分类号	密级		
UDC	编号		

# 中国科学院自动化研究所 博士后研究工作报告

面向对话文本的自然语言理解——在医疗领域的实践

张元哲

工作完成日期 2018年8月— 2020年6月

报告提交日期 2020 年 6 月

中国科学院自动化研究所

2020 年 6 月

# 面向对话文本的自然语言理解——在医疗领域的实践

## Natural Language Understanding towards Dialogue Texts: Practice on Medical Domain

博 士 后 姓 名 张元哲

流动站(一级学科)名称 控制科学与工程

专 业(二级学科)名称 模式识别与智能系统

研究工作起始时间 2018 年 8 月 3 日

研究工作期满时间 2020 年 6 月 20 日

中国科学院自动化研究所

2020年6月

#### 内容摘要

对话文本理解是自然语言处理研究领域的热点和难点问题。对于对话的理解具有重要的科学研究意义和实际应用价值。例如,在医疗、电商、司法领域都存在着大量的对话文本。与普通形式的文本相比,对话文本更加困难,主要体现在:1)口语化严重;2)对话具有交互性,说话人风格不统一;3)对话具有主题性和知识性,隐式信息更多,需要额外知识引导。

本文主要面向对话形式的文本,以医疗领域对话作为实践对象,研究 针对性的背景知识图谱构建、实体抽取、关系分类、实体链接和整体信息 抽取问题,期望以医疗领域的研究作为基础,扩展到更多领域,或者更进 一步,解决多领域的对话理解问题。

本文的主要贡献如下:

首先,探讨医疗知识图谱的构建方法,并提出一种构建症状知识图谱的方法,获得一个医疗症状图谱,作为后续工作的知识依托。

其次,提出面向对话文本的实体和关系标注数据集,医疗命名实体识别仍然采用基于序列标注的方法,明确实体的边界和类别标签,从而将口语化表达转化为规范表达;关系分类则完成了医学属性抽取的工作,采用了多种关系分类方法。此外,本章还尝试将对话文本中出现的实体链接到医疗知识图谱,从而帮助后续的自动诊断、病历质控等落地应用。

最后,提出一个面向整体对话的信息抽取方法,提出相关数据集标注方法,并提出基于深度匹配的信息抽取模型,获得对话内提及的重要医疗信息。实验结果表明本文提出的方法在窗口级别和对话级别上可以达到69.29 的 F 值,说明本方法有较好的研究前景。

综上,本文主要研究一系列针对医疗对话的文本理解方法,以识别实体、关系以及抽取信息为目标,通过在医疗领域的实践,提升针对对话文本的自然语言理解关键技术水平。

关键词:对话文本理解,自然语言处理,知识图谱,信息抽取,医疗对话文本

#### **Abstract**

Dialogue text understanding is a difficult research issue and a hotspot in natural language processing field. The understanding of dialogue has important scientific research significance and practical application value. For example, there are a lot of dialogue texts in the medical, e-commerce, and judicial fields. Compared with ordinary forms of text, the dialogue text is more difficult, mainly reflected in: 1) The dialogue is colloquial; 2) The dialogue is interactive and the speaker style is inconsistent; 3) The dialogue is thematic and knowledgeable, and contains implicit information which need additional knowledge to guide.

This article is mainly for dialogue-based texts, with dialogues in the medical field as practical objects. We research background medical knowledge graph construction, entity recognition, relation classification and overall information extraction. It is expected that the research in the medical field will be expanded to more fields, or further, solve the problem of dialogue understanding in multiple fields.

The main contributions of this article are as follows:

First, we discuss the construction method of medical knowledge, and propose a method of constructing a symptom knowledge graph, as the knowledge support of the follow-up work.

Secondly, we propose an entity and relationship labeling dataset on medical dialogue. Medical named entity recognition still uses sequence-based labeling methods to clarify the boundaries and category labels of entities, thereby converting colloquial expressions into normative expressions; relationship classification has completed the extraction of medical attributes, and uses a variety of relationship classification methods. In addition, this chapter also attempts to link the entities that appear in the dialogue text to the medical knowledge graph, thereby helping subsequent automatic diagnosis, medical record quality control and other applications.

Finally, an information extraction method for the whole dialogue is

proposed, a related dataset annotation method is proposed, and an information extraction model based on deep matching is proposed to obtain important medical information mentioned in the dialogue. Experimental results show that the method can reach an F-measure of 69.29 at the window level and dialogue level, indicating that this method has a good research prospect.

In summary, this article mainly studies a series of text understanding methods for medical dialogues, with the goal of identifying entities, relations and extracting information. Through practice in the medical field, the key techniques of natural language understanding of dialogue texts are improved.

Keywords: Dialogue Text Understanding, Natural Language Processing, Knowledge Graph, Information Extraction, Medical Dialogue Text

### 目录

1	7	概述		1
2	Ā	研究	现状与相关工作	7
	2.	1	医学知识图谱相关研究	7
	2. 2	2 3	对话形式文本理解的相关研究	9
	2. 3	3	医学信息抽取的研究	11
3	١	医疗	知识图谱构建	15
	3. 3	1 7	构建难点	15
	3. 2	2 7	构建方法	17
	•	3. 2. 3	1 知识图谱框架	17
	•	3. 2. 2	2 症状获取	18
	3. 3	3	症状图谱构建的自动方法	20
	•	3. 3. 3	1 同义词挖掘	20
	ć	3. 3. 2	2 上下位词挖掘	21
	ć	3. 3. 3	3 中医症状识别	21
	ć	3. 3. 4	4 症状成分切分	22
	3. 4	4	症状图谱的扩展	25
	3. 5	5	本章小结	27
4	[	医疗	实体与属性抽取	29
	4.	1 í	简单规则抽取	29
	2	4. 1. 1	1 数据集	29
	2	4. 1. 2	2 规则方法	30
	2	4. 1. 3	3 实验结果	30
	4. 2	2 5	实体抽取和关系分类	31
	4	4. 2.	1 数据集	31
	۷	4. 2. 2	2 医疗命名实体识别	34
	2	4. 2. 3	3 关系分类	37

	4.	3	实体链接	38
		4. 3.	1 数据集	38
		4. 3.	2 模型	39
		4. 3.	3 实验结果	39
	4.	4	本章小结	41
5		面向	医疗对话篇章的信息抽取	43
	5.	1	研究动机	43
	5.	2	方法	44
		5. 2.	1 数据集	44
		5. 2.	2 模型	46
	5.	3	实验	50
		5. 3.	1 实验设置	50
		5. 3.	2 基线模型	50
		5. 3.	3 实验结果	50
		5. 3.	4 实例分析	51
	5.	4	本章小结	52
6		总结	与展望	53
致	()	t		55
参	考	文献	<u>.</u>	57
个	人	简质	j	65
<b>.</b> .	<del>, \</del>	·4:±	· · · · · · · · · · · · · · · · · · ·	67

#### 1 概述

对话是人类使用语言交流最自然的形式,是两个人或更多人的交互沟通方式。人类通过对话形式进行交流比使用文字早得多,即使不认识自然语言文字的人,也能通过对话的形式互相交流。对话能够承载很多信息,也是人类传递知识的主要方式,孔子、苏格拉底等都通过对话的形式进行教学。针对对话本身的研究也得到了社会学家、心理学家以及语言学家的关注,这些研究多集中在对话的结构、目的以及社会作用等角度,取得了丰硕的成果[1,2]。随着计算机软硬件的飞速发展,人们对于人工智能有了越来越高的期望,而验证人工智能水平的图灵测试也是采用了对话测试的设定,即如果人类分辨不出与其对话的是人还是机器,则表示机器通过了图灵测试[3]。这足以说明对话对于人类交流的重要程度,以及使机器理解对话并获取对话能力在人工智能领域的难度。

随着互联网的出现和爆发式发展,对话的发生场景不再局限于人与人面对面交流这一种方式,开始逐渐延伸到互联网上,线上交流已经成为了人们主要的沟通方式。从最早的网络聊天室、电子邮件、论坛发帖讨论,到现在的微博及其互动回复、微信等即时通讯工具等,消除了距离带来的限制,依托于互联网的对话正在随时随地发生。中国互联网络信息中心在2020年4月发布的第45次《中国互联网络发展状况统计报告》显示,截至2020年3月,中国的即时通信用户规模已经达到8.96亿,占所有网民比例的99.2%,而且移动网民经常使用的各类APP中,即时通信类用户使用时间最长,占比14.8%。这些惊人的数据都说明,线上对话已经成为了人们日常生活的重要组成部分。在各个垂直细分领域,线上对话也有着广阔的应用,典型的有在线问诊、律师咨询、金融理财客服、网上政务咨询等等不一而足。随着这些线上交流渠道深入每个人的生活,随之产生的对话形式的网络文本数据总量更是呈爆炸式增长,其规模已经远远超过人类能够接收和处理的范畴。

处理对话形式的数据具有多方面的重大意义。从国家战略层面来说,对于对话文本数据的分析有助于掌握人民生活中关注的重点问题和发展 趋势,而且面向对话形式的舆情分析也是未来的研究重点。从行业发展层 面来说,各个垂直细分领域的线上对话都是第一手的用户需求,甚至在某些特定领域,如医疗、司法等,对话本身就是业务的第一个环节,问诊、辩护等都是以对话的形式进行的。对这些领域对话进行分析和理解有助于推进行业的整体发展。从个人生活层面来说,越来越多的事务需要人们在即时通讯工具上处理,人们已经很难从冗长的聊天记录中定位关键信息,已经有很多个人助理类软件能够从对话中提取出时间地点信息等用于提醒,但是其功能依然非常有限。

因此,如何利用机器来理解互联网上的海量对话形式的信息,是一个亟需解决的问题。目前,对于对话文本的研究主要可以分为三类: 1)语言学分析[1],即对话分析,从人类语言学的角度分析和理解对话; 2)对话系统<sup>[4,5]</sup>,即任务型对话和闲聊型对话。任务型对话可解释性稍强,但是应用场景有限。闲聊型对话的目的是让机器学会人类的对话方式,根据人的对话内容给出回应。3)面向对话的阅读理解。阅读理解<sup>[6,7]</sup>相关的工作则注重对于篇章的理解,针对其中的内容中进行提问,如果机器能够回答出相应的问题,则表明其在一定程度上理解了篇章。然而,现在只有很少的工作针对人人对话进行理解<sup>[8,9]</sup>,由于对话本身交互传递信息的特性,适用于普通篇章的阅读理解无法直接适用于对话。现有的两个工作都是关注医疗对话,这是因为医疗领域中,关键信息比较集中,容易针对这些信息提出问题,并且,医疗对话容易找出标准答案,评价相对容易。

本文也将从医疗领域展开研究,作为对于对话文本理解的领域实践。健康医疗是一项关系国计民生的重大事业,健康医疗信息化对于医疗事业现代化发展具有重要作用。近年来,随着全球范围内医院信息系统现代化的稳步推进,健康医疗数据的规模呈爆炸式增长。EMC<sup>[10]</sup>和 IDC<sup>[11]</sup>发布的报告显示,2013年全球健康医疗数据量为153EB,而预计的年增长率为惊人的48%,到2020年,全球的健康医疗数据将达到2.26ZB。中国作为人口大国,健康医疗一直是国家关注的重点,而且随着社会老龄化临近,健康医疗事业近年来受到了社会各界越来越多的关注。为了规范和加速大规模健康数据的生产和使用,2018年9月13日,国家卫生健康委印发了《国家健康医疗大数据标准、安全和服务管理办法(试行)》<sup>[12]</sup>,对医疗健康大数据行业从规范管理和开发利用的角度进行了指导。早在

2014年,国家卫计委就制定了"46312"工程,即建设国家级、省级、地级市、县级 4 级卫生信息平台,依托于电子健康档案和电子病历,支撑公共卫生、医疗服务、医疗保障、药品管理、计划生育、综合管理等 6 项业务应用,构建电子监控档案数据库、电子病历数据库、全员人口个案数据库 3 个数据库,建立一个安全的卫生网络,加强卫生标准体系和安全体系建设。

在互联网问诊中,患者通常需要向医生阐述自身的症状,而医生往往 也要通过多轮次的询问,获取诊断和给出建议所需的信息。在问诊对话中 蕴含的信息,最终都可以通过病历或者某种简化摘要的形式展现。如果计 算机可以针对医疗对话进行理解,在医生方面,可以帮助医生总结问题、 辅助问诊,甚至可以帮助医生在现实问诊中缓解书写病历的压力;对于医 院方面,可以帮助院方进行病历质量检查、医学生考试等;在患者方面, 则可以帮助患者理解问诊结果,也可以帮助患者在已有的所有在线问诊 对话中找到和自己病情相关的对话内容作为参考。

和普通的自然语言理解任务相比,面向医疗对话文本的自然语言理解面临着以下几个难点:

- (1)需要依托于医学知识图谱框架,而面向病历生成的中文医学知识图谱框架并不成熟。例如,从无结构的自然语言医疗对话文本中抽取信息可以看作是预定义的信息抽取,这就需要一个支撑的知识图谱框架,来定义哪些内容需要抽取,抽取出的实体和属性应该如何映射到知识图谱框架中。举一个具体例子,如果已有医学知识图谱显示,"心绞痛"这一疾病具有"发作位置"、"持续时间"、"诱因"等属性,那么在抽取时就能够有的放矢,抽取相关内容。虽然健康医疗行业以及相关科学工作者对于医学知识图谱的需求非常强烈,但是目前还没有普遍公认的、广泛应用的中文医学知识图谱框架。需要指出的是,现在已有的英文医学疾病图谱ICD-10<sup>[13]</sup>、UMLS<sup>[14]</sup>等,也只是属于医学术语词典。
- (2) 医疗对话文本语料的获取非常困难。和其它自然语言处理任务相比,医疗语料获取环节的难点体现于: a) 医疗相关的真实数据获取困难。由于涉及到患者的个人隐私以及医院的隐私政策,公开的医疗语料需要去掉一切可以推断出患者身份的信息,即需要信息脱敏。目前,很少有

公开可用的医疗对话文本语料。b) 缺少针对医疗对话文本的标注规范。针对医疗对话文本应该如何标注,现在尚没有公认的规范,其部分原因也是由于缺少相应的医学知识图谱框架。c) 和其它领域的对话语料不同,即使有了标注规范, 医疗领域的对话语料也需要医学专业的相关从业人员帮助标注与审核,和其它领域相比专业性更强,标注规范更严格、周期更长、成本也更高。

- (3)口语化的对话形式文本给自然语言处理带来额外的困难。对话的开放性给自然语言处理带来了更大的挑战,其难点具体表现为: a)对话语言口语化严重,一般情况下,语法结构不规整,而基础的自然语言处理工具,例如分词、词性标注、句法分析等工具的效果和文本的风格密切相关,而日常的口语化文本则由于其特有的复杂性和多样性,会造成这些基本的自然语言处理工具效果大打折扣。b)对话中存在大量冗余信息。人们日常对话具有随意性,语气助词较多,因此在对话中会有一些不会在普通文本中出现的内容,例如口头语和重复的连接词等,属于典型的冗余信息。
- (4) 医疗对话文本的领域专业性特点给信息抽取带来额外的困难。 医疗对话是特定场景下的对话,因此不但有普通对话的特点,还有医疗场景的特点。对于医疗场景的对话,其难点在于: a) 医疗知识密集,对话中存在大量的医学术语,而且有很多口语化的不规范的医学用语,这给自然语言处理的基础工具提出了更高的要求。例如,患者在描述症状位置时,经常会使用日常用语,如"尾巴骨"、"后脊梁"等,而其对应的实际医学术语应为"尾椎骨"、"脊柱"。b) 对话含有隐含知识,医疗对话场景下的冗余信息不容易分辨。以门诊问诊为例,医生为了判断患者的病情或排除一些可能疾病,需要提出一些问题,其中有些可能最终不会体现在病历中。还有一些对话是和本次问诊无关的,如安慰患者的话语等,在这里都可以视为冗余信息,这在医疗对话中是大量存在的。c) 否定信息在医疗对话中非常关键,因为在医疗对话中,经常会有阴性症状的表述,这就要求对否定信息的识别要更准确。例如,"头一般不疼"表示"头痛"为"阴性症状",而阴性症状通常是非常重要的指征,如果错误识别就是对事实的颠倒,严重情况下会导致医疗事故。

本文针对对话形式文本的自然语言理解问题,以医疗对话为实践领域,探索一套完整的对话理解体系,包括背景知识图谱构建、医疗实体识别与连接、医疗关系分类以及对话整体信息抽取。

具体地,本文的主要安排如下:

本章主要介绍了面向对话的自然语言理解研究的必要性和主要研究内容,以及医疗对话文本理解的相关背景介绍。

第二章,将对医疗对话的自然语言理解涉及的相关国内外研究工作 进行综述。

第三章,探讨医疗知识图谱的构建方法,并提出一种构建医疗知识图谱的方法,获得一个医疗症状图谱,作为后续工作的知识依托。

第四章,提出面向对话文本的实体和关系标注数据集,并使用己有模型,实现医疗命名实体识别、关系分类、以及症状实体链接。

第五章,提出一个面向整体对话的信息抽取方法,标注相关数据集,并提出基于深度匹配的信息抽取模型,获得对话内提及的重要医疗信息。 第六章将对本文工作进行总结,并对未来工作做出展望。

#### 2 研究现状与相关工作

#### 2.1 医学知识图谱相关研究

建立医疗领域的知识图谱是医务工作者以及医学科研人员一直以来的目标。目前,影响力大的英文医学知识图谱(或医学术语词典)有:

SNOMED-CT<sup>[15]</sup>:全称为 Systematized Nomenclature of Medicine-Clinical Terms,可译为医学系统命名法一临床术语,是一部经过系统组织编排的,便于计算机处理的医学术语集。它涵盖多方面的临床信息,如疾病、操作、药物等,可以在不同的学科、专业和不同地点之间实现对于临床数据的索引、存储、检索和聚合。采用该术语词典有助于组织病历内容,减少临床和科学研究中数据采集、编码及使用方式上的差异。

ICD-10<sup>[13]</sup>: 全称为 The International Statistical Classification of Diseases and Related Health Problems 10th Revision,可以译为国际疾病伤害及死因分类标准第十版。ICD-10 的研究起始于 1983 年,并于 1992 年完成,是世界卫生组织依据疾病的某些特征,按照规则将疾病分门别类,并用编码的方法来表示的系统。

UMLS<sup>[14]</sup>:全称为 Unified Medical Language System,可以译为一体化医学语言系统,是对生物医学科学领域内许多受控词表的一部纲目式汇编,使得不同的术语系统之间能够彼此转换。UMLS 被看作由生物医学概念构成的本体,并且具有适用于自然语言处理的工具,主要面向医学信息学领域的开发研究人员使用。

TAMBIS<sup>[16]</sup>:全称为 Transparent Access to Multiple Bioinformatics Information Sources,该项目提供了 TAMBIS 本体(TaO),描述了很大范围内的生物信息学任务和资源,并且是动态增长的。该本体不含有实例,只含有生物信息学相关的概念以及概念之间的关系,如果想要获取实例,需要借助其它外部的数据库,而该本体可以作为查询实例的接口使用。

中文方面,国内的医务工作者以及医学研究者也正在积极构建中文的医学知识图谱:

OMAHA<sup>[17]</sup> (Open Medical and Healthcare Alliance) 是为了解决医疗

健康领域数据共享这一共性难题,满足业界对数据结构化、标准化的迫切需求,行业内领先的相关机构和个人发起的开放医疗与健康联盟。该联盟提出了 OMAHA 七巧板医学术语集样例数据,涉及领域包括疾病诊断、症状、检验检查、基因、基因突变等,术语集包括概念、术语、关系(子类关系和属性关系)和映射四大核心构件。目前已积累 76 万概念,100万术语,254万关系。

OpenKG<sup>[18]</sup>是中国中文信息学会语言与知识计算专业委员会所倡导的开放知识图谱项目,旨在促进中文知识图谱数据的开放与互联,其发布的医疗相关的知识图谱有中文症状库、中医医案知识图谱、中医经方知识图谱等。

CMeKG<sup>[19]</sup>(Chinese Medical Knowledge Graph)是利用自然语言处理与文本挖掘技术,基于大规模医学文本数据,以人机结合的方式研发的中文医学知识图谱。 CMeKG 的构建参考了 ICD、MeSH<sup>[20]</sup>等权威的国际医学标准以及规模庞大、多源异构的临床指南、行业标准、诊疗规范与医学百科等医学文本信息。包括 6310 种疾病、19853 种药物、1237 种诊疗技术及设备的结构化知识描述。

侯丽等<sup>[21]</sup>对基于本体的智库系统构建进行了探讨,并且构建了以知识节点为对象的疾病库、药物库、检查库三大医学知识库,并且建立了各个库之间的关联。此外,王昊奋等<sup>[22]</sup>从现有中文开放链接数据中获取医疗信息,并搜集了主流医学站点中的医疗知识,进行本体构建和知识整合,发布了中文医疗链接数据。

以上工作都是中文医学知识图谱领域重要的进展,但是,从生成中文电子病历的角度来看,当前的中文知识图谱,主要存在着框架覆盖度不够的问题。特别是针对特定科室的电子病历,对于其相关的疾病、症状、检查、治疗和用药等要求较高,难以用现有的知识图谱框架指导抽取。

综上,目前,国际上通用的医学知识图谱为 SNOMED-CT、ICD-10、UMLS等,其框架不支持面向电子病历生成的信息抽取。中文方面,医学知识图谱目前获得了研究者越来越多的关注。但是同样,当前的中文知识图谱框架并不适用于生成电子病历。

#### 2.2 对话形式文本理解的相关研究

针对直接从医疗对话文本中抽取医学信息,并生成电子病历这一任务,唯一能找到的发表工作是 Finley 的工作<sup>[23]</sup>,是一个原型演示系统,其作者表示目前尚未有公开可用的版本。任务分为五个步骤解决:对话角色标注、语音识别、知识抽取、处理结构化数据、自然语言生成。其中最重要的就是知识抽取,知识抽取部分使用的方法主要有词典匹配、正则表达式、词典解释匹配、有监督的机器学习方法等。这篇文章只有简单介绍,没有方法细节和实验结果。

Wei 等[24]使用了儿科在线问诊的对话语料,主要目的是做一任务型 对话系统。其对儿科在线问诊对话文本进行了疾病标注,其采用了序列标 注的标注方法,以字为粒度对疾病进行了标注,然后学习得到序列标注模 型,可以抽取出对话文本中的疾病症状等实体。然后使用强化学习进行对 话策略学习,以自动诊断为目的,得到更好的任务型对话系统。Mayfield 等[25]也针对医疗对话文本,研究哪些对话内容是给出了信息的。Wang 等 [26]对儿科医患对话进行详尽的标注和分析。从对话分析的角度,分两方 面设计标注框架:对话的层级结构以及对话中的动作。会话分析 (Conversation Analysis)是话语分析(Discourse Analysis)的一个分支。会话 分析与很多领域有关,包括语用学、言语行为理论、互动的社会语言学、 交际理论、社会心理学等。而医患对话也是自然语言会话的一类。从结构 上来说,可以分为话轮,邻接对,序列以及阶段,从动作上来说,本文主 要对和抗生素有关的动作进行了定义。标注的语料来自2013年在中国医 院采集的视频,共318份,人工转化为对话。利用标注好的数据,作者进 行了医疗决策过程分析和抗生素使用情况分析。从标注情况来看,我们要 做如此详尽的标注是很困难的,但是其中阶段(现病史、既往史等对话分 段)的标注还是可行的。

任务型对话系统<sup>[27,28]</sup>的目标通过对话完成某个特定领域的任务,如订餐、订机票等。其主要关注的是人机对话,核心由三个部分组成:自然语言理解、对话管理、自然语言生成。自然语言理解的作用是理解用户的输入,如意图分类、槽位填充等;对话管理包含对话状态跟踪和对话策略两部分,对话状态跟踪是指根据目前对话的状态、系统的旧状态,来更新当

前的对话状态,而对话策略则是给出系统下一步应该进行哪些动作;自然语言生成模块则是根据系统策略,生成相应的自然语言。其中,对话状态跟踪任务[29,30,31]的相关研究对本申请有一定的借鉴意义。

传统的对话状态跟踪方法通常采用人工设计的规则来选择结果<sup>[32]</sup>,这些方法由于不够灵活会经常出现错误<sup>[33]</sup>,随着训练数据规模的增长,基于统计学习的方法逐渐崭露头角<sup>[34]</sup>。对话状态跟踪挑战(Dialog State Tracing Challenge, DSTC)<sup>[35]</sup>是专门针对这项任务提出的挑战评测,在评测榜单上,复杂的规则方法<sup>[36]</sup>,机器学习模型如条件随机场<sup>[37]</sup>、最大熵模型<sup>[38]</sup>等方法都占据一席之地。

随着深度学习的方法渗透到各个领域,Henderson等[39]在 2013 年提出了信念追踪的深度学习,其方法使用滑动窗口输出可能候选的的概率分布,并易于做领域迁移。Mrkšić等[40]在 2015 年提出了多领域 RNN 对话状态跟踪模型,其方法训练一个泛化的信念跟踪模型,继而 Mrkšić等[41]于 2016 年提出了一个神经信念跟踪器(Neural Belief Tracker, NBT)来检测槽-值对。Zhong等[42]在 2018 年了提出全局-局部自注意力机制,把槽-值选择看成一个候选排序打分问题,考虑了全局和局部两个方面,全局部分对所有槽都适用,而局部部分则是每个槽都有自己的编码器及参数。作者认为这种设计可以解决比较稀有的槽-值对预测问题。此外,编码器还使用自注意力机制来获取注意力上下文。最终得分的获取参考两方面,即用户自己描述的部分,以及系统提出的问题和用户的回答。

Sun 等人<sup>[43]</sup>通过在中国的英语考试,包括高考、大学考试、测试题等构建了基于对话的多项选择阅读理解数据集 DREAM,用于评估机器在现实生活中文本为对话场景下的理解能力。相较于一般的阅读理解数据集,DREAM 有文本数据为对话形式、更加需要多布推理和常识推理的能力,对话形式的数据相较于文本形式有着如下区别 1) 对话比较随意,存在很多不精确表达; 2) 经常出现话题转移的现象;3) 对话者之间有更频繁的信息交互现象。DREAM 利用了高质量的考试资源,需要机器从三个带选项中间选择一个正确的答案,很考验机器的跨句推理能力。

Liu 等人[44,45]利用医院护士 对于患者回访的记录作为种子数据生成模板,进一步生成大量的模仿护士-患者对话的对话数据,解决了长期以

来医疗领域数据难获取,隐私问题难以解决的问题。 特别的是护士患者对话中涉及大量的口语现象,以及明显的语言现象,例如:省略、重复、追问、确认等;数据集涉及胸痛、咳嗽等9中常见的症状,每个症状包含发作时间、诱因、严重程度、频率、位置等5种属性;Liu等人在这个数据集上进一步按照对话中的语言现象对于数据按照轮次级别进一步进行了整理,将省略、重复、确认等句子合并到一个轮次中,并且提出了轮次级别和词级别的层级注意力机制,一定程度上解决了对话数据中数据不规整的现象,并且使得机器更进一步理解对话中的知识。

综上,和医疗对话文本相关的工作数量也不多,有些从会话分析的角度进行分析,有些从信息抽取的角度进行分析。任务型对话虽然关注的是人机对话,但是其中对话管理部分也值得借鉴。综上,针对医疗对话文本的工作仍然处于起步阶段,但是可以看出其是一个极具潜力的研究方向,并拥有广阔的实用场景。

目前已有的和对话相关的阅读理解模型方面,采用的仍然是针对篇章的模型,没有考虑到对话中多轮次特征,也未考虑对说话人的角色进行有效区分,同时也没有引入相应的医学专业词典或知识图谱等,仍有很大的提升空间。

#### 2.3 医学信息抽取的研究

对病历进行信息抽取,从而得到结构化的病历表示,属于健康医疗数据的处理阶段,这个过程产生的结构化数据是计算机可以直接使用的,因此具有巨大的价值,所以针对病历信息抽取的研究工作很丰富。

最具有代表性的数据集是 I2B2 2010 评测任务<sup>[46]</sup>中给出的,该评测首次对电子病历命名实体进行了系统的分类,其依据实参照 UMLS<sup>[14]</sup>定义的语义类型,把命名实体分为了三类: 医疗问题(包括疾病和症状)、治疗、检查。这种分类充分体现了面向问题的思想,医疗手段是为了治疗医疗问题,检查是为了确认医疗问题<sup>[47]</sup>。

最早追溯到 1968 年, Weed<sup>[47]</sup>提出面向问题组织电子病历就是为了医务人员便于诊断推理<sup>[48]</sup>。1994 年 Friedma 等<sup>[49]</sup>开发了初版 MedLEE (Medical language extraction and encoding) 系统, 其方法基于词汇和语法

的规则,识别 X 射线报告里的疾病名和疾病的修饰成分,并与医疗实体词典 MED11 建立映射。IBM 公司开发的 MedKAT (Medical knowledge analysis tool)<sup>[50]</sup>可以识别癌症病历中的疾病,并提出了癌症疾病知识表示模型。 梅奥诊所的 cTAKES<sup>[51]</sup>(Clinical text analysis and knowledge extraction system)利用 SNOMED CT、RxNORM13 以及 UMLS 来抽取病历中的实体。随着机器学习研究的不断深入,基于机器学习的方法得到了研究者的关注。基于机器学习的方法主要可以分为基于分类的方法和基于序列标注的方法两个研究范式,而后者的效果一般更好。叶枫等<sup>[52]</sup>首次对中文电子病历进行识别,采用条件随机场模型对疾病、临床症状、手术操作这三类命名实体进行识别,并且加入了人工词典辅助特征,其缺点是语料规模较小,且识别的类别很少。de Bruijn 等<sup>[53]</sup>采用半马尔科夫模型进行四类标注的序列标注用,并且引入了上下文特征,以及 UMLS、cTAKES 的结果。

Rajani 等[54]把医学实体抽取看成是一个实体链接任务,具体来说,把自然语言的片段对应到 UMLS 的概念上,使用比较常用的集成方法中的模型融合方法,即训练一个模型用于组合其他各个模型,该方法的子系统有 8 个,在元分类器上,加入了附加属性。最终的抽取结果不只看子系统的选择,也要符合外部知识。

Ling 等<sup>[55]</sup>的主要目的进行自动诊断,其创新点是使用了强化学习的框架,把任务分解为概念(实体)识别和诊断两个过程。强化学习的奖励也由两部分组成,即概念识别的准确率和最终诊断的准确率。在概念识别这一步,引入了两个无结构的外部资源: Wikipedia 和 MayoClinic。外部资源主要起到了扩展概念的作用,不再局限于数据中出现的概念。深度 Q 网络用来优化,强化学习的状态主要是当前抽取的概念,动作为接收全部新概念、拒绝全部概念等。本文的数据集使用的是 TREC clinical decision support track 2015<sup>[56]</sup>,由描述、诊断总结和诊断组成。使用强化学习做医学实体抽取的方法值得关注。

针对医疗对话的信息抽取工作最具代表性的是谷歌最近的工作<sup>[57]</sup>。 这项工作从医疗对话中抽取出症状和状态。该工作定义了 186 个症状, 状态只有三种,即出现、未出现、未知。文章提出了 Span-Attribute Tagging (SA-T)模型:首先利用序列标注方法,标注出起止位置,然后结合上下文 更丰富的特征,分别预测症状以及症状状态;另外还提出了序列到序列模 型,直接把多轮对话输入,输出症状和症状状态。总体来看,这项工作最 大贡献在于定义了新任务和数据集,但数据集并未公开。

综上,针对医疗文本的医学信息抽取是一项相对成熟的任务。由于 I2B2 评测具有公开可用的数据集,因此得到了大部分研究者的青睐。从 相关研究可以看出,自然语言处理技术的进步对这项任务的效果提升起 到了很大的作用:从规则匹配,到传统机器学习模型,再到深度学习模型,效果不断提升。另一方面,目前针对中文的医疗文本信息抽取工作也取得了一定的进展。但是,由于医疗对话文本和医疗文本数据本身存在巨大差 异,因此,这些方法是否适用于医疗对话文本仍有待探索。

#### 3 医疗知识图谱构建

本章主要构建症状知识图谱,症状图谱是指以症状为结点组成的医疗知识图谱。症状是日常生活中最常用的,特别是对于医学不甚了解的普通人。对于患者来说,症状是其最直观的感受,是寻医问药的切入点;对于医生来说,症状是问诊的主要内容,是诊断、鉴别诊断的重要线索和主要依据;对于计算机处理来说,症状是人工智能问诊流程的起点。和其他类别例如疾病、检查、药物等相比,症状表述更加丰富,症状与症状之间的联系也非常紧密。构建症状图谱具有多方面的意义,例如,在医患口语对话中,症状的识别、状态的判断以及属性抽取均需要症状图谱的支撑;在病历质控方面,不仅需要了解症状与疾病、药物等的关联,还需要知道和当前症状密切相关的症状,这就需要症状图谱的帮助;在辅助诊断方面,一般需要从症状出发,设计诊断逻辑。



图 3-1: 症状图谱的应用示意图

#### 3.1 构建难点

现有很多工作是从大数据中抽取出症状,根据其在文本中的关系,以 及医生和教材的帮助,组织成为症状图谱,但是这样构建的症状知识图谱 都会遇到以下难点:

1) 症状定义边界困难。对于出现在症状图谱中的结果,需要是简洁准确的,尽管有这个前提,对于症状的界定仍然很困难,例如"胃痛"、"胃胀痛"、"胃部隐痛"等,是否按独立症状收录入图谱?如果过于严

格,那么症状图谱的覆盖度将会不足,而如果过于宽松,症状图谱中将会有很多相似的症状,为以后的链接等应用带来困难。

2) 症状和体征、检查检验结果、疾病等难以界定。互联网上的医疗数据纷繁复杂,有些定义不准确,这就导致了症状知识图谱中掺杂了疾病等其他内容。当然,有些情况的界定本身就是比较模糊的,医生也持有不一致的意见。因此,获取的症状图谱纯净度会较低。如表 3-1 所示,症状图谱中不只含有症状的情况时有发生。

表 3-1: 一些易和症状混淆的情况

12.5-	
症状	小腹痛 张口困难 心前区隐痛
体征	对光反射迟钝 干啰音 肌病步态
疾病	肾动脉硬化 脂溢性皮炎 病理性近视
检查检验结果	红细胞分布宽度偏低 直立性尿蛋白 尿中组织胺排泄增加

3) 心理和中医症状比较难以处理,这些症状在知识图谱中的会比较独立,很难与其他症状连接,因此不宜放入图谱中。一些例子如表 3-2 所示。

表 3-2: 心理症状和中医症状例子

心理症状	病理性偷窃
	精神洁癖
	考试综合症
	类戒断反应
	皮肤不仁
中压动机	皮肤不仁 睥生痰核
中医症状	
中医症状	睥生痰核

#### 3.2 构建方法

构建方法首先需要确定症状知识图谱的框架(schema),然后在各个 医学来源挖掘症状词表,最后将这些症状挂载在知识图谱框架上。

#### 3.2.1 知识图谱框架

症状知识图谱可以从解剖部位、科室等定义。我们调研目前比较流行的几个医疗知识网站,如 A+医学百科<sup>[58]</sup>、快速问医生<sup>[59]</sup>、寻医问药网<sup>[60]</sup>、99 健康网<sup>[61]</sup>等,发现利用部位维度作为框架主体是最符合应用需求的。在咨询专业医生后,除了部位维度外,我们还定义了原子维度和全身性维度。

部位维度:这里并没有采用医学解剖部位,因为这些部位名称过于专业,无论是问诊还是病历中,很少会有这些部位的专业名词出现,因此实用性不高。我们采用几个网站给出的部位做了综合,并且根据医生的建议做了修改。最终确定头颈部范围、腹部范围、臀部范围、胸部范围、背部范围、四肢范围这6个作为大部位。

原子维度:为了解决很多症状都具有同一个核心症状这个特点,我们使用原子症状,在另一个维度把各种同类症状组织起来。例如,我们设立了眩晕问题,头晕、晕眩等都属于这个原子症状,夜间出汗、出汗减少等都是属于出汗问题。

全身性维度: 很多出现在不同部位的症状, 其实都属于某个特定的全

身性维度。例如,小腿肌肉疼痛和小臂肌肉疼痛,都可以归于肌肉这个维度,肩周疼痛和膝盖疼痛都可以归于关节这个维度。

整体的症状知识图谱框架如图 3-2 所示。

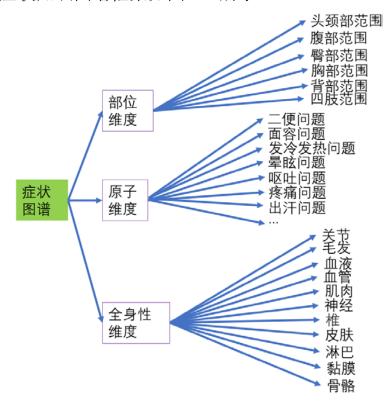


图 3-2: 症状知识图谱的构建框架

#### 3.2.2 症状获取

我们从几个医学健康网站爬取了标注为症状的实体,具体获取的数据情况如表 3-3 所示。

1, 3-,	5. 犹取的症状数据
来源网站	数目
39 健康网[62]	7815
99 健康网[61]	6060
寻医问药网[60]	6864
飞华健康网[63]	10448
快速问医生[59]	5740
A+医学百科 <sup>[58]</sup>	6885
Total(去重)	12359

表 3-3: 获取的症状数据

我们按照刚才所说的框架,请标注人员将这些症状实体挂载到图谱框架上,得到初版的症状知识图谱,症状总数为9723个。

从部位维度来看,症状的分布如图 3-3 所示,这里只列举到 2 级层次部位。总体来说,获取的症状中,头颈部范围的症状最多,背部范围的症状最少。

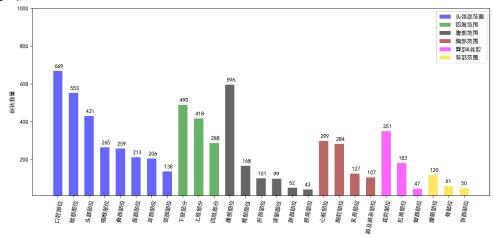


图 3-3: 部位维度症状分布

从原子维度,我们把各个症状在原子症状的维度上的分布统计展示在图 3-4 中。从图中可以看到,疼痛问题是症状中最多的,后面依次是发肿问题、二便问题等。

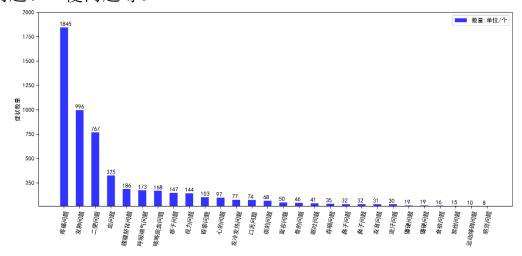


图 3-4: 原子症状维度症状分布

从全身性维度,我们也做了症状分布的统计,如图 3-5 所示,可以看到,骨骼相关问题、皮肤相关问题最多,血液相关问题最少。

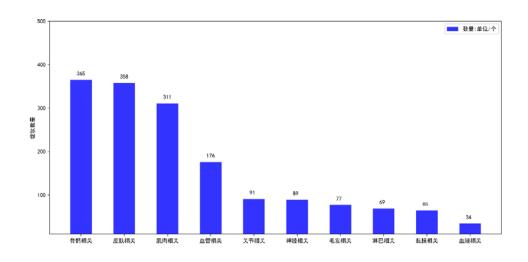


图 3-5: 全身性维度症状分布

通过这三个维度,可以覆盖 87.71%的症状,覆盖率较高。其余还有 12.29%的症状,无法通过这三个维度描述,因此无法通过图谱框架组织 起来。整体的覆盖情况如图 3-6 所示。

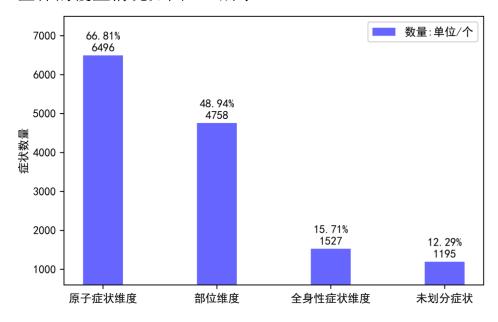


图 3-6: 三个维度症状覆盖的症状情况

#### 3.3 症状图谱构建的自动方法

#### 3.3.1 同义词挖掘

在这里,我们使用模板得到同义词集合:

• A(即是|即为|即指|即|又叫做|又叫作|又称为|亦称|又叫|简称|也叫作|也叫做|也叫|俗称为|俗称|又名|又名为|又称)B

- A 医学是上(又称为|又叫做|又叫作|也称为|也叫作|也叫做|称为|称|叫做|叫作|叫|也名)B
- A(是)B 的俗称

最终获得同义词词典: 396 个集合(每个集合一般为 2 个词或 3 个词),具体情况如表 3-4 所示。

 同义词的集合个数
 396

 含有 2 个同义词的集合数量
 379

 含有 3 个同义词的集合数量
 15

表 3-4: 同义词获取情况

#### 3.3.2 上下位词挖掘

使用模板得到上下位词集合

- A 是 B 的一种
- A 是一种 B
- A 分为 B 和 C

最终获得 858 个集合 (一个上位词+若干下位词构成一个集合)。 图 3-7 是上下位词的一些例子。

1. 腹肌强直可分为随意性肌强直和非随意性肌强直。随意性肌强直呈对称性......

腹肌强直可分为随意性肌强直和非随意性肌强直 随意性肌强直呈对称性,吸气时肌强直(呼气时肌肉松弛),使用肌肉松弛技巧有效,直立位腹肌收缩无腹痛。非随意性肌强直常为非对称性,呼气和吸气均存在腹肌强直,使用肌肉松弛技巧无效,直立位腹肌收缩感疼痛。腹肌强直是

#### 2.分离性漫游症是漫游症的一种,是有短暂失忆或者长久失忆的一种漫游症状.....

分离性漫游症是 漫游症的一种 是有短暂失忆或者长久失忆的一种漫游症状,一般,漫游者会失去自己曾经所在的环境记忆,而获取了以前的某段记忆作为自己的记忆或新的状态,去生活在另外一个环境,完全忘记自己而把自己认为是自己曾经记忆中的某个人。

图 3-7: 上位词抽取的例子

#### 3.3.3 中医症状识别

如本章开始时所述,中医症状暂不纳入症状知识图谱,但是,从网上获取的症状中有很多是中医症状,因此,需要过滤掉这些症状,为了减轻标注负担,我们训练了自动识别中医症状的分类器。我们标注了 2000 条数据,将标注的数据按比例分成 8:1:1 的训练集、开发集和测试集。测试

集大小为 220, 其中非中医为 171 条, 中医为 49 条。我们采用现在已有主流文本分类方法, 实验结果在表 3-5 中列出。

Model	Micro-P	Micro-R	Micro-F
FastText (2016, Armand. etc)	0.9052	0.5918	0.6028
CNN (2014, Yoon. etc)	0.9185	0.8382	0.8699
DPCNN (2016, Rie Johnson. etc)	0.9412	0.8717	0.9006
RNN (2016, Pengfei. etc)	0.3886	0.5000	0.4373
RNN + Att (2016, Peng Zhou. etc)	0.9477	0.8921	0.9162
Transformer (2016, Rie Johnson. etc)	0.9452	0.8440	0.8823

表 3-5: 中医识别性能

可以看出,中医识别的 F 值可以达到 0.91,可以有效帮助标注人员识别中医症状。

#### 3.3.4 症状成分切分

已有的上下位关系仅仅表示一个症状是另一个症状的上位词,但是,在症状图谱中,特别是在我们构建的三个维度的图谱中,可以有更加丰富的上下位表示。例如,"出血"是"牙龈严重出血"的上位症状,除此之外,我们还知道,这种上位关系蕴含了部位上位(牙龈)和程度上位(严重)。为了捕捉这种更加丰富的上下位关系,我们对症状成分切分进行了尝试。图 3-8 是症状切分和上下位关系表示的例子。

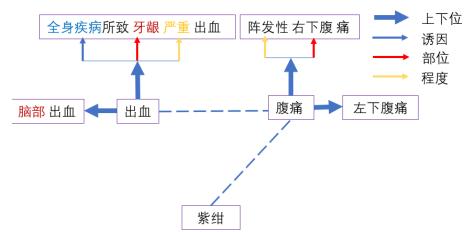


图 3-8: 通过症状成分切分进行更丰富的上下位关系表示

我们把症状成分切分任务形式化为序列标注任务。标注的数据集的是从各个医疗网站爬取的 9723 个症状中,选择了 2000 条进行了最初标注。在去除了中医、心理以及无法划分的整体症状(比如唐氏综合征),剩余的有效标注条数为 1401 条,然后按照划分为 8:1:1 的比例划分为训

练集、开发集和测试集。

最终的数据集使用了 5 种标签,即原子症状(Atom)、器官(Orgn)、性质(Prop)和群体(Grop)和其他(O),使用 BIO 标注方法。

数据集	症状数目	B-Atom	I-Atom	B-Orgn	I-Orgn	B-Prop	I-Prop	B-Grop	I-Grop	О	
Train	1120 条	969	1187	652	828	238	452	47	58	6477	10908
		(8.8%)	(10.8%)	(5.9%)	(7.6%)	(2.1%)	(4.1%)	(0.4%)	(0.5%)	(59.4%)	
Test	142 条	105	117	77	111	36	62	4	5	777	1294
		(8.1%)	(9.0%)	(6.0%)	(8.6%)	(2.8%)	(4.8%)	(0.3%)	(0.4%)	(60.0%)	
Dev	139 条	113	135	88	106	34	59	4	4	756	1299
		(8.7%)	(10.4%)	(6.8%)	(8.2%)	(2.6%)	(4.5%)	(0.3%)	(0.3%)	(58.2%)	

表 3-6: 数据集详细情况

由于已经把成分切分任务形式化为了序列标注任务,我们使用目前在序列标注中表现较好的 BERT+CRF 模型来实现,图 3-9 为模型框架,下面进行详细介绍。

#### 1) BERT 编码器

BERT 神经网络仅用于表示输入文本,而不会在训练阶段进行微调。除了 Dropout 之外,冻结所有层的参数,这会减少过度拟合。我们使用官方的基于 BERT 的中文预训练模型。

#### 2) BERT 加权

由于我们不对 BERT 进行微调,因此应针对特定序列标记任务调整 其输出。

#### 3) LSTM 层

使用LSTM层来学习输入序列中标记之间的依赖性。

#### 4) 多头注意力

使用自注意力机制可以帮助每个字符学习序列中的其他依赖项。

#### 5) 线性层

使用 tanh 激活函数的线性层。

#### 6) NCRF ++

CRF++通过在相邻标签之间添加转移分数来捕获标签依赖性。NCRF++是支持句子级最大对数似然损失训练的 CRF。在 N-best 输出下扩展了解码算法。

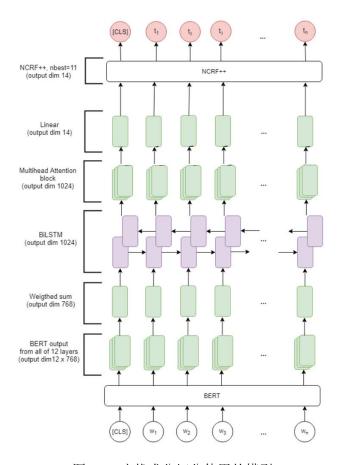


图 3-9: 症状成分切分使用的模型

下表为整体实验结果,以及各个标签的具体结果。

表 3-7: 症状成分划分实验结果

	Precision	Red	call	F1-score	Support					
B_Atom	0.809	0.838		0.823	113					
I_Atom	0.848	0.854		0.851	135					
B_Orgn	0.932	0.932		0.932	88					
I_Orgn	0.890	0.915		0.902	106					
B_Prop	0.794	0.871		0.831	34					
I_Prop	0.857	0.871		0.864	59					
B_Grop	1.000	1.000		1.000	4					
I_Grop	1.000	1.000		1.000	4					
O	0.851	0.803		0.826	756					
总体情况										
Micro avg	0.858	0.858		).858	0.858					
Macro avg	0.88	7	0.898		0.892					
Weighted avg	g 0.859	0.859		).858	0.858					

下面将对实验结果进行分析:

首先,根据预测结果,可以得到序列标注的混淆矩阵,如图 3-10 所

示。

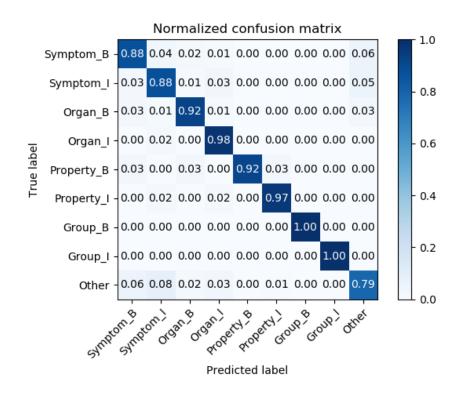


图 3-10: 症状成分切分的实验结果混淆矩阵

通过对混淆矩阵的分析,我们可以发现问题主要出在症状(Symptom)和其他(Other)上,总结主要错误原因如下:

- 原子症状定义不明确,对于划分造成了困难
- 症状作为修饰词的情况偶尔出现
- 原子症状未被学习到

#### 3.4 症状图谱的扩展

本章构建的症状知识图谱,一定会存在症状覆盖不全的情况,因此,本小节使用症状成分划分算法,找到新症状在已有图谱中的位置,以便扩展症状知识图谱。如图 3-11 所示,新的症状"下腹部有出血红斑"是已有图谱的很多结点的下位词。

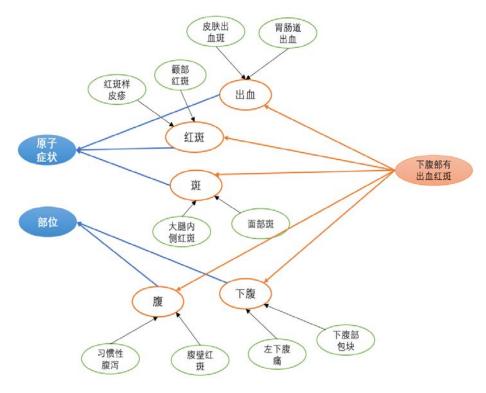


图 3-11: 新症状映射扩展的例子

设图谱里所有症状集合为  $S{S1, S2, \cdots Sn}$ , 图谱里的三个维度的非症状结点为  $N(n1, n2, \cdots, nk)$ , 新症状为 S'。

在相似症状连接方面:将 S 中所有症状提前进行成分划分,新症状 S'也进行成分划分,然后将新症状的成分划分结果与所有的图谱症状进行 比较打分 (如下图所示),然后对所有分值进行排序,选出最好的几个。则打分的规则如下:

$$Score = \frac{\sum T_w \times S_w \times \frac{len(part_i)}{len(S_i)} \times \frac{len(part_j)}{len(S')}}{len(part_{list_S}) \times len(part_{list_{S'}})}$$

其中, $S_i \in S$ , $part_i \in S_i$ , $part_j \in S'$ , $T_w$ 为症状成分的权重,例如,原子症状权重为 1.0,部位为 0.7,性质为 0.6, $S_w$ 为症状陈各分匹配相似性的权重,例如相等为 1.0,包含为 0.5。注意仅当症状成分类型一样时才进行比较打分。

在图谱上下位连接方面: 只支持对图谱中三个维度(原子症状、位置、全身)的最末端的非症状结点的上下位连接。实际上,即使用字符串匹配方法: 若新症状S'结点 $n_i$ ,则认为新症状S'中为结点 $n_i$ 的下位。由于此任务尚未有标准测试集,因此无法做定量评价,部分新症状的映射结果在图 26

3-12 中展示,其中,symptom 表示和其最相近的症状,group 则显示其在三个组织维度上的上位词。



图 3-12: 部分新症状的映射结果

#### 3.5 本章小结

本章主要关注医疗知识图谱的构建,重点阐述了症状知识图谱的构建过程。在人工构建方面,基于医疗症状的特点,提出以三个维度为主要框架的症状知识图谱,覆盖 9723 个常见症状。在此基础上,本章总结了症状知识图谱构建中涉及的自动构建方法,包括同义词和上下位词挖掘、中医症状判断和症状成分切分算法等。在症状图谱扩展方面,探索了利用症状成分切分进行新症状定位的方法。医疗知识图谱的构建是一个长期的过程,本章仅从症状角度进行图谱构建,在此过程中探索医学图谱通用的构建方法。

# 4 医疗实体与属性抽取

医疗实体抽取是目前开展的比较多的研究工作<sup>[64,65,66]</sup>,但是基本都是针对文献以及电子病历等相对标准化表达的医疗文本,直接针对口语对话的抽取仍然很少<sup>[67]</sup>。而在口语化文本的关系抽取方面,则面临着更加复杂的问题,如关系的定义以及篇章级别的抽取等,目前尚未有相关工作直接研究。因此,在本章我们提出面向口语对话医疗文本的实体抽取和链接,以及关系分类方法。主要分为三部分,首先介绍简单的规则抽取方法,然后介绍实体和关系的自动抽取和分类方法,最后介绍实体链接的工作。

#### 4.1 简单规则抽取

在任务实践过程中,我们发现有些信息通过简单的规则即可抽取。例如,一般信息,如睡眠情况,二便情况等,在对话中的询问方式具有通用性,因此存在使用规则进行抽取的可能。

例子:

现病史中一般信息: 饮食、睡眠、二便、体重、体力、精神

既往史:血压、血糖、血脂

个人史:居住城市与时长、是否吸烟、是否饮酒、是否接触有害物质、 是否接触疫区、过敏史、手术史

婚姻史与月经史:是否结婚、结婚年龄、配偶健康情况、配偶是否亡故、是否近亲结婚、是否绝经

家族史: 有何亲属患何疾病

任务可以形式化定义如下:

输入: 医患对话文本

输出: 部分结构化数据, 如一般信息状况, 既往史情况等

# 4.1.1 数据集

本子任务使用规则方式,因此所需数据为真实录音数据。我们对真实

医患对话录音数据进行了标注, 并依次构建规则。

我们标注了 60 段医患口语对话,其中使用 30 段对话用于设计和调试规则,另外 30 段对话作为测试集,检验效果。部分标注情况如表 4-1 所示:

	25	2	体重	未见变化	20
/ <del>/</del>	24	1		增加	1
结婚年龄	27	1		支架	5
	23	1		搭桥	4
	如常	11	手术史	否	3
	不佳	3		子宫切除	2
饮食	尚可	2		白内障	2
	佳	1	毒物接触史	否	7
	正常	1	吸烟	否	13
	如常	20	WX M24	是	13
二便		1		子女,健康	13
	尿急			父亲, 高血压	9
饮酒	否	14	家族史	父亲,冠心病	7
W/H	是	10	3.37.2	姐姐,冠心病	5
	不佳	10		母亲, 冠心病	5
睡眠	如常	8	疫区接触史	否	5
一	佳	2	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	杏	12
	尚可	1		磺胺	2
	否	5	过敏史	青霉素	2
外伤史	摔伤	1		黄安	1
71132	骨折	1		核桃	1

表 4-1: 简单规则抽取的对话标注情况

# 4.1.2 规则方法

我们构建了一系列的启发式规则。一些典型的规则如下:

- 追踪多轮对话中对关键词,如饮食、家族史等的肯定否认状态 过程
- 关键词和正则模板匹配关键常见语句表达方式,抽取出信息
- 使用具有表征信息的词,如药物、手术等的出现频数进行综合打分,根据分值进行综合判断

# 4.1.3 实验结果

实验的结果如表 4-2 所示。

Slot 空 (P/R/F) 不佳/否 (P/R/F) 饮食 0.700 7 19 1.000 0.823 1.000 0.789 0.882 0.750 0.666 4 睡眠 0.500 4 0.833 0.937 0.882 16 0.777 0.700 0.736 10 二便 0.500 0.666 6 0.937 0.882 0.909 0.285 0.444 7 1.000 17 1.000 体重 0.833 0.666 6 0.882 0.789 0.833 0.750 0.666 5 19 高血压 0.000 0.000 0.000 0 0.750 0.900 0.818 20 0.666 0.400 0.500 10 高血糖 0.000 0.000 0.000 2 1.000 1.000 1.000 10 0.875 0.777 0.823 18 7 高血脂 6 0.761 0 941 0.842 17 1.000 0.428 1.000 输血中 1.000 1.000 18 0.000 0.000 0.000 0 1.000 0.916 0.956 12 1.000 1.000 1.000 1.000 1.000 1.000 4 0.000 0.000 0.000 0 26 吸烟 0.750 0.960 1.000 0.857 3 0.923 1.000 12 1.000 0.866 0.928 15 饮酒 0.800 1.000 0.888 4 0.909 0.833 0.869 12 0.928 0.928 0.928 14 9 毒物接触史 0.954 1.000 0.976 21 0.000 0.000 0.000 0 1.000 0.888 0.941 疫区接触史 1.000 1.000 1.000 26 0.000 0.000 0.000 0 1.000 1.000 1.000 4 0.950 0.833 0.833 外伤史 0.904 0.926 21 0.500 3 0.833 6 手术史 1.000 1.000 1.000 4 0.916 1.000 0.956 22 1.000 0.500 0.666 4 过敏史 1.000 3 0.857 0.882 0.909 17 1.000 1.000 0.818 0.900 10 0.937

表 4-2: 简单规则抽取的评估结果

从结果中可以看出,使用简单的方法可以保证抽取出简单的信息,并 且针对出现较多的情况,规则抽取的 F 值可以满足需求。

具体地,表现较好的类别有饮食/睡眠/二便,这三种在肯定表达时比较容易,否定表达时则表现差些。因为在否定表达时表达要丰富些。二便否定表达情况很多而且共同点太少。睡眠有时候也不好处理,主要是在描述病情时常常出现睡觉情况,容易与病情情况搅在一起;三高在阳性时各种特征(用药,测量,表述句式)比较明显,容易判断;吸烟/饮酒在正负例上表现尚可。

而也有一些类别的表现较差,比如体重,体重判断正负相对难些,主要是表达方式多变,现有的例子不足,而一个个添加表达规则又低效,一旦遇到新例就会失效;手术史效果非常差,可能的原因是,手术种类很多,手术名称多变缩写太多(切瘤子/去瘤/粉瘤切除),手术前后没有独特的提示词(做了、切了),难以判断是否作过(是要做,说是做)等。

# 4.2 实体抽取和关系分类

# 4.2.1 数据集

除简单的易于抽取的信息之外,面对更加复杂的情况,还需要抽取对话中的实体信息。这和从医疗文献中进行实体识别是一样的,但是口语化的医学实体在文本的出现会更加灵活多变。仅仅获取几种疾病是否出现

这样的结构化信息对于生成病历来说是远远不够的。为了能够获取更加 丰富的结构化信息,我们必须要考虑更加复杂的疾病属性。例如,在真实 病历中,经常会出现某疾病于何时发作,发作持续时间,发作诱因或者性 质等。为此,我们的目标就是除抽取疾病本身外,进一步抽取疾病的性质。

为了更好的进行抽取工作,我们定义了标注的范围,在表 4-3 列出。

表 4-3: 标注内容示例

标注 任务	内容	详情举例
实体标注	10 类:疾病或症状、时间、部位、频率、药物、用法用量、检查、 等术、性质、非实体信息	背痛、出汗、恶心、发绀、发 热、房颤、房间隔缺损、 腹 不适、感冒、高血压、高血 糖、高血脂、冠心病、呼吸困 难、甲亢、僵硬、咳嗽、流 消、呕吐、贫血、 室血、 第、 呕吐、贫血、 实 至 。 以 ,
关系 标注	16 类:疾病或症状、手术、检查、药物相关的 关系	疾病或症状持续时间、疾病或症状每次发作时间、疾病或症状开始或发作时间、疾病或症状结束时间、疾病或症状相关部位、疾病或症状发作频率、疾病或症状性质、疾病或症状性质、疾病或症状加重条件、疾病或症状缓解条件、用药持续时间、用药开始时间、用药结束时间、用药用法用量、检查时间、手术时间

为了使这样复杂的标注效率更高,我们借助于标注工具 brat,将待标注数据导入到标注工具中,交给标注人员标注。使用 brat 标注工具的示例如下,左侧为 brat 的界面,右侧为标注后得到的结果,以结构化文本表示:

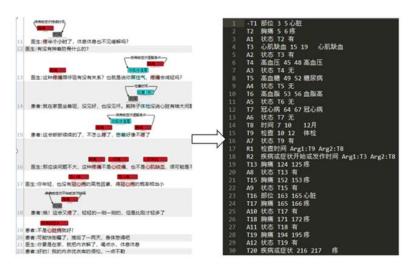


图 4-1: brat 标注工具,图中的色块即为我们定义的需要标注的实体,实体间的连线即为实体时间的属性关系。

#### 数据标注工作有以下难点:

- 实体总共需要标注 53 种,其中 47 种实体还需要同时标注状态,44 种实体还需要同时标注状态和 ID,并且状态和 ID 的标注需要结合上下文才能确定。
- 关系总共需要标注 16 种。关系多,容易漏标。
- 口语化描述、错别字不易标注。
- 标注规范随着标注样本的增多需保持更替。

# 表 4-4 为实体标注情况:

表 4-4: 实体标注数据统计

实体类型	标注数量
疾病或症状(父类)	5460
疾病或症状(子类)	4515
检查	2120
手术	274
药物	1776
时间	1190
频率	106
部位	846
非实体信息	258
用法用量	210
性质	378

对于关系, 我们定了如下关系, 标注的具体情况如下表所示。

表 4-5: 关系标注数据统计

关系类型	标注数量
疾病或症状开始或发作时间	514
疾病或症状结束时间	39
疾病或症状每次发作时间	24
疾病或症状持续时间	405
疾病或症状发作频率	105
疾病或症状相关部位	810
疾病或症状性质	699
疾病或症状触发条件	209
疾病或症状加重条件	26
疾病或症状缓解条件	33
手术时间	7
检查时间	123
用药开始时间	25
用药结束时间	4
用药持续时间	63
用药用法用量	147
总计	3233

# 4.2.2 医疗命名实体识别

我们首先进行的是医疗命名实体识别 (NER) [68], 从春雨医生[69]的对话中来抽取实体。通过数据筛选和预处理,我们得到的命名实体识别数据情况如表 4-6 所示。

表 4-6: 用于 NER 的数据统计

Е	Entity		检查		药物		手术			时间		性质		部位	
3	数量		2120		1776		274			1190		378		846	
心衰	高血压	心肌缺血	呼吸困难	发热	头晕	水肿	高血脂	先天性 心脏病	心律不齐	感冒	胸闷	冠心病	头痛	高血糖	呕吐
533	372	325	310	283	227	208	201	191	146	133	132	128	121	96	93
咳嗽	糖尿病	腹部不适	出汗	室间隔缺损	恶心	胸痛	心慌	房间隔缺损	贫血	战栗抽 搐	心肌梗 死	胃部不 适	发绀	心脏肥大	心绞痛
92	91	78	65	57	54	52	51	50	49	44	44	40	38	37	29

命名实体识别模型我们使用 BERT+CRF<sup>[70]</sup>模型,来进行医疗命名实体识别。

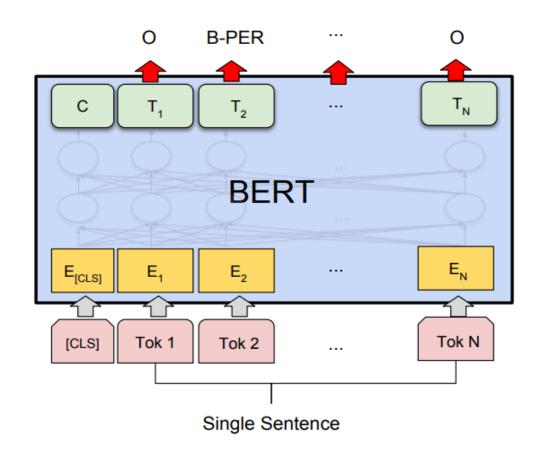


图 4-2: BERT+CRF 模型结构

下表为实验结果,我们给出了每个类别的具体结果,其中 Support 表示该标签在测试集中出现的次数。

表 4-7: NER 实验结果

		P	R	F	Support
	B-	0.94	0.98	0.96	1493
疾病	I-	0.96	0.97	0.96	971
	E-	0.95	0.98	0.96	1423
	B-	0.85	0.92	0.89	995
疾病-有	I-	0.84	0.92	0.88	688
	E-	0.86	0.82	0.89	948
	B-	0.68	0.72	0.7	322
疾病-无	I-	0.7	0.74	0.72	176
	E-	0.65	0.73	0.69	302
	B-	0.95	0.96	0.95	827
检查	I-	0.98	0.94	0.96	1143
	E-	0.95	0.95	0.95	821
	B-	0.95	0.97	0.96	658
药物	I-	0.96	0.99	0.97	1349
	E-	0.95	0.98	0.96	654
	B-	0.9	0.84	0.87	82
手术	I-	0.92	0.96	0.94	119
	E-	0.9	0.87	0.88	82
	B-	0.48	0.89	0.83	269
部位	I-	0.78	0.83	0.8	121
	E-	0.79	0.86	0.83	216
	B-	0.75	0.84	0.79	128
性质	I-	0.76	0.89	0.82	53
	E-	0.73	0.88	0.8	125
	B-	0.84	0.91	0.88	432
时间	I-	0.85	0.92	0.88	318
	E-	0.09	0.92	0.16	431

#### 实验结果分析:

- NER 工具在很多类别上都取得了不错的效果,如疾病(有)、 检查和药物等,达到可以直接应用的水平。
- 医疗对话文本中存在大量的口语话现象,相对来说,传统的人名、地名实体表达更加整齐规范。例如: "性质:一阵一阵的痛", "针扎了一下,一堆密密麻麻的特别小"。
- 标注数据量和标注质量对于影响很大:在经过对开发集的错误分析并且反馈给标注人员重新标注测试集后,结果明显上升;数据使用滑动窗口后,数据量增加,效果也进一步提升。

#### 4.2.3 关系分类

用于关系分类的数据情况如下表所示。

疾病或症状性质

关系类 关系三 对话数 对话轮 字符数 实体数 型 量 数 元组 922 403115 22429 12986 4 2164 关系类型 Train Test 疾病或症状相关部位 717 137 疾病或症状开始或发作 464 60 时间 疾病或症状持续时间 376 56

表 4-8: 用于关系分类的数据情况

关系分类模型方面,由于在医疗对话中,需要考虑篇章级的关系,DocRed<sup>[71]</sup>模型,可以大量跨句子关系实例,但是在的目前我们的数据尚未能按照其方式进行组织。我们采用 CNN/LSTM/FFN/Transformer<sup>[72]</sup>等模型,模型利用的特征有上下文表示、头尾实体类别的表示、头尾实体间距离表示,以及上下文与头尾实体的相对位置表示。实验结果在表 4-9 中列出。

312

42

	P	R	F1
Transformer	0.6416	0.6373	0.6395
CNN	0.6703	0.6271	0.6480
LSTM	0.7981	0.5763	0.6693
BiLSTM	0.7171	0.7390	0.7279

表 4-9: 关系分类的实验效果

- □ 下表为错误分析,可以看出,错误主要来自于:
  - 未充分利用上下文语境信息,无关系的实体配对也容易预测出关系。
  - 疾病或症状未考虑实体 id (未标注 id),因为多个 mention 可能对应同一个实体,导致其他同义的疾病或症状 mention 与其他类型实体匹配时会预测出关系,有正有误。

真实\预测	Na	部位	性质	开始或发 作时间	持续时间
Na	0	33	14	19	18
部位	19	118	0	0	0
性质	5	0	37	0	0
开始或发 作时间	23	0	0	37	0
持续时间	28	0	0	0	26

表 4-10: 属性抽取的混淆矩阵

### 4.3 实体链接

医疗实体链接<sup>[73,74,75]</sup>任务的目标是将对话文本中的实体提及对应到 医学知识图谱的实体上。我们把已有的标注数据中的症状和疾病对应到 ICD-10上,即以 ICD-10 作为待链接的医疗知识图谱。选择 ICD-10 的原 因主要是因为其结构相对简单,标注难度相对小,后续可以继续利用我们 自己构建的医疗知识图谱。

# 4.3.1 数据集

实体链接的数据集情况如下:

all Train Test Dev 19 类 随机划分 165 452 140 Mention 532 Mentions(未去重) 532 6081 813 797 Context 5801 4600 600 601 40 类 随机划分 Mention 929 828 244 259 Mentions(未去重) 929 11425 1278 1508 10846 8700 1000 1146 Context

表 4-11: 症状实体链接的数据情况

#### 4.3.2 模型

我们采用几种不同的模型:

作为基线模型,使用基于字符串的模型,利用 Jaccard 系数来判断链接对象。

有监督模型方面,使用基于 BERT-WWM<sup>[76]</sup>的模型,并采用下述两种模型设定:

- BERT-WWM (Only Mention):考察无上下文情况的性能。
- BERT-WWM (Mention + Context): 考察 mention 在上下文语境中的性能。

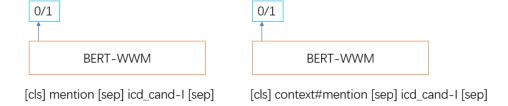


图 4-3: 实体链接模型示意图

### 4.3.3 实验结果

实验说明:

- 在症状实体链接的测试文件,使用训练好的 NER 模型进行预测。
- 整个句子作为评价单位, NER 预测合法提及, 进一步进行链接。
- 评价按照标准实体链接指标评价。

acc@n: 测试集预测结果排序 top n/836 如果等于正确链接 ICD-10 实体, 则算正确。

1) NER 实验结果, 在下表中列出

表 4-12: 命名实体识别的实验结果

NER		19 类		40 类				
指标	P	R	F	P R F				
结果	0.9258	0.9359	0.9278	0.9474	0.9450	0.9461		

# 2) 实体链接的实验结果,在下表中列出

表 4-13: 实体链接的实验结果

			19 类		40 类			
		acc@1	acc@5	acc@50	acc@1	acc@5	acc@50	
	String- based	0.29	0.40	0.45	0.40	0.52	0.67	
方法	Mention- based	0.28	0.80	1.0	0.42	0.94	1.0	
	BERT	0.90	0.99	1.0	0.92	1.0	1.0	

注: acc@n: 代表对于整个候选列表进行预测,如果正确答案在 topN中,则 topN 正确数量加一。

# 3) 整体结果

表 4-14: 阶段管道流程实验结果

	19 类												
	acc@1			acc@5			acc@50						
P	R	F	P	R	F	P	R	F					
0.848	0.857	0.852	0.924	0.934	0.929	0.925	0.935	0.930					
				40 类									
	acc@1			acc@5			acc@50						
P	R	F	P	R	F	P	R	F					
0.877	0.874	0.875	0.947	0.945	0.946	0.947	0.945	0.946					

从实验结果可以看出,使用 NER 和实体链接连接式的方式,最终取得 F 值 0.877 的效果。

#### 4.4 本章小结

本章主要关注医疗实体与医疗属性的抽取方法,为了实现良好的抽取效果,我们首先进行了数据集标注工作,不仅标注了疾病、症状、检查、手术等实体,还标注了各实体之间存在的关系。在此数据集基础上,本章分别研究了医疗命名实体识别,以及医疗关系分类两个问题。医疗命名实体识别仍然采用基于序列标注的方法,明确实体的边界和类别标签,从而将口语化表达转化为规范表达;关系分类则完成了医学属性抽取的工作,采用了多种关系分类方法。此外,本章还尝试将对话文本中出现的实体链接到医疗知识图谱,从而帮助后续的自动诊断、病历质控等落地应用。

# 5 面向医疗对话篇章的信息抽取

#### 5.1 研究动机

正如前面章节所提到的,传统的实体抽取和关系分类对于医疗对话标注的要求非常苛刻,为此,本章提出一个面向医患对话文本的整体信息抽取系统,它可以从对话中抽取出症状、检查、手术、一般信息及其相应的状态。我们收集并标注了医疗对话数据集,采用窗口式标注,和序列标注相比,减缓了标注难度。另外,针对医疗问诊对话文本的难点,我们提出的模型采用深度匹配神经网络,能够考虑到对话的结构信息。实验结果表明,本文提出的方法对于解决医疗对话信息抽取问题具有很好的研究前景。

健康医疗是一项关系到国计民生的重大事业,也在人工智能未来研究中有着举足轻重的位置。近年来,随着全球范围内医疗信息系统现代化的稳步推进,健康医疗数据的规模呈爆炸式增长,越来越多的医院开始全面推广电子病历。和普通的纸质病历相比,电子病历更容易保存和查询,无论对于患者和医院都提供了更大的便利。一方面,患者可以更加便捷地查看自己的健康记录;另一方面,医生也能迅速查看患者以往的就医记录,或者查找相似患者的诊疗方法。更重要的是,由于电子病历便于计算机直接处理,所以医学工作者可以更轻松地展开疾病的各种调查研究,例如,流行病学研究、寻找同类病人群体等。

尽管电子病历能够带来如此多的好处,但是它同时有一个不容忽视的缺点,那就是电子病历的产生十分耗费医生的精力[77]。据调查,美国的医生平均每天要花费多达 11 个小时来和临床文档系统打交道,甚至还要再多花费 1.4 小时的个人时间才能完全处理好电子病历[57],这无疑使得电子病历成为了医生们的沉重负担,它带来的更严重问题是,医生由于要耗费如此多的时间和精力写电子病历,和患者做有效交流的时间必然会减少,这些都是电子病历不可忽视的弊端。

如果使用自动的方法,直接将医患的对话内容转化为电子病历,将极 大缓解医生的压力,使得医生有更多的精力去和患者进行充分的交流,问 诊质量也将会提高。从对话中首先抽取出重要医学信息是很有必要的一步,我们认为和端到端模型相比,这一中间环节是必不可少的,主要基于两个原因: 1)结构化信息是非常有用的,而端到端模型则没有这个中间步骤; 2)端到端模型需要更大的数据量来进行训练,而医患对话的标注语料难以大规模获取。

在本章中,我们首先使用了滑动窗口的方式对于语料进行标注,然后提出一个基于深度匹配网络的抽取系统。我们从中文在线医疗问诊网站春雨医生收集医患对话文本,为了信息标注的方便,我们只关注心内科这一个科室的对话,因为这个科室的对话的问诊内容更加丰富。标注规则的制定要同时兼顾效率和可行性。我们定义了4种大类,即症状、检查、手术和一般信息,其中一般信息是指睡眠情况、饮食情况等。与此同时,还定义了它们相应的状态。这种标注方式主要有两个好处:1)和序列标注相比,这种标注方式更加简单,对于标注人员的医学背景知识要求也更低。2)可以标注多轮对话共同描述同一信息的情况。我们共标注了1120段对话,获得了超过40000个标签。

我们提出的抽取模型则包括 4 个主要组成部分,即编码模块、匹配模块、集成模块和打分模块。我们做了详尽的实验,实验结果表明本文提出的方法在窗口级别和对话级别上可以达到 69.29 的 F 值,说明本方法有较好的研究前景。

# 5.2 方法

# 5.2.1 数据集

我们使用互联网医疗咨询对话,即春雨医生的在线医疗对话文本作为语料。我们的主要标注原则是尽量详尽地标注对话中提到的医学信息。目前,在自然语言处理领域经常被使用的标注范式是序列标注,即 BIO 的标注形式<sup>[78,79]</sup>。但是,这种方法也存在缺点,如:标签信息通过多轮描述时无法用序列标签标注,如表 5-1 所示,这种情况在医患对话中并不罕见。

表 5-1: 对话标注示例

对话	标签
患者: "医生, 你帮我看看这个是不是	症状:早搏,状态:阳性
早搏?"	症状:心慌,状态:阳性
医生: "从心电图看你这个是,平时有	症状:胸痛,状态:阳性
没有心慌、气短?"	症状: 呼吸困难, 状态: 阳性
患者: "有的, 医生我这个能不能做消	手术:射频消融,状态:医生建议
融啊?"	检查:心电图,状态:患者已做
医生: "需要考虑。"	
患者: "我这胸口老是一阵一阵的疼"	

为此,我们采用滑动窗口的方式进行标注。具体地,我们定义了4个主要类别,称为大类,对于每个大类,分别定义了更具体的子类,以及它们相应的状态。在医学领域,状态信息是非常重要的。例如,某个症状的阴性或阳性对于特定的诊断会起到非常关键的作用。状态信息对于各个大类不尽相同,例如,疾病的状态有阴性、阳性和未知等,检查的状态则有患者已做和医生建议等。具体的标注范围在表 5-2 中列出。

表 5-2: 对话标注的具体内容

大类	子类	状态
症状	背痛, 出汗, 呃逆, 恶心, 发绀, 发热, 乏力, 腹部不	阳性, 阴性, 医生诊
	适,感冒,高血糖,高血压,高血脂,冠心病,行动不	断有, 医生诊断无,
	便,呼吸困难,甲亢,僵硬,咳嗽,流涕,呕吐,贫血,	未知
	水肿, 糖尿病, 头痛, 头晕, 胃部不适, 先天性心脏	
	病,心慌,心肌病,心肌梗死,心肌缺血,心肌炎,心	
	绞痛, 心律不齐, 心衰, 心脏肥大, 房颤, 房间隔缺	
	损, 室间隔缺损, 胸闷, 胸痛, 休克, 晕厥, 早搏, 战	
	栗抽搐	
检查	B 超, CT, CTA, 彩超, 测血压, 超声, 核磁共振, 甲	患者已做, 患者未
	状腺功能, 平板, 肾功能, 体检, 心电图, 心肌酶, 胸	做, 医生建议, 医生
	片, 血常规, 造影	不建议, 未知
手术	介入, 射频消融, 搭桥, 支架	患者已做, 患者未
		做, 医生建议, 医生
		不建议, 未知
一般信息	睡眠, 饮食, 精神状态, 大小便, 吸烟, 饮酒	正常, 异常, 未知

标注是以窗口为单位的,这是因为如果以整段对话作为输入的话,过长的对话内容会给标注带来很大困难。注意窗口的信息可以通过一定的

更新规则得到整段对话的信息,和对话状态跟踪<sup>[36,37]</sup>所做的工作比较相似。我们采用的滑动窗口大小为 5,滑动步长为 1。我们请 3 名研究生来进行标注,并特别邀请了一位医生作为标注指导。生成的窗口被随机分配给 3 名标注人员。

最终,我们标注了 1120 段对话,共 18212 个窗口。我们按照 800/160/160 的比例划分对话为训练集、开发集和测试集。标注共获得了 46151 个标签,每个窗口平均有 2.53 个标签,每段对话平均有 41.21 个标签。具体的标签分布情况见表 5-3。

	对话	窗口	症状	检查	手术	一般信息
训练集	800	12931	21420	8879	839	1363
开发集	160	2587	4254	1680	119	259
测试集	160	2694	4878	1869	264	327
共计	1120	18212	30552	12428	1222	1949

表 5-3: 标签分布统计

同时,我们还请所有标注人员同时标注抽样的 100 个窗口,以进行标注一致性检验。标注的一致性系数(cohen's kappa)达到了 0.91,说明本文的标注方式具有很强的可行性。

我们采用传统的信息抽取评价指标来进行评价,即准确率 P, 召回率 R, 和 F 值。由于在真实应用场景中,系统是针对整段对话进行抽取的,所以我们除了评价窗口级别的指标外,还评价了对话级别的指标。为了进一步探索模型的表现,我们还对不同的粒度进行了评价,即只看大类粒度,细化到子类粒度以及全部子类及其状态上的表现。

### 5.2.2 模型

在这一小节,我们将详细介绍本文新提出的方法。因为大类、子类及 其状态是预先定义好的,本文采用一个深度匹配模型来从医患对话中抽 取信息。深度匹配模型在自然语言处理领域的多个任务中都得到了成功 的应用,例如机器阅读理解等<sup>[80,81,82]</sup>。和分类模型相比,匹配模型可以利 用候选端更多的信息。

我们定义一个窗口中的句子为 n 个句子, $\{U(1), ... U(n)\}$ ,这里 $U(i) \in$ 

 $R^{l \times d_{emb}}$ 。我们提出的方法框架如图 5-1 所示,共有 4 个主要模块,下面将分别进行详细介绍。

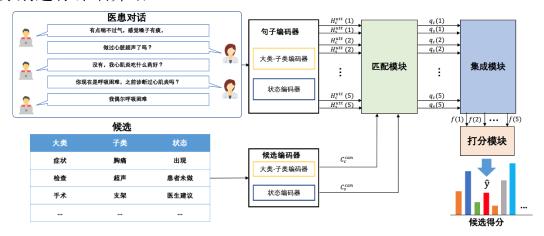


图 5-1: 模型框架图

#### 1) 编码模块

编码模块主要利用了带有注意力机制的双向长短时记忆(LSTM)模型<sup>[83]</sup>。假设输入为 $X \in R^{l \times d_{emb}}$ ,编码器将进行如下工作:

$$H = BiLSTM(X) \in R^{l \times d_{rnn}}$$

$$a_i = WH_i + b \in R$$

$$p = softmax(a) \in R^{(l \times l)}$$

$$c = \sum_i p_i H_i \in R^l$$

为了表示方便,我们可以采用如下表示,H,c = encoder(X)。H包括了输入序列中每个字符的上下文表示,而 c 是一个单独的压缩表示,是整个序列信息的加权表示。

对于句子 U,大类-子类 V 和状态 S 都可以通过这个编码器进行表示:

$$H_c^{utt}(i), C_c^{utt}(i) = encoder_c^{utt}(U(i))$$
 $H_s^{utt}(i), C_s^{utt}(i) = encoder_s^{utt}(U(i))$ 
 $H_c^{can}, c_c^{can} = encoder_c^{can}(V)$ 
 $H_s^{can}, c_s^{can} = encoder_s^{can}(S)$ 

这里上角标 utt 和 can 分别表示输入句子端编码器以及候选端编码器,下角标 c 和 s 则分别表示类别编码器和状态编码器。U, V, S 则分别由不同参数的编码器进行编码操作。

#### 2) 匹配模块

在这个模块中,大类-子类表示被视为一个查询(query),用来计算其与原始句子的注意力。

$$a_c(i,j) = c_c^{can} H_c^{utt}(i,j) \in R$$

$$p_c(i) = softmax(a_c(i)) \in R^l$$

$$q_c(i) = \sum_i p_c(i,j) H_c^{utt}(i,j) \in R^{d_{rnn}}$$

同时,状态表示则被视为另一个查询,也用来计算其与原始句子的注意力。

$$a_{s}(j) = c_{s}^{can} H_{s}^{utt}(i, j) \in R$$

$$p_{s}(i) = softmax(a_{s}(i)) \in R^{l}$$

$$q_{s}(i) = \sum_{i} p_{s}(i, j) H_{s}^{utt}(i, j) \in R^{d_{rnn}}$$

这里,(i,j)表示第i个句子里的第j个词。这个步骤的主要目的是,针对每个候选,捕捉句子中最相关的信息。

#### 3) 集成模块

为了判断一个候选是否在对话窗口中被提到,我们需要同时获取大类-子类信息和它们的状态信息。特别地,我们需要匹配每个大类-子类向量 $q_c(i)$ 和 $q_s$ 。

在多数情况下,大类-子类信息和它们的状态信息在一个句子中出现,但是在对话中,也有在不同句子中出现的情况,特别是在医患问答的场景下。所以,我们的方法必须要考虑不同话轮句子之间的关系。基于这个想法,我们使用了两种策略。

第一个策略针对大类-子类信息和状态信息出现在同一个句子中的情况。此时U(i)的表示是将 $q_c(i)$ 和 $q_s(i)$ 简单拼接。

$$f(i) = concat(q_c(i), q_s(i)) \in R^{2 \times d_{rnn}}$$

f(i)包含了大类-子类信息和状态,可以预测其和候选的得分。

第二个策略则考虑了句间关系,为了获得相关的状态信息 $q_c(i)$ ,我们把 $q_c(i)$ 作为一个查询,去获取它对于状态表示 $q_s$ 的注意力,具体过程如下:

$$a = q_c W q_s^T \in R^{l \times l}$$

$$p = softmax(a, 1) \in R^l$$

$$\widetilde{q_s} = pq_s \in R^{l \times d_{rnn}}$$

$$f(i) = concat(q_c(i), \widetilde{q_s}(i)) \in R^{2 \times d_{rnn}}$$

其中,W是可更新的的随机初始化的矩阵,形状为 $R^{d_{rnn}\times d_{rnn}}$ 。考虑到句子在对话中的位置是一个重要信息,我们在 $q_c$ 和 $q_s$ 的头部引入位置嵌入表示。

#### 4) 打分模块

集成模块的输出是打分模块的输入。我们使用每个句子的特征 f(i)来为候选打分,并且考虑得分最高的那个句子做为整个窗口对于该候选的打分。

$$s^{utt}(i) = feedforward(f(i))$$
  
 $y = sigmoid(max(s^{utt}))$ 

这种简单的策略合理性在于,如果窗口内的某个句子对于该候选的得分很高,那么这个对话窗口很有可能提到了这个候选。

#### 5) 学习

模型的损失函数可以用交叉熵损失表示:

$$L = \frac{1}{kl} \sum_{k} \sum_{l} -y_{l}^{k} \log\left(\widehat{y_{l}^{k}}\right) + \left(1 - y_{l}^{k}\right) \log\left(1 - \widehat{y_{l}^{k}}\right)$$

其中,下标k表示第k个训练样本,下标l表示第l个候选。

#### 6) 预测

由于任务被定义为匹配任务, 所以可以有多个预测结果。具体地, 在

预测过程中,我们的方法保留了所有得分超过 0.5 的候选作为答案。需要注意的是,在预测时同样需要把对话切分为窗口。我们可以用简单的规则把多个窗口的结果合并为对话级别的窗口,即把所有的窗口集合结果统一到一个集合中,状态则采用较新窗口的结果。

#### 5.3 实验

#### 5.3.1 实验设置

我们采用固定的 300 维的向量<sup>[84]</sup>表示中文字符,使用 Adam<sup>[85]</sup>优化器来进行模型优化。BiLSTM 的隐藏层大小为 400.前馈神经网络的层数为 4 层。同时,前馈神经网络模型采用了 dropout<sup>[86]</sup>技术,比例设为 0.2。我们通过开发集上的 F 值来确定训练停止时机。

#### 5.3.2 基线模型

我们采用了如下几个基线模型作为实验对比。

基线 Baseline:使用最简单的分类模型来做这个任务。输入为窗口内的所有句子的拼接,同样采用 BiLSTM 来获取语义表示,然后使用自注意力机制来获取单一向量表示。然后通过一个前馈神经网络分类器。输出的标签包括了所有的可能候选。

分类器 Classifier: 为了和更强的基线模型对比,我们复用了本文提出的框架,实现一个更先进的分类模型。和本文提出方法的主要区别是, $q_c$ 和 $q_s$ 的获取方式不同,在这里,不再使用匹配模块,二是直接用 $c_c^{utt}$ 和 $c_s^{utt}$ 。

# 5.3.3 实验结果

根据实验设定,本文的实验结果如表 5-4 所示。其中,标注 single 的表示只考虑单个句子,而标注 multi 的表示考虑句间的相互作用。

	窗口级别									:	对话级别							
模型		大类			子类			全部			大类			子类			全部	
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Baseline	67.21	63.78	64.92	60.89	49.2	53.81	53.13	49.46	50.69	93.57	89.49	90.96	83.42	73.76	77.29	61.34	52.65	56.08
Classifier-single	80.51	76.39	77.53	76.58	64.63	68.30	68.20	61.60	62.87	97.14	91.82	93.23	91.77	75.36	80.96	71.87	56.67	61.78
Classifier-multi	80.72	77.76	78.33	76.84	68.07	70.35	67.87	64.71	64.57	96.61	92.86	93.45	90.68	82.41	84.65	68.86	62.50	63.99
Ours-single	78.62	73.55	74.92	76.67	65.51	68.88	69.4	64.47	65.18	96.93	90.16	92.01	94.27	79.81	84.72	75.37	63.17	67.27
Ours	80.42	76.23	77.77	77.21	66.04	69.75	70.24	64.96	66.4	98.86	91.52	92.69	95.31	82.53	86.83	76.83	64.07	69.28

表 5-4: 实验结果

从实验结果中我们能得到如下结论:

- 本文提出的方法在全面评价上取得了最好的 F 值。F 值达到了 66.40 和 69.28,在这种复杂任务上是一个可以使用的结果。
- 考虑多轮之间关系的模型取得了更好的效果,这个结果也说明我们提出的方法可以有效捕捉丰富的对话特征,及其句子和候选的关系。
- 基于匹配的模型表现要优于分类模型。正如前文所述,我们认为这是因为匹配模型可以更多地利用候选端的信息。
- 本文提出的方法以及分类器方法都由于基线模型,这说明了本文 所提出框架的有效性。

#### 5.3.4 实例分析

在这一小节,我们将分析一些具体的例子来分析本文所提出方法的有效性。为了便于理解和说明,我们采用了可视化的表示方法。图 5-2 中的三个图分别表示了: a) 大类-子类对于各个句子的注意力热图,可以看出对于心脏病相关的词语注意力得分较高; b) 状态对于各个句子的注意力热图,可以看出其对于表示其状态的词语注意力程度更高; c) 第 4 个句子和其它各个句子的注意力热图,可以看出它和第 5 句的关联更大,这也是符合事实的。这几个例子都很好地说明了我们提出的方法在注意力捕捉方面是成功的。



图 5-2: 实例分析。(a)大类-子类对于各个句子的注意力热图; (b) 状态对于各个句子的注意力热图; (c) 第 4 个句子和其它各个句子的注意力热图。

# 5.4 本章小结

本章主要针对医疗对话中的整体信息抽取问题,提出一个新的数据集,包括从互联网上收集的已标注的1120段医患在线对话,共标注了超过40000个标签。模型方面,提出一个面向医患对话的对话医学信息抽取方法,采用深度匹配网络模型,其创新性在于可以考虑对话轮次之间的相互影响。我们提出的方法可以达到69.28的F值,实验结果表明本文提出的方法对于解决医疗对话信息抽取具有很好的研究前景。

# 6 总结与展望

本文主要面向对话形式的文本,以医疗对话作为实践领域,研究了背景知识图谱构建、实体抽取与关系分类,以及对话整体信息抽取问题,主要贡献总结如下:

首先,探讨医疗知识图谱的构建方法,并提出一种构建医疗知识图谱的方法,获得一个医疗症状图谱,作为后续工作的知识依托。

其次,提出面向对话文本的症状实体链接数据集,并使用基于预训练语言模型的 方法,得到症状实体链接工具。

最后,提出一个面向整体对话的信息抽取方法,提出相关数据集标注方法,并提出基于深度匹配的信息抽取模型,获得对话内提及的重要医疗信息。

综上,本文主要研究一系列针对医疗对话的文本理解方法,以识别实体、抽取信息为目标,通过在医疗领域的实践,提升针对对话文本的自然语言理解关键技术水平。

本文目前只对医疗对话理解中最基础的任务上做了探索,我们未来的研究工作包括更加全面的知识图谱构建,对话中的共指消解、篇章关系抽取等。期望以医疗领域的研究作为基础,扩展到更多领域,解决多领域的对话理解问题。

### 致谢

博士后的经历是一段难忘的时光,回到研究所熟悉的科研环境,尽管只有两年的时间,却收获满满。

在此出站报告完成之际,感谢在博士后期间帮助过我的老师和同学们。感谢我的博士后合作导师赵军研究员,您不仅教会了我做科学研究的方法,从您身上我学到了如何做人和做事。我还要感谢研究组的刘康研究员,无论是文章撰写还是科研项目研发,都少不了您悉心的指导。

感谢我的同事,何世柱副研究和陈玉博副研究员,同在一个办公室, 一起学习前沿技术,攻克项目难题,每天的朝夕相处让我在你们身上学习 到很多。感谢实验室助理李文婷,以及实验室的所有同学们。

感谢自动化所博士后主管部门的支持。感谢国家自然科学基金青年 基金的资助,以及精神疾病诊断与治疗北京市重点实验室开放研究课题 的资助。

最后,感谢我的家人们给予我的无条件支持,感谢你们的鼓励和陪伴。

# 参考文献

- [1] Hutchby, I., & Wooffitt, R. (2008). Conversation analysis. Polity.
- [2] 冉永平. (2000). 话语标记语的语用学研究综述. 外语研究, (4), 8-14.
- [3] Turing, A. M. (2009). Computing machinery and intelligence. In Parsing the Turing Test (pp. 23-65). Springer, Dordrecht.
- [4] Sarikaya, R., Crook, P. A., Marin, A., Jeong, M., Robichaud, J. P., Celikyilmaz, A., ... & Boies, D. (2016, December). An overview of end-to-end language understanding and dialog management for personal digital assistants. In 2016 IEEE Spoken Language Technology Workshop (SLT) (pp. 391-397). IEEE.
- [5] Chen, Y. N., Celikyilmaz, A., & Hakkani-Tür, D. (2017). Deep learning for dialogue systems. Proceedings of ACL 2017, Tutorial Abstracts, 8-14.
- [6] Rajpurkar P, Zhang J, Lopyrev K, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv:1606.05250, 2016.
- [7] Richardson M, Burges C J C, Renshaw E. Mctest: A challenge dataset for the open-domain machine comprehension of text. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 193-203.
- [8] Liu Z, Lim J H H, Sahimi N F A B, et al. Fast Prototyping a Dialogue Comprehension System for Nurse-Patient Conversations on Symptom Monitoring. arXiv preprint arXiv:1903.03530, 2019.
- [9] Liu Z, Chen N. Reading Turn by Turn: Hierarchical Attention Architecture for Spoken Dialogue Comprehension. Proceedings of the 57th Conference of the Association for Computational Linguistics. 2019: 5460-5466.
- [10] https://www.dellemc.com/en-us/index.htm
- [11] https://www.idc.com
- [12]http://www.cac.gov.cn/2018-09/15/c\_1123432498.htm
- [13]ICD-10 Classifications of Mental and Behavioural Disorder: Clinical Descriptions and Disgnostic Guidelines. Geneva. World Health Organisation. 1992.

- [14] Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl\_1), D267-D270.
- [15]Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics, 121, 279.
- [16] Barnes, T., Harris, E. E., Holloman, M., Arastu, H. H., Leinweber, C., Ju, A. W., & Naves, J. (2017). A quality verification tool to assure complete pre-treatment electronic medical records (EMR) for patients undergoing radiation therapy.
- [17] http://www.omaha.org.cn/
- [18] http://openkg.cn/
- [19]http://zstp.pcl.ac.cn:5050/
- [20] Lipscomb, C. E. (2000). Medical subject headings (MeSH). Bulletin of the Medical Library Association, 88(3), 265.
- [21]侯丽, 钱庆, 黄利辉, 等. 基于本体的临床医学知识库系统构建探讨, 医学信息学杂志, 2011.
- [22]王昊奋, 张金康, 程小军. 中文开放链接医疗数据的构建., 中国数字医学, 2013.
- [23] Finley, G., Edwards, E., Robinson, A., Brenndoerfer, M., Sadoughi, N., Fone, J., ... & Suendermann-Oeft, D. (2018). An automated medical scribe for documenting clinical encounters. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 11-15).
- [24] Finley, G., Edwards, E., Robinson, A., Brenndoerfer, M., Sadoughi, N., Fone, J., ... & Suendermann-Oeft, D. (2018). An automated medical scribe for documenting clinical encounters. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 11-15).
- [25] Mayfield, E., Laws, M. B., Wilson, I. B., & Penstein Rosé, C. (2013). Automating annotation of information-giving for analysis of clinical conversation. Journal of the American Medical Informatics Association, 21(e1), e122-e128.
- [26] Wang, N., Song, Y., & Xia, F. (2018). Coding Structures and Actions with the COSTA Scheme in Medical Conversations. In Proceedings of the BioNLP 2018 workshop (pp. 76-86).

- [27] Sarikaya, R., Crook, P. A., Marin, A., Jeong, M., Robichaud, J. P., Celikyilmaz, A., ... & Boies, D. (2016, December). An overview of end-to-end language understanding and dialog management for personal digital assistants. In 2016 IEEE Spoken Language Technology Workshop (SLT) (pp. 391-397). IEEE.
- [28] Chen, Y. N., Celikyilmaz, A., & Hakkani-Tür, D. (2017). Deep learning for dialogue systems. Proceedings of ACL 2017, Tutorial Abstracts, 8-14.
- [29] J. D. Williams. A belief tracking challenge task for spoken dialog systems. In NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data, pages 23–24, 2012.
- [30] Yoshino, K., Hiraoka, T., Neubig, G., & Nakamura, S. (2016, January). Dialogue state tracking using long short term memory neural networks. In Proceedings of Seventh International Workshop on Spoken Dialog Systems (pp. 1-8).
- [31] Sun, K., Chen, L., Zhu, S., & Yu, K. (2014, December). A generalized rule based tracker for dialogue state tracking. In 2014 IEEE Spoken Language Technology Workshop (SLT) (pp. 330-335). IEEE.
- [32] D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. A form-based dialogue manager for spoken language applications. In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, volume 2, pages 701–704. IEEE, 1996.
- [33] J. D. Williams. Web-style ranking and slu combination for dialog state tracking. In SIGDIAL Conference, pages 282–291, 2014.
- [34] S. Young, M. Gai, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. 24(2):150–174, 2010.
- [35] J. D. Williams. A belief tracking challenge task for spoken dialog systems. In NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data, pages 23–24, 2012.
- [36] Z. Wang and O. Lemon. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In SIGDIAL

- Conference, pages 423–432, 2013.
- [37]S. Lee and M. Eskenazi. Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In SIGDIAL Conference, pages 414–422, 2013.
- [38] J. Williams. Multi-domain learning and generalization in dialog state tracking. In SIGDIAL Conference, pages 433–441, 2013.
- [39] M. Henderson, B. Thomson, and S. Young. Deep neural network approach for the dialog state tracking challenge. In Proceedings of the SIGDIAL 2013 Conference, pages 467–471, 2013.
- [40] Mrkšić, N., Séaghdha, D. O., Thomson, B., Gašić, M., Su, P. H., Vandyke, D, Young, S. Multi-domain dialog state tracking using recurrent neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 794–799, Beijing, China, July 2015. Association for Computational Linguistics.
- [41] Mrkšić, N., Séaghdha, D. O., Wen, T. H., Thomson, B., & Young, S. (2016). Neural belief tracker: Data-driven dialogue state tracking. arXiv preprint arXiv:1606.03777.
- [42]Zhong, V., Xiong, C., & Socher, R. (2018). Global-locally self-attentive dialogue state tracker. arXiv preprint arXiv:1805.09655.
- [43] Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., & Cardie, C. (2019). DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension. Transactions of the Association for Computational Linguistics, 7, 217-231.
- [44] Liu Z, Lim J H H, Sahimi N F A B, et al. Fast Prototyping a Dialogue Comprehension System for Nurse-Patient Conversations on Symptom Monitoring. arXiv preprint arXiv:1903.03530, 2019.
- [45] Liu Z, Chen N. Reading Turn by Turn: Hierarchical Attention Architecture for Spoken Dialogue Comprehension. Proceedings of the 57th Conference of the Association for Computational Linguistics. 2019: 5460-5466.
- [46] Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on

- concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5), 552-556.
- [47] Weed L L. Medical records that guide and teach. New England Journal of Medicine, 1968, 278(12): 593–600.
- [48] 杨锦锋, 于秋滨, 关毅, & 蒋志鹏. (2014). 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 40(8), 1537-1562.
- [49] Friedman C, Alderson P O, Austin J, Cimino J J, Johnson S B. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association, 1994, 1(2): 161–174.
- [50] Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen P C. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. Journal of biomedical informatics, 2009, 42(5): 937–949.
- [51] Savova G K, Masanz J, Ogren P V, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen Piet C. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Information Association, 2010, 17(5): 507–13.
- [52] 叶枫, 陈莺莺, 周根贵, 李昊旻, 李莹. 电子病历中命名实体的智能 识别. 中国生物医学工程学报, 2011, 30(2): 256-262.
- [53] de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at I2B2 2010. Journal of the American Medical Informatics Association, 2011, 18(5): 557–562.
- [54] Rajani, N. F., Bornea, M., & Barker, K.(2017). Stacking with auxiliary features for entity linking in the medical domain. BioNLP 2017, 39-47.
- [55] Ling, Y., Hasan, S. A., Datla, V., Qadir, A., Lee, K., Liu, J., & Farri, O. (2017, November). Diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning: A preliminary study. In Machine Learning for Healthcare Conference (pp. 271-285).
- [56] Palotti, J., & Hanbury, A. (2015). TUW@ TREC clinical decision support track 2015.

- Vienna University of Technology Vienna Austria.
- [57] Du, N.; Chen, K.; Kannan, A.; Tran, L.; Chen, Y.; and Shafran, I. 2019. Extracting symptoms and their status from clinical conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 915–925. Florence, Italy: Association for Computational Linguistics.
- [58] http://www.a-hospital.com/
- [59] https://www.120ask.com/
- [60]http://3g.xywy.com/
- [61] http://www.99jk.com.cn/
- [62]http://www.39.net/
- [63] https://www.fh21.com.cn/
- [64] Hahn, U., Romacker, M., & Schulz, S. (2002). MEDSYNDIKATE—a natural language system for the extraction of medical information from findings reports. International journal of medical informatics, 67(1-3), 63-74.
- [65] Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010).
  MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association, 17(1), 19-24.
- [66] Mykowiecka, A., Marciniak, M., & Kupść, A. (2009). Rule-based information extraction from patients' clinical data. Journal of biomedical informatics, 42(5), 923-936.
- [67] Jagannathan, V., Mullett, C. J., Arbogast, J. G., Halbritter, K. A., Yellapragada, D., Regulapati, S., & Bandaru, P. (2009). Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. International journal of medical informatics, 78(4), 284-291.
- [68] Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M., & Xu, H. (2014). A comprehensive study of named entity recognition in Chinese clinical text. Journal of the American Medical Informatics Association, 21(5), 808-814.
- [69] https://www.chunyuyisheng.com/
- [70] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, January). BERT: Pre-training

- of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1).
- [71] Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., ... & Sun, M. (2019, July). DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 764-777).
- [72] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
- [73] Zheng, J. G., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J., & Ji, H. (2015). Entity linking for biomedical literature. BMC medical informatics and decision making, 15(S1), S4.
- [74] Wang, M., Zhang, J., Liu, J., Hu, W., Wang, S., Li, X., & Liu, W. (2017, October). Pdd graph: Bridging electronic medical records and biomedical knowledge graphs via entity linking. In International Semantic Web Conference (pp. 219-227). Springer, Cham.
- [75]赵亚辉. (2017). 临床医疗实体链接方法研究 (Master's thesis, 哈尔滨工业大学).
- [76] https://github.com/ymcui/Chinese-BERT-wwm
- [77] Wachter, R., and Goldsmith, J. 2018. To combat physician burnout and improve care, fix the electronic health record. Harvard Business Review.
- [78] Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.
- [79] Ma, X., & Hovy, E. (2016, August). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1064-1074).
- [80] Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.
- [81] Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541.
- [82] zhu, C.; Zeng, M.; and Huang, X. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. arXiv preprint arXiv:1812.03593.

- [83] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
- [84] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 3111–3119.
- [85] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.
- [86] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

# 个人简历

张元哲,2011年本科毕业于北京航空航天大学,同年9月保送至中国科学院自动化研究所硕博连读攻读博士学位,导师是赵军研究员,2016年获工学博士学位,专业为模式识别与智能系统。2016-2018曾任百度自然语言处理算法工程师,2018年8月在中国科学院自动化所模式识别国家重点实验室作博士后,研究方向为机器阅读理解、知识图谱和自然语言处理。在ACL、AAAI、EMNLP等项级国际会议和重要学术期刊上发表论文多篇。主持国家自然科学基金青年科学基金:医疗对话文本中的信息抽取关键技术研究。参与国家自然科学基金重点项目、973计划、863计划等多个科研项目。参与研发的基于知识图谱的问答、对话篇章理解等工具和软件在华为、百度、云知声等多家单位得到应用。面向医疗对话文本的信息抽取系统获得2019全国博士后人工智能发展与应用论坛特等奖。作为核心骨干研发的"大规模开放域文本知识获取与应用平台"项目获得2019年度北京市科学技术进步奖一等奖(已公示)。

# 联系方式:

电子邮件: yzzhang@nlpr.ia.ac.cn

通讯地址:北京市海淀区中关村东路95号智能化大厦7层,100190

电话: 010-82544736

# 论文发表情况

#### ■ 博士后在站期间发表的论文

- Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao,
   Shengping Liu, Kang Liu, Jun Zhao. MIE: A Medical Information
   Extractor towards Medical Dialogues. In ACL, 2020.
- Zhixing Tian, Yuanzhe Zhang, Xinwei Feng, Wenbin Jiang, Yajuan Lyu, Kang Liu, Jun Zhao. Capturing Sentence Relations for Answer Sentence Selection with Multi-Perspective Graph Encoding. In AAAI, 2020.
- Delai Qiu, Yuanzhe Zhang, Xinwei Feng, Xiangwen Liao, Wenbin Jiang, Yajuan Lyu, Kang Liu, Jun Zhao. Machine Reading Comprehension Using Structural Knowledge Graph-aware Network. In EMNLP, 2019.
- Delai Qiu, Liang Bao, Zhixing Tian, Yuanzhe Zhang, Kang Liu, Jun Zhao, Xiangwen Liao. Reconstructed Option Rereading Network for Opinion Questions Reading Comprehension. In CCL, 2019 (Best Paper Award).

#### ■ 博士生期间发表的主要论文

- **Yuanzhe Zhang**, Shizhu He, Kang Liu, Jun Zhao. A Joint Model for Question Answering over Multiple Knowledge Bases. In AAAI, 2016.
- **Yuanzhe Zhang**, Xuepeng Wang, Shizhu He, Kang Liu, Jun Zhao, Xueqiang Lv. IAMA Results for OAEI 2013. In ISWC, 2013.
- Yuanzhe Zhang, Xuepeng Wang, Siwei Lai, Shizhu He, Kang Liu, Jun Zhao, Xueqiang Lv. Ontology Matching withWord Embeddings. In CCL, 2014.
- Yanchao Hao; Yuanzhe Zhang; Kang Liu; Shizhu He; Zhanyi Liu; Hua
   Wu; Jun Zhao. An End-to-End Model for Question Answering over

- Knowledge Base with Cross-Attention Combining Global Knowledge. In ACL, 2017.
- Jun Zhao, Kang Liu, Shizhu He, and **Yuanzhe Zhang**, Question Answering over Knowledge Bases, in IEEE Intelligent Systems, 2015.
- Shizhu He, Kang Liu, **Yuanzhe Zhang** and Jun Zhao, Questioning Answering over Linked Data Using First-order Logic, in EMNLP, 2014.
- Shizhu He, **Yuanzhe Zhang**, Kang Liu and Jun Zhao, CASIA@V2: A MLN-based Question Answering System over Linked Data, in Proceedings of the fourth workshop on question answering over linked data (QALD-4) in CLEF, 2014.
- 王雪鹏,**张元哲**,刘康,徐立恒,赵军,刘树林,吕学强.基于网络语义标签的多源知识库实体对齐算法,In CCIR 2014 (优秀论文).