述，获取其中蕴含的浅层词法信息，将为中医临床辅助辨证、中医临床医案分析等下游任务提供丰富的医学语义信息。如图1所示，由方位词“左”、专有名词“脉”和形容词“细”所构成的脉诊描述，说明了患者的左心室收缩和舒张功能存在异常，推断患者可能“气血亏虚”(杨杰, 2006)。

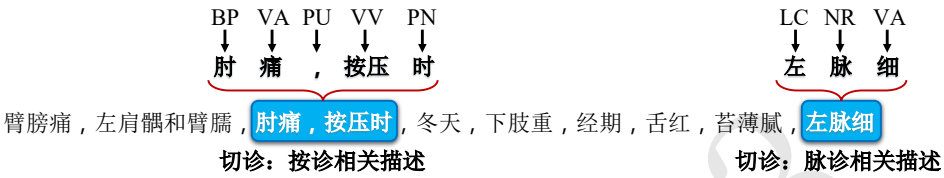

图 1. 中医临床记录中的切诊描述及其包含的浅层词法特征信息

中医临床记录中实体及其关系的信息抽取研究已广泛开展。Wang等人(2014)开展了基于统计序列标注模型的中医临床症状信息抽取研究。Ruan等人(2020)开展了基于深度学习的中医临床“病-证-药”实体关系抽取研究。此外，中医临床四诊信息抽取的研究尚处在起步阶段，2022年王亚强等人(2022)首次开展了中医临床记录四诊描述抽取研究。目前，中医临床切诊信息抽取与词法分析的研究尚未见相关报道。

中医临床切诊描述具有其领域特殊性和文本描述方式的独特性：

首先，切诊包括按诊和脉诊两种方法，两种方法的描述方式各有不同。按诊描述倾向叙述过程，用语与其它三诊描述类似，如图1中的“肘痛，按压时”，叙述了“按压过程患者肘部的疼痛状态”，且其中包含的词语均在其它三诊中常见。脉诊描述倾向描述位置和状态，如图1中的方位词“左”和形容词“细”，但其用语也有特殊性，如“脉”是脉诊描述中的专有名词。

其次，由于中医临床切诊描述的口语化特点，且通常采用简短的短语或短句描述，使其具有较强的稀疏性。如图1中所示，按诊描述“按压时肘痛”被口语化地记录为“肘痛，按压时”。此外，在本文构造的语料中，切诊信息简短，文本长度平均包含2.97个字[2]。

第三，中医临床记录具有简短性和独特的语言特色(Wang et al., 2012)，与一般领域相比，中医临床切诊描述中的词和词类定义存在不同。例如，简短的脉诊描述“脉滑”表达了“脉”往来流利，应指圆“滑”的脉象状态，其中包含“脉”和“滑”两个单字词，并且“脉”是脉诊描述中的专有名词，“滑”被用作形容词，这些词法定义在一般领域中并不常见。

最后，中医临床切诊信息抽取与词法分析任务存在类别分布不均衡的问题。在本文构造的语料中，中医临床记录的平均长度为38.90个字，而中医临床切诊描述包含的字数平均占比仅为9.30%。此外，中医临床切诊描述中长度大于2的词远少于词长为1的词数（如图2所示）。并且如图3所示，词类标签的计数分布同样存在长尾现象。

这些中医领域的特殊性和中医临床切诊描述的独特性，给中医临床切诊信息抽取与词法分析带来巨大挑战。因此，本文首次开展了中医临床切诊信息抽取与词法分析的研究，主要贡献包含以下三个方面：

1. 以“宾州中文树库分词指南”(Xia, 2000a)为基础，根据中医领域的特殊性和中医临床记录描述的独特性，建立了“中医临床切诊描述的中文分词指南”。此外，通过融合“信息处理用现

[1]中医临床四诊方法包括“望诊”、“闻诊”、“问诊”和“切诊”四种方法(李灿东, 2021)
[2]本文将中医临床记录中的标点符号也视为“字”

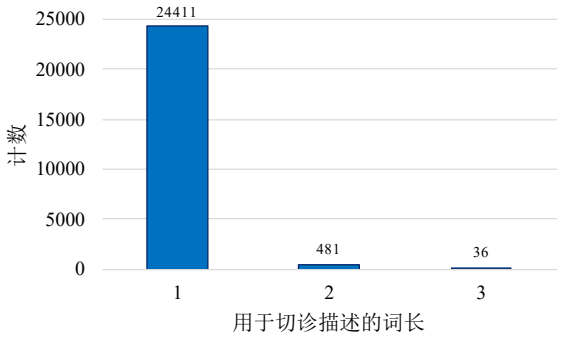

图 2. 中医临床切诊描述的不同词长计数

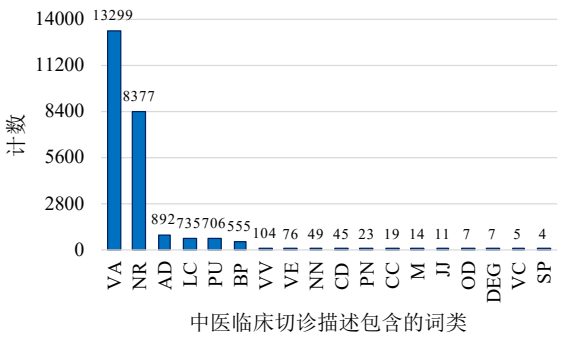

图 3. 中医临床切诊描述包含的词类计数

代汉语词类标记规范”[3]、“中文电子病历词性标注规范”(杨锦锋et al., 2016)等规范，结合中医领域知识，建立了“中医临床切诊描述的词类标注指南”。

2. 提出了一种语料构建框架，以贡献1中的两份“指南”为基础，聘请专家对前期收集的10594条中医临床记录进行了切诊描述标注、中文分词标注和词性标注，构建了可用于中医临床切诊信息抽取和浅层词法分析的多任务联合抽取语料。首次开展了面向中医临床切诊描述的中文词法分析研究，为中医临床切诊描述的语法分析以及面向中医临床记录的深层语义分析奠定良好基础。

3. 基于构建的中医临床切诊信息抽取和词法分析语料，本文将中医临床切诊信息抽取、中文分词和词性标注采用同级多任务共享编码参数方式进行多任务学习建模，开展了中医临床切诊信息抽取、中文分词和词性标注联合信息抽取研究初探，验证了多任务学习方法在中医临床切诊信息抽取、中文分词和词性标注联合信息抽取方面的有效性，发现了新问题，为未来的进一步研究指明了方向。

实验结果表明，采用本文提出的标准化语料构建框架（参见第3章描述）分别构建中医临床切诊信息抽取、中文分词和词性标注语料，均能使标注专家对标注任务的理解快速达到一致，各语料最终的标注一致性IAA（Inter-Annotator Agreement(Ron and Massimo, 2008)）值均达到0.94以上（为强一致性标注结果(Fleiss, 1981)）。该结果说明，本文设计的标注指南正确，语料构建质量高。此外，基于同级多任务共享编码参数方法构建中医临床切诊信息抽取、中文分词和词性标注联合抽取模型，能提升各任务的抽取性能，F值最高提升了2.2%（在最复杂的中医临床切诊词性标注任务上取得），说明了多任务联合建模方法是未来开展中医临床切诊信息抽取、中文分词和词性标注研究的重要方向。

2 相关工作

2.1 中医临床信息抽取

近年来，中医临床信息抽取研究受到广泛关注。Zhang等人(2022)对2010年至2021年间，中医文本信息抽取任务进行了综述，中医临床信息抽取是其中最重要的任务之一。中医临床信息抽取主要围绕中医临床记录中的实体（如疾病、症状、体征等(Liu et al., 2015; Guan et al., 2021)）信息抽取任务展开，少量研究关注实体间的关系抽取任务(Bai et al., 2022)。针对中医临床记录所包含的语言学信息（如词法信息(Jiao et al., 2018)）和临床语义信息（如中医临床四诊信息(李灿东, 2021)）抽取的相关研究较少。近期，王亚强等人(2022)首次开展了中医临床记录四诊描述抽取的研究。本文在此基础上，进一步围绕极具中医特色的中医临床切诊信息抽取与词法分析展开探索。

面向中医临床记录的词法分析研究尚未见报道，更不用说针对中医临床切诊描述的词法分析研究。通用领域的研究发现，词法分析的结果可作为补充信息，为下游任务提供丰富的语言学特征(Sagot and Alonso , 2017)，并有助于深层语义分析研究的开展(Guo et al., 2016; Kurita et al., 2017)。然而，由于中医领域的特殊性和中医临床记录描述的独特性，直接应用已有的

[3]URL: http://www.moe.gov.cn/ewebeditor/uploadfile/2015/01/13/20150113085826365.pdf

通用词法分析模型(Jiao et al., 2018; Sun et al., 2009)无法有效开展中医临床切诊描述的词法分析。因此，本文围绕中医临床切诊描述的词法分析进行了初探研究。

2.2 中医临床语料构建

中医临床信息抽取通常采用有监督学习方法实现(Zhang et al., 2022)，中医临床切诊信息抽取与词法分析研究需要带标注的语料库支撑。中医临床四诊信息抽取的研究处于起步阶段，支撑中医临床切诊信息抽取研究的语料初具规模(王亚强et al., 2022)。而中医临床切诊描述的词法分析研究尚未开展，由于中医领域的特殊性和中医临床记录描述的独特性，迫切需要利用中医临床记录构建语料以支撑相关研究。本文在前期构建的中医临床四诊信息抽取语料(王亚强et al., 2022)的基础上，构建了中医临床切诊信息抽取与词法分析语料。

语料构建耗时、费力，中医临床切诊信息抽取与词法分析语料的构建还需要中医专家的参与，这也为该任务提出了挑战。Zhang和Wang等人(2020)提出了一种面向中医临床信息抽取的语料构建框架，能有效提升构建的效率和质量。因此，本文针对中医临床切诊信息抽取与词法分析任务，改进了该语料构建框架，与中医专家共同探讨并制定了各任务的语料标注规范，构建了中医临床切诊信息抽取与词法分析语料，实验结果表明，本文所构建的语料质量高。

2.3 多任务联合信息抽取

多任务学习是一种通过相关任务之间共享表示，以提升原始任务模型泛化能力与性能的方法(Crawshaw, 2020)。中医临床切诊信息抽取、中文分词和词性标注是三项典型的相关任务，均可以采用序列标注方法建模(王亚强et al., 2022; Jiao et al., 2018)，三项任务之间具有共同的输入（即中医临床记录），可以共享输入的嵌入表示和中间层编码表示信息，最后根据任务的不同预测序列标注输出。因此，中医临床切诊信息抽取与词法分析可以采用基于多任务学习的序列标注方法(Rei, 2017; Lin et al., 2018; Fang et al., 2023)联合建模。本文将中医临床切诊信息抽取、中文分词和词性标注三项任务视为同级任务，借鉴Pham等人(2019)提出的“同级任务共享模型”对三项任务联合建模，实现三项任务通过共享表示的方式，促进各任务对应的序列标注模型的泛化能力和标注性能的提升。

3 语料构建方法

中医临床切诊信息抽取与词法分析采用有监督序列标注方法实现，有监督模型需要高质量的带标注语料的支撑。而中医临床切诊信息抽取与词法分析的研究尚处于起步阶段，相关研究所需的高质量的带标注语料尚有待构建。因此，本文基于前期研究工作所构建的中医临床四诊信息抽取语料，进一步的构建中医临床切诊信息抽取与词法分析语料。

3.1 构建方法框架

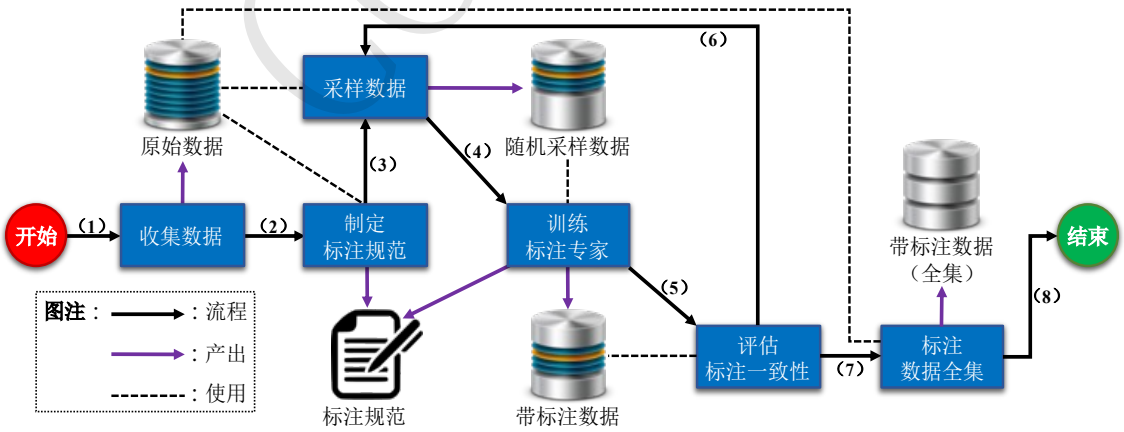

图 4. 中医临床记录切诊信息抽取与词法分析语料构建方法框架

本文借鉴Zhang和Wang等人(2020)提出的语料构建框架，分别用于中医临床切诊信息抽取、中文分词和词性标注语料的构建过程，构建方法框架如图4所示。框架中包含六个关键步骤，包括:

(1) 收集数据：根据任务需要，从中医专家日常诊疗过程中收集的中医临床记录，随机的选取指定数量的临床记录，形成原始数据。本文构造语料的10594条中医临床记录即从中医专家日常诊疗过程中收集的数据随机抽取得到。

(2) 制定标注规范：根据任务需要，以原始数据为研究对象，标注专家共同探讨，制定新的标注规范，或者借鉴并改进已有的同类任务的标注规范，制定适用于本任务的标注规范。本文采用了后者方式，分别制定了中医临床切诊信息标注规范、中文分词标注规范和词性标注规范。通常，在本步骤形成的标注规范为初稿，后期会随着对任务和数据的深入理解，对标注规范进行迭代更新。

(3) 采样数据：从原始数据中随机采样N条中医临床记录数据用于后续步骤(4)训练标注专家。通常N为一个较小的自然数，本文中$N = 100$。如需重复采样数据来反复训练标注专家，一般采用无放回采样方法，以保证未来训练时可遇见相似但不相同的待标注数据。

(4) 训练标注专家：参与制定标注规范的三位标注专家，根据当前任务的标注规范，独立地对相同的无标注随机采样数据进行标注，得到带标注数据。此外，在标注的过程中，标注专家如发现新的标注规则，将记录并在任务结束后与其他专家讨论，确认后更新标注规范。

(5) 评估标注一致性：采用IAA度量各标注专家的标注一致性，其结果用于评价专家在当前任务上的认知一致性。当IAA值小于阈值α时，将重复步骤(3)至(5)。当IAA值连续三次大于α时，认为各专家在未来独立的标注任务中，能够以相同的标准，保质地完成数据标注任务，则进入最后的第(6)步骤。对IAA的解释及本文设置参见章节3.3。

(6) 标注数据全集：各专家在达到对标注任务的认知一致后，原始数据将平均分配给各标注专家完成数据标注任务。一般地，各份数据会包含N'条相同的数据，用于计算在构建最终的带标注数据全集时，各专家的标注一致性，以确保最终的语料的质量。通常N'也是一个较小的自然数，本文中$N' = 100$。

3.2 标注规范

本文以中医临床切诊描述为案例，融合、重写和扩展同类任务的标注规范，并在语料构建方法框架中迭代完善所形成的中医临床切诊信息标注规范、中文分词标注规范以及词性标注规范。各规范主要的扩展原则如后文所述，标注规则具体内容参见(词法分析相关规范, 2023)。

3.2.1 切诊信息标注规范

本文沿用了前序工作(王亚强et al., 2022)中制定的中医临床四诊信息标注规范及其数据，构建形成了中医临床切诊信息抽取语料。在前序工作中，已形成中医临床四诊信息抽取语料，本文重点关注四诊中最具中医特色的切诊信息抽取，因此，在中医临床四诊信息抽取语料的基础上，本文统一将该语料中的其它三诊（即望诊、闻诊和问诊）的标注直接删除，最终形成中医临床切诊信息抽取语料。

3.2.2 中文分词标注规范

根据“中医四诊操作规范第4部分：切诊”国标(李灿东et al., 2021)对中医临床切诊相关概念的定义，三位标注专家融合通用领域的“信息处理用现代汉语分词规范”(靳光瑾et al., 2006)与“宾州中文树库分词指南”(Xia, 2000a)，以中医临床切诊描述为案例，重写并扩展制定了中医临床切诊描述的中文分词标注规范。

与通用领域的分词规则不同，中医临床切诊描述的中文分词规范以能够获取细粒度中医临床语义信息(Zhang et al., 2020)为成词的基本原则。因此，本文将通用领域中对人名、组织或国家名称等实体类描述的分词规则进行了修改，扩展制定了对中医切诊术语的细粒度分词规则。例如，以“脉细”为例，根据“宾州中文树库分词指南”的规则，“脉细”作为中医学实体不会被进一步分词，但为获取其中包含的细粒度中医学语义信息，“脉细”将进一步地切分为“脉”和“细”两个词语，进而为下游任务获取“脉”的状态为“细”这一细粒度中医临床修饰语义信息形成铺垫。

在中医临床切诊描述中，常见关于“大小”的描述内容，例如“周围硬块约5*5cm”，其中，“5*5”表达了中医专家通过中医按诊后，获得并记录的硬块大小信息。尽管“5”为数

词，“*”表示数学符号，在通用领域的规范中，它们均可以单独成词，并表示更细粒度的大小测算语义信息，但是为保留中医专家判断硬块大小的中医临床语义信息，将“大小”这类通常由几个通用领域的词语组成的字符串定义为词，不作细粒度切分。

3.2.3 词性标注规范

根据“中医四诊操作规范第4部分：切诊”国标(李灿东et al., 2021)对中医临床切诊相关概念的定义，三位标注专家融合通用领域的“信息处理用现代汉语词类标记规范”(靳光瑾et al., 2006)和“宾州中文树库词性标注指南”(Xia, 2000b)，以中医临床切诊描述为案例，重写并扩展制定了中医临床切诊描述的词性标注规范。

本文以语法功能作为划分词类的主要依据，并在标注的过程中，结合上下文语义综合判别词类。例如，“停”常用含义为“停止、停下”，表示“动作、行为和心理状态”，因此为动词。然而，在中医临床记录“7次1停”中，结合上下文医学语义信息，判定“停”在此处的语法功能为“动词量词”，因此被归为“量词”词类。

根据中医临床切诊描述中包含的词语特点，将专有名词等17种通用领域常见词类（表1词性标注的标签名称）保留在中医临床切诊描述的词性标注规范中。此外，单独新增了“身体部位”词类，具有表达“切按部位”以及发现“病症的部位”的中医临床语义信息，与通用领域的名词加以区分的目的。例如，在切诊描述“手热额冷”中，身体部位“手”和“额”分别体现了患者“热”和“冷”病症部位，因此需要将这类具有中医临床语义信息的词与其他通用领域的名词加以区分。

3.3 标注一致性度量方法

在构建方法框架中，IAA具有重要作用。一方面，IAA值用于衡量多位标注专家在进行相同标注任务时，对该任务认知是否达到一致。另一方面，IAA值也间接的用于评价所制定的标注规范的质量，即用多位标注专家在相同的标注规范指导下是否能够保持对相同任务的认知一致，来说明标注规范的全面性、代表性和准确性。

本文采用了Fleiss' Kappa方法来计算IAA值，其具体计算方式参见(Fleiss, 1971)。IAA值越大，说明多位标注专家在完成相同标注任务时，对该任务的认知歧义越低，即认知一致性越强，说明未来多位标注专家在完成该类任务时，能够做出一致的判断，从而保证标注的质量。反之亦然。通常，$IAA \geq 0.75$表示一致性优秀(Fleiss, 1981)。本文在构建方法框架中，将IAA的达标值α设置为0.85。并且，为避免偶然性，本文在构建方法框架中的步骤(5)中要求IAA值连续三次大于α，才能进入步骤(6)。

4 多任务联合信息抽取框架

中医临床切诊信息抽取、中文分词和词性标注是三项相关任务，它们在对输入数据的编码层面具有可共享的信息，解码后获得不同任务各自的输出结果。这是典型的多任务学习应用场景，通过相关任务之间的信息共享，提升原始任务模型泛化能力与性能(Crawshaw, 2020)。作为初探，本文自然地将Pham等人(2019)提出的“同级任务共享模型”（Same-level-Shared Model）应用到中医临床切诊信息抽取、中文分词和词性标注三项任务的联合信息抽取建模，模型框架如图5所示。

中医临床切诊信息抽取、中文分词和词性标注的多任务联合信息抽取可自然地基于序列标注方法建模。如图5所示，给定输入中医临床记录字序列$\boldsymbol{w} = \{w_1, w_2, \cdots, w_n\}$和输出标注序列$\boldsymbol{l}_j = \{l_{j,1}, l_{j,2}, \cdots, l_{j,n}\}$。其中，$n$为$\boldsymbol{w}$包含字$w_i$的数量及标签序列$\boldsymbol{l}_j$的长度，$j \in (1, T)$，$T$为多任务联合信息抽取框架中任务的数量，通常大于2，本文中$T = 3$。$w_i \in V$，V为训练数据中字符的集合，$l_{j,k} \in L_j$，L_j表示任务j的标签集。在本文实验数据中，字符包括中文字、标点、数字，中医临床切诊信息抽取（$j = 1$）、中文分词（$j = 2$）和词性标注（$j = 3$）任务的标签集如表1所示。

在联合信息抽取框架中，主要包括两部分。一是编码层，实现多任务共享的中医临床记录输入的向量表示（字的$\boldsymbol{w}_i$的独热表示$\boldsymbol{x}_i$和嵌入$\boldsymbol{e}_i$）变换和中间信息融合编码（$\boldsymbol{h}_i$）学习。二是解码层，根据任务的不同，综合利用共享的编码信息$\boldsymbol{h}_i$，通过线性变换形成序列标注推理模型的输入，最终输出预测的序列标注结果。

本文中，中医临床记录输入的嵌入表示变换采用了通用的BERT(Devlin et al., 2018)实现；为实现上下文信息融合，中间信息融合编码采用了BiLSTM(Shin and Lee, 2019)；各任务的序

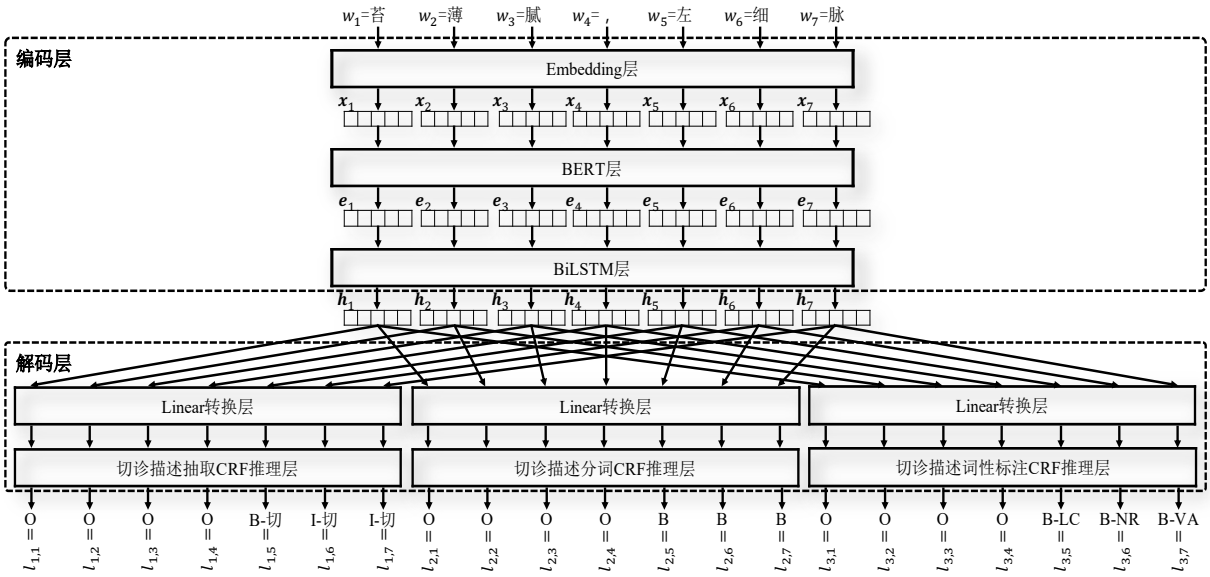

图 5. 中医临床切诊信息抽取、中文分词和词性标注联合信息抽取框架

任务名称	标签数量	标签名称(类型)
信息抽取	3	切诊描述开始位置(B-切), 切诊描述中间位置(I-切), 其他位置(O)
中文分词	3	词语开始位置(B), 词语中间位置(I), 其他位置(O)
词性标注	37	专有名词(NR), 普通名词(NN), 能愿动词(VV), 有字动词(VE), 系动词(VC), 表语形容词(VA), 名词修饰语(JJ), 基数词(CD), 序数词(OD), 量词(M), 副词(AD), 代词(PN), 方位词(LC), 身体部位(BP), 并列连词(CC), 属格标记(DEG), 句末助词(SP), 标点符号(PU), 其他(O) （**说明**：每种词类有“B”和“I”两种标签，分别表示该词类标签的开始和中间位置）

表 1. 中医临床切诊信息抽取、中文分词和词性标注任务的序列标注标签集定义

列标注结果预测采用了CRF（Conditional Random Fields(Ma and Hovy, 2016)）模型实现。

多任务学习优化的目标是最小化多个任务的加权损失函数，同时保证模型参数在多个任务中具有共享性。因此，多任务优化的目标函数为：

$$L_{global} = \arg\min_{\hat{\boldsymbol{\theta}}_1,\cdots,\hat{\boldsymbol{\theta}}_T} \sum_{j=1}^{T} \beta_j \cdot L_j(\boldsymbol{\theta}_j)$$

其中，β_j为超参数，表示任务j的损失E_j在多任务条件下的全局损失L_{global}中的贡献度，$\sum_{j=1}^{T} \beta_j = 1$。

5 实验

5.1 实验数据与模型设置

本文所构建的中医临床切诊信息抽取与词法分析语料是基于中医专家日常诊疗过程中收集的中医临床记录完成，共有10594条数据，包含了412123个字，字典大小为2453。其中，最短和最长的中医临床记录分别包含2个和311个字。中医临床记录的长度及该长度出现的频数关系如图6所示。从图中可以看出，大部分中医临床记录的长度集中在11个字至62个字之间，属于短文本。此外，中医临床记录中，片段描述之间通常采用逗号分隔，简单地以中文逗号和英文逗号为分隔统计，共包含片段描述91023条，片段描述的长度及该长度出现的频数关系如图7所示。从图中可以看出，大部分中医临床记录的片段描述的长度集中在1个字至9个字之间，属于短文本。这些特点进一步使得中医临床切诊信息抽取及词法分析联合信息抽取任务具有挑战。

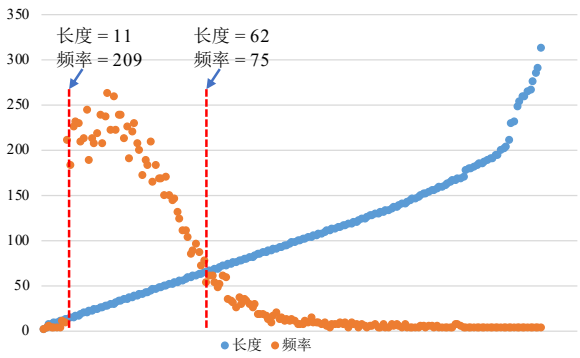

图 6. 中医临床记录的长度及该长度出现的频数关系

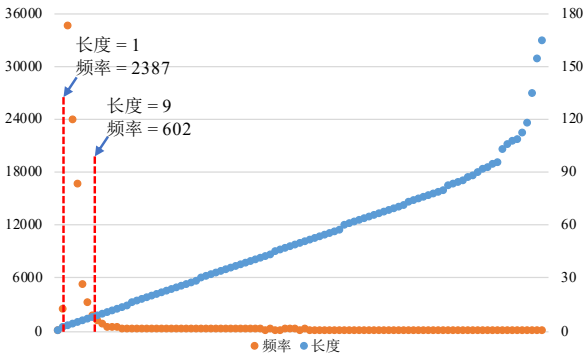

图 7. 中医临床记录中片段描述的长度及该长度出现的频数关系

本文的多任务联合信息抽取框架是基于BERT+BiLSTM+CRF改造，如图5所示，负责编码的BERT层和BiLSTM层为参数共享部分，CRF推理层分别执行各标注任务。框架的训练采用AdamW优化器，其超参数β_1和β_2分别设为0.9和0.999，为避免过拟合，Dropout设置为0.1，批量大小设置为32，Epochs设置为100，学习率设置为5e-5，最大句子长度设置为256，中间层嵌入表示向量$\boldsymbol{e}_i$和$\boldsymbol{h}_i$的大小设置为128。

5.2 语料构建效率与质量

本文基于图4所描述的中医临床切诊信息抽取与词法分析语料构建方法框架完成了中医临床切诊信息抽取、中文分词和词性标注语料三个语料的构建。如前文所述，中医临床切诊信息抽取语料基于前序工作构造的语料得到（具体方法参见3.2.1），该语料也采用图4所述框架构造得到，语料构建质量（即IAA值的相关信息参见(王亚强et al., 2022)）。

基于图4所描述的中医临床切诊信息抽取与词法分析语料构建方法框架，三位标注专家经过了三轮的标注训练，即达到对中医临床切诊描述的中文分词和词性标注两项标注任务的认知一致性，三轮标注的IAA值结果如表2所示。

	中文分词标注任务的IAA值	词性标注任务的IAA值
第一轮训练	0.9281	0.8543
第二轮训练	0.9564	0.9361
第三轮训练	0.9523	0.9381
最终标注结果	0.9450	0.9470

表 2. 中文分词和词性标注任务上的标注训练及最终标注的IAA值

从表2的结果可以看出，本文提出的中医临床切诊信息抽取与词法分析语料构建方法框架具有良好的语料构建效率，可以高效地完成三份语料的构建，经过三轮的标注训练，即能达到对标注任务的认知一致。

此外，从表2的结果还可以看出，三位标注专家在每项任务上均可以在第一轮就达到对标注任务的较高认知一致性。在中文分词标注任务中，三位第一轮训练后的IAA值达到0.9281。在词性标注任务中，三位第一轮训练后的IAA值达到0.8543。这些结果说明，本文制定的标注规范具有良好的代表性和可操作性。

进一步观察2的结果可以看出，中医临床切诊描述的词性标注任务的IAA值在不断提升，但在标注训练过程中，中医临床切诊描述的词性标注任务的IAA值始终低于中文分词标注任务的IAA值。同时观察表1可以发现，中医临床切诊描述的词性标注任务标签数量远高于其它两项任务。上述结果表明，中医临床切诊描述的词性标注任务相较于中医临床切诊信息标注任务和中文分词标注任务更有难度。

本文采用随机无放回采样策略构造每轮标注专家标注训练数据，以保证训练过程的客观性，从最终的标注结果来看，本文所构建的中医临床切诊信息抽取、中文分词和词性标注语料

质量高，各标注任务最终的IAA值均高于0.94，为强一致性标注结果(Fleiss, 1981)。从中文分词和词性标注任务训练过程获得的IAA结果来看，词性标注任务的IAA值始终在提升，说明本文制定的标注指南将词性标签给予了清晰的定义。比较特殊的是，中文分词任务的IAA值尽管第一轮训练时即达到0.9281，然而训练过程中的IAA值有波动上升的趋势。该结果说明，中医临床切诊描述的中文分词任务尽管相对简单，但由于中医临床记录描述存在不规范的现象，使其词语边界的界定常存在歧义性，指南需要持续更新。

5.3 语料特点分析

表3中分别统计了排名前十的中医临床切诊信息抽取语料中包含的频繁字和非频繁字，中医临床切诊描述的中文分词语料中包含的频繁字和非频繁字以及频繁词和非频繁词。基于表3，对比两份语料中的频繁字与非频繁字之间的差异可以明显看出，中医临床切诊描述具有中医特色，并且中医切诊在中医临床辨证论治过程中具有重要意义，“脉”、“细”、“略”等具有中医切诊特色的字频繁出现在中医临床记录中。

中医临床切诊信息抽取语料				中医临床切诊描述的中文分词语料							
频繁字	计数	非频繁字	计数	频繁字	计数	非频繁字	计数	频繁词	计数	非频繁词	计数
痛	9740	腴	1	脉	7915	眉	1	脉	7915	石	1
脉	8052	悉	1	细	4486	此	1	细	4486	皆有	1
苔	6647	络	1	弦	2093	人	1	弦	2093	一直	1
舌	6410	哽	1	弱	1123	到	1	弱	1123	好	1
薄	4726	绳	1	滑	945	皮	1	滑	945	眉	1
便	4658	死	1	沉	940	光	1	沉	940	裂纹	1
黄	4631	割	1	数	889	六	1	数	889	耳门	1
细	4564	呵	1	略	598	柔	1	略	598	下	1
干	4424	屋	1	软	585	时	1	软	585	裂	1
略	4200	回	1	平	532	或	1	平	532	凹陷性	1

表 3. 中医临床切诊信息抽取语料和中文分词语料中排名前十的频繁与非频繁字和词的统计数据

此外，从表3中“脉”、“细”、“略”等具有中医切诊特色的字的频数，在中医临床切诊描述的中文分词语料和中医临床切诊信息抽取语料中的对比发现，频数并不相等。该结果说明，中医临床切诊描述中的字存在歧义性，这为中医临床切诊信息抽取、中文分词以及词性标注任务均带来一定的挑战。

对比表3中，中医临床切诊描述的中文分词语料中的频繁字和非频繁字，以及频繁词和非频繁词可以发现，中医临床切诊描述单字成词的现象较严重，这与中医临床记录具有简短性的特点(Wang et al., 2012)有关，这一特点使得单字词发生兼类现象(陆俭明, 1994)的风险增加，一定程度上会对中医临床切诊描述的词性标注任务形成挑战。

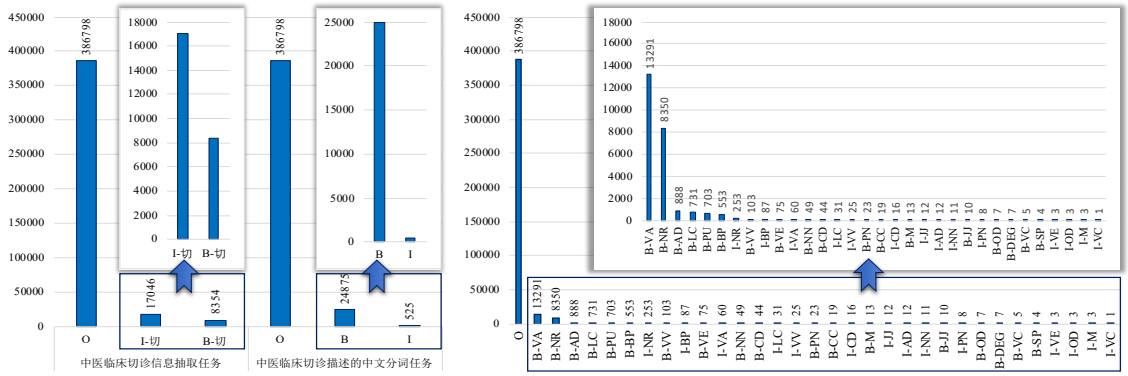

图 8. 中医临床切诊信息抽取、中文分词和词性标注语料标签分布

中医临床切诊信息抽取、中文分词和词性标注任务的标签分布如图8所示，从图中可以清晰的观察到，三项任务均面临严重的标签分布有偏问题，其中其他类标签“O”最多。并且，即使将其他类标签“O”删除，三项任务依然面临标签分布有偏的问题，长尾分布始终保持，该问题将对基于序列标注的多任务联合信息抽取带来挑战。

5.4 多任务联合信息抽取

本文评价多任务联合信息抽取结果分别采用了对象级（即切诊描述片段对象、词语对象、词性对象的整体）以及标签级的准确率（P）、召回率（R）和F度量值（F）的宏平均结果性。两类P、R和F的宏平均计算方法参见(Wang et al., 2014)，实验结果如表4所示。

	任务损失贡献度	切诊抽取任务			中文分词任务			词性标注任务		
	$(\beta_1, \beta_2, \beta_3)$[1]	P	R	F	P	R	F	P	R	F
对象级	(1, 0, 0)	93.94	93.18	93.55	-	-	-	-	-	-
	(0, 1, 0)	-	-	-	80.24	**71.01**	**74.6**	-	-	-
	(0, 0, 1)	-	-	-	-	-	-	45.42	38.18	40.33
	(1/3, 1/3, 1/3)	**94.00**	93.02	93.51	79.80	68.97	73.19	43.62	35.72	37.73
	(0.2, 0.3, 0.5)	93.98	**93.24**	**93.58**	79.80	69.86	73.95	**49.28**	**40.20**	**42.80**
	(0.1, 0.2, 0.7)	93.96	92.90	93.43	**80.25**	68.39	72.97	47.58	38.95	41.58
标签级	(1, 0, 0)	96.85	95.31	96.06	-	-	-	-	-	-
	(0, 1, 0)	-	-	-	86.19	**80.81**	**82.96**	-	-	-
	(0, 0, 1)	-	-	-	-	-	-	50.81	43.03	45.21
	(1/3, 1/3, 1/3)	**96.87**	95.25	96.04	**86.25**	79.28	82.04	47.58	39.72	41.38
	(0.2, 0.3, 0.5)	96.67	**95.50**	**96.08**	85.74	79.88	82.35	**53.19**	**45.53**	**47.41**
	(0.1, 0.2, 0.7)	96.83	95.20	96.00	85.97	78.89	81.70	52.97	42.50	45.18

[1] $(\beta_1, \beta_2, \beta_3)$被设置为(1, 0, 0)、(0, 1, 0)和(0, 0, 1)时，多任务联合信息抽取框架退化为基于BERT+BiLSTM+CRF的单任务的信息抽取模型，“1”所在维度为当前执行的抽取任务。

表 4. 多任务联合信息抽取结果

如表4所示，相对于单任务的信息抽取模型，无论是对象级还是标签级，采用多任务联合信息抽取方法，在不同的$(\beta_1, \beta_2, \beta_3)$设置下，总能取得$P$、$R$或$F$的结果提升。特别是在复杂的中医临床切诊描述的词性标注任务上，效果提升明显，在$(\beta_1, \beta_2, \beta_3)$被设置为(0.2, 0.3, 0.5)时，$F$值在对象级和标签级分别提升了2.47%和2.2%。

从表4可以看出，多任务联合信息抽取方法在中医临床切诊描述的中文分词任务上效果不明显，仅P值结果有少量提升。此外，中医临床切诊描述的词性标注任务的总体性能不高，还有巨大的提升空间，这与其任务复杂性有关，标签数量多，且中医临床记录描述有其独特性。如何有效提升多任务联合信息抽取的性能是未来有待深入研究的方向。

6 总结

切诊极具中医特色，中医临床切诊信息抽取与词法分析为基于中医临床记录的辨证论治上下游任务研究提供丰富的中医临床医学语义信息。本文首次开展了中医临床切诊信息抽取与词法分析研究，围绕中医临床切诊信息抽取、中文分词和词性标注任务，构建了万余条带标注的多任务高质量语料，为中医临床记录的自然语言处理和文本挖掘开辟了新方向。本文基于同级多任务共享编码参数的多任务学习框架，开展了中医临床切诊信息抽取、中文分词和词性标注联合信息抽取的研究初探，形成了基线结果，并发现了系列问题，为后续研究指明了方向。

参考文献

Tian Bai, Haotian Guan, Shang Wang, Ye Wang, and Lan Huang. 2022. Traditional Chinese medicine entity relation extraction based on CNN with segment attention. *Neural Computing and Applications*,34: 2739-2748.

Michael Crawshaw. 2020. Multi-Task Learning with Deep Neural Networks: A Survey. *arXiv preprint arXiv:2009.09796*.

Jacob Devlin, MingWei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378.

J. L. Fleiss 1981. Statistical methods for rates and proportions. 2nd ed. *John Wiley and Sons*, IBSN: 978-0-471-26370-8.

Qin Fang, Yane Li, Hailin Feng, and Yaoping Ruan. 2023. Chinese Named Entity Recognition Model Based on Multi-Task Learning. *Applied Sciences*, 13(8): 4770.

Yu Guan, Huan Li, and Wenjing Xu. 2021. A Traditional Chinese Medicine Terminology Recognition Model Based on Deep Learning: A TCM Terminology Recognition Model. *Proceedings of the 6th International Conference on Big Data and Computing*, 15-20.

Zhen Guo, Yujie Zhang, Chen Su, Jinan Xu, and Hitoshi Isahara. 2016. Character-Level Dependency Model for Joint Word Segmentation, POS Tagging, and Dependency Parsing in Chinese. *IEICE TRANSACTIONS on Information and Systems*, 99(1): 15-20.

Honglan Liu, Xiaona Qin, and Bin Fu. 2015. The Symptoms and Pathogenesis Entity Recognition of TCM Medical Records Based on CRF. *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, 1479-1484. IEEE.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 799-809.

Zhenyu Jiao, Shuqi Sun, and Ke Sun. 2018. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *arXiv preprint arXiv:1807.01882*.

Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. 2017. Neural Joint Model for Transition-based Chinese Syntactic Analysis. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1204-1214.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *arXiv preprint arXiv:1603.01354*.

Thai-Hoang Pham, Khai Mai, Nguyen Minh Trung, Nguyen Tuan Duc, Danushka Bolegala, Ryohei Sasano, and Satoshi Sekine. 2019. Multi-Task Learning with Contextualized Word Representations for Extented Named Entity Recognition. *arXiv preprint arXiv:1902.10118*.

Marek Rei. 2017. Semi-supervised Multitask Learning for Sequence Labeling. *arXiv preprint arXiv:1704.07156*.

Artstein Ron and Poesio Massimo. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational linguistics*, 34(4): 555-596.

Chunyang Ruan, Yingpei Wu, Guang Sheng Luo, Yun Yang, and Pingchuan Ma. 2020. Relation Extraction for Chinese Clinical Records Using Multi-View Graph Learning. *IEEE Access*, 8: 215613-245622.

Benoît Sagot and Héctor Martínez Alonso. 2017. Improving neural tagging with lexical information. *Proceedings of the 15th International Conference on Parsing Technologies*, 25-31, Pisa, Italy, September, Association for Computational Linguistics.

Xiao Sun, Degen Huang, and Fuji Ren. 2009. Chinese lexical analysis based on hybrid MMSM model. *International Journal of Innovative Computing, Information and Control*, 5(12 (A)).

Youhyun Shin and Sang-goo Lee. 2019. Learning Context Using Segment-Level LSTM for Neural Sequence Labeling. *IEEE/ACM Transactions on Audio, Speech, and language processing*, 28:105-115.

Yaqiang Wang, Zhonghua Yu, Yongguang Jiang, Yongchao Liu, Li Chen, and Yiguang Liu. 2012. A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records. *Journal of Biomedical Informatics*, 45(2):210-223.

Yaqiang Wang, Zhonghua Yu, Li Chen, Yunhui Chen, Yiguang Liu, Xiaoguang Hu, and Yongguang Jiang. 2014. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study. *Journal of Biomedical Informatics*, 47:91-104.

Fei Xia. 2000. The segmentation guidelines for the Penn Chinese Treebank (3.0).

Fei Xia. 2000. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). *IRCS Technical Reports Series*, 38.

Tingting Zhang, Yaqiang Wang, Xiaofeng Wang, Yafei Yang, and Ying Ye. 2020. Constructing fine-grained entity recognition corpora based on clinical records of traditional Chinese medicine. *BMC Medical Informatics and Decision Making*, 20(1):1-17.

Tingting Zhang, Zonghai Huang, Yaqiang Wang, Chuanbiao Wen, Yangzhi Peng, and Ying Ye. 2022. Information Extraction from the Text Data on Traditional Chinese Medicine: A Review on Tasks, Challenges, and Methods from 2010 to 2021. *Evidence-Based Complementary and Alternative Medicine*, 2022.

靳光瑾, 肖航, 郭曙伦, 富丽, 章云帆, 于桂英, 陈玉泉, 王立. 2006. 信息处理用现代汉语词类标记规范. 中华人民共和国国家质量监督检验检疫总局; 中国国家标准化管理委员会, GB/T 20532-2006.

李灿东. 2021. 中医诊断学. 中国中医药出版社

陆俭明. 1994. 关于词的兼类问题. 中国语文, 28-34.

李灿东, 陈研, 郭宇博, 苏祥飞, 王天芳, [朱文锋], 郑进, 顾星, 王洋, 林雪娟, 甘慧娟, 李宇涛. 2021. 中医四诊操作规范　第4部分：切诊. 国家市场监督管理总局; 国家标准化管理委员会, GB/T 40665.4-2021.

谭同来. 2010. 中华医学切诊大全. 山西科学技术出版社.

王亚强, 李凯伦, 蒋永光, 舒红平. 2022. 基于批数据过采样的中医临床记录四诊描述抽取方法. 第*21*届全国计算语言学大会论文集, 611-622.

杨杰. 2006. 基于脉动信息获取的中医脉诊数字化、可视化探讨. 北京中医药大学

杨锦锋, 关毅, 何彬, 曲春燕, 于秋滨, 刘雅欣, 赵永杰. 2016. 中文电子病历命名实体和实体关系语料库构建. 软件学报, 27(11): 2725-2746.

中医临床切诊描述词法分析相关规范. 2023. *https://github.com/xx-Jiangwen/Guideline.*