基于深度神经网络的自动文本摘要研究

近年来，自动文本摘要已经成为了人工智能和自然语言处理领域的重要研究方向之一。自动文本摘要旨在提取出原始文本中的关键信息，并生成一段语义通顺且简洁准确的摘要，其目的是为了提高用户浏览信息的效率。随着深度学习的发展，当今的自动文本摘要模型主要基于序列到序列框架构建。然而，目前序列到序列框架在自动文本摘要中的应用也存在着诸多问题，例如集外词生成困难、无法有效地对单词之间的联系进行建模、缺乏对关键信息提取过程的建模等。针对这些问题，本文对基于序列到序列框架的自动文本摘要模型进行了改进，主要研究内容总结为以下两点：

（1）提出了基于改进子词单元的生成式文本摘要方法。该方法通过改进的子词分割算法将一个完整的单词分割成不相交的子词单元并构建词汇表，这样不仅可以减小词汇表的规模，而且单词被分割为子词单元后，具有相同含义但形态不同的单词会体现更强的关联性，例如受单复数、时态等因素影响的单词。此外，由于集外词可以通过不同的子词单元组成，所以该算法也可以有效地缓解集外词难以生成的问题。我们在文本摘要数据集Ｇｉｇａｗｏｒｄ、ＣＮＮ／ＤａｉｌｙＭａｉｌ和ＸＳｕｍ上进行了实验，实验结果验证了该方法能对单词之间的联系进行更好的建模并且能够缓解集外词难以生成的问题。

（2）提出了基于分层信息过滤的生成式文本摘要方法。该方法使用动态路由算法基于编码器的输出动态地计算全局向量，然后使用全局向量指导分层信息过滤算法从两个层面对输入文本中的噪声进行过滤：词层面和语义层面。首先，我们使用全局向量和编码器的输出计算出输入文本中每个词的权重，根据权重来选择输入文本中的关键词。然后，我们使用双门单元对输入文本中的语义噪声进行过滤。具体来说，双门单元包含两个门：过滤门和补充门，过滤门对输入文本中的语义噪声进行初步的过滤，补充门将一部分原始信息添加到被过滤的文本表示当中形成最终的文本表示，这样可以避免输入文本中的信息被过度过滤的问题。我们在文本摘要数据集Ｇｉｇａｗｏｒｄ和ＣＮＮ／Ｄａｉｌｙ　Ｍａｉｌ上进行了实验，实验结果验证了我们提出的方法在噪声过滤上的有效性。

关键词：自动文本摘要；子词单元；噪声过滤；深度神经网络；自然语言处理

基于预训练模型及强化学习的自动文本摘要研究

自人类社会进入信息化时代以来，自动文本摘要技术就成为研究者重点关注的研究方向。这项技术旨在对长文本进行压缩、提炼，在保证主旨语义不变的情况下，形成更简洁、凝练的短文本。随着深度学习自然语言处理等领域的迅速兴起，将深度学习与文本摘要技术相结合的研究方法方兴未艾。以往的工作仅仅将两者进行简单结合，存在文本词嵌入表示不够精确、泛化性能差以及训练成本高等问题，从而限制了文本摘要的性能提升。本文针对所存在的问题，从特征融合和模型迁移微调两方面将预训练模型的相关研究内容应用于文本摘要任务，并将抽取式与生成式这两种摘要范式相结合来提升处理长文本的能力，通过大量实验证明，本文所提出的方法在性能上均优于基线模型。本文的主要工作包括如下两部分：

（1）提出基于 BERT 词嵌入及强化学习的文本摘要生成模型

针对文本嵌入表示不够精细的问题，将预训练模型 BERT 的词嵌入表示应用于文本摘要任务，利用 BERT 在海量数据上训练得到的丰富语义特征，来提高摘要任务的性能。在训练过程中，借助强化学习桥接两个子模型来训练统一的摘要模型。一方面该方法在抽取子模型的训练过程中使用自注意力机制来获取文档向量，而且在抽取过程中考虑了语句信息的显著性以及冗余度等问题。另一方面在生成子模型训练过程中，进行句子级别注意力与单词级别注意力的不一致性计算，提升了生成子模型的性能。在公开标准数据集上，大量的实验结果证明该模型取得了目前最好的生成式摘要性能。此外，对比实验证明了预训练模型的词嵌入以及强化学习在模型中的有效性。

（2）提出基于预训练模型微调的混合式文本摘要方法

针对泛化性能差以及训练成本高的问题，该模型使用基于预训练模型微调的方式对文本摘要采取先抽取后生成的方法，充分利用抽取式摘要与生成式摘要两种方法的优点。一方面在训练抽取模型时使用预训练模型 RoBERTa 微调，然后连接抽取分类器，该分类器设定抽取概率的阈值，然后对文档中的句子进行二

分类预测。另一方面在训练生成模型时使用编码器-解码器结构的T5 预训练模型微调，对前面的抽取语句进行摘要生成，在保持语义不变的前提下，生成更简短凝练的摘要。在多个公开数据集上，模型的优良性能得到了实验数据的验证支持。

关键词：预训练模型，强化学习，文本摘要

深度学习在文本特征提取中的研究与应用

文本数据作为最常见的数据形式之一，涵盖范围广、数据量大，又有密度不均的特点，不同平台中存储的文本数据结构不定相同，结构化文本数据、半结构化文本数据和非结构化文本数据并存。文本挖掘工作便是从这些海量、密度不均、异构的文本数据中提取隐含的、有价值信息以供决策或预测等使用。特征提取作为文本挖掘工作中不可缺少的关键步骤，其提取出的特征项的优劣会直接影响文本挖掘后续工作，将深度学习人工神经网络模型用于特征提取步骤不仅可以提高特征提取工作效率，借助复杂模型还可以提升特征项的质量，以更好反映文本数据特性。本文主要工作如下:

首先对现阶段常用的传统特征提取有关技术进行阐述，其次阐述了深度学习相关理论和自然语言处理部分技术，分析了深度学习在特征提取方面可能存在的优化效果，以及分析了深度学习人工神经网络模型RNN、LSTM 在文本特征提取方面存在的局限性,最后针对前述问题提出一种结合 BERT 预训练模型和 HAN 网络的特征提取模型，从文本数据预处理方面与层次化分析文本数据视角，进行文本特征提取工作，通过设置五组模型在中英各两个数据集上进行二分类的对比实验，验证本文提出的特征提取模型的有效性及一定的优越性，实验结果表明本文提出的特征提取模型在对上下文信息的学习上、语义学习、词的多层特征学习上均有良好表现，同时对长距离依赖问题有着一定弱化效果。

关键词：文本特征提取　深度学习　BERT 预训练模型　HAN 网络

基于 BERT 模型的垃圾分选专利分析与研究

伴随着知识产权逐步成为国家、地区及企业在知识经济时代背景下持续发展的战略性资源与核心竞争力。专利文献数量与日俱增，如何针对大量专利文献进行快速阅读及分析也成为目前诸多专家学者的研究方向。

近年来，由于深度学习的不断发展，自然语言处理技术（NLP）得到了全新的技术突破，语言训练模型在对于处理非结构化数据尤其是文本处理及分析的方面愈加完善。对于专利文本领域而言，将自然语言处理技术应用于专利文本数据的处理与分析中是值得探究的。

本文对如何使用自然语言处理技术（NLP）对专利文本进行处理及分析进行了研究。首先，本文详细阐述了专利分析方法、自然语言处理技术、语言训练模型以及文本自动生成相关技术及方法。其次，借鉴文本挖掘与自动文本摘要的思想和技术，本文提出一种基于大规模语言训练模型 BERT 与 LSTM 神经网络相结合的方法实现专利文献辅助文档的自动生成。本设计整体采用 Encoder-Attention-Decoder 架构，其中 Encoder 层使用大规模语言预训练模型 BERT 实现专利文本的特征提取及特征向量化，解决了传统的非 Transformer 语言模型中难以解决的一词多义等问题，并能够提取到更深层次的语言特征信息，得到更细粒度的文本编码表示。Decoder 层使用双层 LSTM 神经网络作为文本序列输出模型，将模型理解后的数据息解码成为自然语言文本，实现文档的自动生成。并在编码器与解码器之间加入 Attention 机制，避免了语义信息的缺失。最终，本文以垃圾分选技术领域的专利作为本实验的研究对象，以垃圾分选专利的发明内容作为数据集进行实验研究。实验结果表明通过该模型生成的专利辅助文档具有一定的归纳总结作用，能够体现该类专利的共性、特点及技术组成。通过阅读辅助文档可大大减少人工阅读量，同时为专利阅读者及专利申请人提供了相关技术信息。

关键词：NLP　专利分析　BERT　LSTM 辅助文档生成

基于融合特征表示的癌症研究趋势分析算法研究

癌症研究在生命科学和医学领域中至关重要，许多国家和组织每年花费在癌症研究的经费高达数十亿美元。随着大数据时代的到来，先进的设备和技术对于研究人员已是触手可及，但是随着海量信息的增长，医生及生物研究人员很难处理日益增长的信息过载问题。因此，随着生物医学研究论文的爆炸式增长，采用机器学习访求可以帮助研究者们在庞大的文本数据资源中快速获取信息，使他们更有效的了解近年的癌症研究重点和发展历程以提升人类医疗的集体经验。

随着计算机技术的发展，自然语言处理的研究越来越火热，其在机器翻译、信息抽取、文本情感分析、自动问答和个性化推荐等领域都扮演着重要的角色。其中文本特征表示的相关研究是自然语言处理领域的重中之重。人类的文本语言逻辑无法被机器所识别，因此我们需要通过机器学习算法将文本转化为机器能够理解的形式。文本特征表示形式多种多样，一直以来是人们关注的重点，其对文本表示的准确性是下一步相关应用工作的基础。

本文将自然语言处理领域与医学领域相结合，利用融合的特征表示方法对癌症研究趋势进行分析。首先，本文对传统的文本特征表示模型进行了融合改进，并通过机器学习分类与聚类算法对不同文本特征表示模型得到的文本向量进行相关实验，进而衡量不同向量对文本信息的表示准确性。实验结果表明 Tr-W2v 算法得到的融合文本向量在分类实验中表现效果最佳，而 Ti-W2v 算法得到的融合文本向量在聚类实验中表现效果最佳，本文对于此结果结合相关算法特点也进行了直观的解释。其次，本文基于融合改进的文本向量进一步提出了相似度趋势分析、关键词趋势分析和改进的关键词趋势分析等多种癌症趋势分析模型。其中相似度趋势分析模型分析了近年中国五大高发癌症的相似度走向趋势。随后，以肺癌数据为例，本文提出的关键词趋势分析模型分析了肺癌整体的研究方向和区域。为进一步解决关键词趋势分析模型的不足，本文提出的改进的关键词趋势分析模型从肺癌相关的基因蛋白、肺癌相关的治疗药物和方法以及肺癌相关的其他热点等多角度进行了更细致的分析。根据本文得到的癌症趋势分析结果，医生及生物医学研究者可以从大量癌症研究论文中了解到每年的不同热点区域差异和相关的联系趋势。这可以在很大程度上减少相关人士阅读大量论文和追踪热点的工作量，并在一定程度上辅助指导他们快速搜集信息和进一步开展工作。

关键词： 特征表示，特征融合，自然语言处理，文本挖掘

基于神经网络的智能医疗诊断研究

随着社会发展越来越快，人们的生活节奏也紧随着社会的变化而变化。与此同时，人们对医疗健康问题的关注也随之递增。我国当前医疗问题主要为“医疗资源不平衡”，“看病难”，“看病贵”，“医疗误诊率高”等等，这些问题一直是我国医疗　　界的难题。随着信息技术地快速发展和进步，人工智能、云计算和互联网等新兴信息技术给人们带来了解决医疗难题的新思路。因此，数字医疗的概念已经得到了越来越多人的认可，其中医疗诊断是数字医疗的核心问题之一。医疗诊断的准确率及效率，与人们的生命健康息息相关，是一个不容忽视的重要问题。将信息科技运用于医疗诊断中，因此显得意义重大。

本论文将多种神经网络模型应用到医疗诊断中，包括全连接神经网络，卷积神经网络。通过获取的医疗数据，运用ｊｉｅｂａ分词工具和当前流行的数据预处理技术，对错综复杂的医疗数据进行整理，并使用ｐａｎｄａｓ或\Ｖｏｒｄ２Ｖｅｃ等工具将医疗数据对中文数据进行量化处理，将数据转换成ｏｎｅ－ｈｏｔ二元变量或是稠密向量等计算机算法可识　　别格式。利用量化后的数据，训练全连接神经网络模型，卷积神经网络模型和Ｗｏｒｄ２ＶｅＣ＋卷积神经网络模型，最后对比三个模型的医疗诊断准确率，同时与决策树模型进行比较和分析，并将准确率最高的模型部署到基于智能医疗诊断的医疗综合服务系统中。经过数据预处理得到量化数据，模型训练，最终结果显示W〇ｒｄ２ＶｅＣ＋卷积神经网络的准确率高于其他模型，约为８９％。基于神经网络的智能医疗诊断系统还有许多方面有待完善，例如预测准确率的提升，症状关联性改进，数据预处理方法等等，未来将从上述几个方面着手，进行更深入地探索。

关键词：全连接神经网络；卷积神经网络；自然语言处理；数据预处理

基于人工智能的肺癌辅助诊断系统的设计与实现

经调查研究得知，近年来在体检中发现肺癌的人数逐年增多，但是选择住院治疗的人数却比较少，大约７５％的患者在确诊时已经是肺癌中晚期，其治愈几率大大降低。因此，尽早诊断出肺癌对于降低因肺癌造成的死亡率具有重要意义。另一方面，面对大量的患者，临床医生有着相当大的工作量，需要先筛选肺癌患者，然后根据自身经验对患者进行相应的治疗。然而在我国，培养一位经验丰富的临床医生所需要花费的时间周期较长。综上所述，在医学领域，研究运用人工智能技术对医学数据分析、辅助临床医生进行肺癌筛选诊断和治疗，具有重要理论意义和实际应用价值。

本文在分析了目前深度学习方法之后发现，国内外研究者主要针对肺部ＣＴ图像进行肺癌预测，忽略了放射科医生给出的ＣＴ图像描述和检验报告，尤其是检验报告，这样会丢失部分信息。考虑上述问题，本文设计了一种新颖的文本和图像的多模态学习的肺癌辅助诊断方案。该方案与目前已有的方法不同，是基于ＣＴ图像、放射科医生给出的ＣＴ图像描述、检验报告三部分进行多模态融合。其主要实现要点是将图像部分先预处理，再利用Ｒｅｓｎｅｔ网络建模；将ＣＴ图像描述部分利用自然语言处理技术进行分词、预训练、建模；将检验报告利用多层感知机建模；最后经三部分融合。实验验证，基于文本和ＣＴ图像的多模态方法的准确率要比基于ＣＴ图像的单模态方法的准确率提高３％，这说明，ＣＴ影像仍是肺癌诊断的主要信息，而检查描述和检验结果作为补充信息加入到模型中，可以很好的提升模型的精确度。

再基于本文设计的文本和图像的多模态学习方式，设计实现了肺癌辅助诊断系统。充分将多模态方式与计算机辅助诊断融合，帮助临床医生筛选诊断患者。该系统能够快速实现数据预处理，肺癌辅助诊断判断，诊断信息录入，以及患者过往史病例等信息的查询。该系统的实现减轻了临床医生的工作量，提高了其工作效率，还为临床医生提供了一个全方位观察和诊断患者，同时，肺癌患者也能及时了解自身情况。

关键词：肺癌，卷积神经网络，自然语言处理，多模态学习

中文电子病历医学实体识别算法研究

电子病历是患者完整病程的数字化记录，对帮助医生分析病案和医疗决策具有重要意义。结构化电子病历由于选择复杂、限制医生思维以及病例高度重复等问题，慢慢被医生使用自然语言进行书写的非结构化和后结构化电子病历所取代。结构化的电子病历是医疗大数据分析的基础，因此，将自然语言书写的电子病历转化为具有一定规则的结构化数据是现在医学信息学研究的重要方向。深度学习方法的出现和使用也使针对电子病历的自然语言处理成为研究热点。本文研究基于深度学习的命名实体识别技术，可完成对医学文本中的实体名词进行识别与提取，从而达到电子病历的后结构化目的。

在命名实体识别任务中，词嵌入作为最重要的预训练方法，将上下文中的词语信息转化为数学空间中的向量。不同于英文从词语或句子级别的角度进行研究，中文词嵌入的研究重点在于挖掘中文词语与字符的内在偏旁部首与笔画信息。故本文提出一种融合词信息与子词信息作为词嵌入的模型，使用字符与笔画组合来构成子词信息部分，对比现有词嵌入方法结合更多词语的内在信息。通过外部评估的方式，在四种不同命名实体识别模型中进行测试，结果表明，本文提出的融合模型比仅使用单一词语作为词嵌入，在模型的F1 值评估指标上平均提高 1%。

由于中文电子病历命名实体识别的研究需要大量标注数据，而聘请医生和具有相应知识背景的专家来进行数据标注，在人力物力上耗费巨大且投入产出比极低。所以本文提出一种基于众包标注的医学实体识别模型，将众包标注的电子病历作为输入进行模型训练，利用对抗学习的思想降低众包之间的差异并提升模型的泛化能力，通过与其他对众包语料进行投票后的命名实体识别模型进行对比实验，F1 值有 2%-3%左右的提升，并且在准确率和召回率上也取得更好的效果。

本文基于 Dev Ops 的理念设计并开发电子病历标注系统，实现了 Web 端的电子病历的标注与医学术

语词典等应用，并在服务器端对应用服务、数据库及服务器等节点进行监控，运用 Docker 容器技术实现从代码提交、测试到服务部署的 CI/CD 流水线。最后通过 API 并发测试，通过监控模块对硬件及节点状态进行实时监控，并在负载达到预设压力时进行报警通知，从而验证了系统的稳定性。

关键词：电子病历，自然语言处理，命名实体识别，词嵌入，众包标注

基于医学挂号系统和问答匹配模型研究与实现

随着互联网的发展，网络数据井喷式增长，出现了很多医疗社区，越来越多人开始使在医疗问答社区进行信息获取和问诊。医疗社区利用在线医生提供就诊服务的同时，也会利用积累的数据提供信息检索服务，但主要都是基于搜索引擎的关键词匹配返还一堆相关问题和答案文档，不能深入理解用户问句的语义信息。问答系统因为能返回一个确定的答案而不是一堆仍需用户筛选的文档，成为了当下研究热点。

基于深度学习的问答系统是将答案确认过程看成一个问句和候选答案语义匹配的过程，目前研究主要有两个研究方向：利用多种网络组合来增强网络特征提取能力；用注意力机制识别问题意图，突出问题和答案的交互信息。但是利用多种组合的深度神经网络来提取问答句深层特征时，往往利用网络最后一层的特征作为句子特征表示，而没有有效利用中间各层提取的特征；在利用注意力机制捕获问题答案相关信息时，也是用最终的问句特征对答案特征加权筛选，而没有考虑中间每一相同层间问题、答案的交互信息。

考虑到去医院就诊挂号确定科室时，病人受限于自己掌握的医疗知识有时并不清楚自己需要挂的科室；同样，用户通过医疗社区进行咨询就诊时，也需要按科室类别（如内科，五官科，妇产科）选择医生，但有的用户并不能确定自己的病属于哪一科，据我所知，到目前为止还没有相关方面的研究。

针对以上问题，本文主要研究工作如下：

（1）本文首次将深度学习和自然语言处理技术应用于解决医疗挂号问题，将病人看病挂号选择科室的问题建模成问句分类问题。基于医疗社区积累的数据，本文利用 Bi-LSTM 结合注意力机制训练了一个分类器对病人问题分类。此外还用 LSI 文本相似度技术找到带标签数据集中最相似的问题，根据最相似问题的标签对分类器的结果进行校验。该挂号系统能方便用户线下就诊时确认挂号科室以及在网上医疗社区按科室找到自己需要的医生。

（2）本文提出了一种多层融合的层间交互的问答匹配模型，利用多层特征融合的深度神经网络和多层层间注意力机制来解决问答匹配问题。多层的 Bi-LSTM+CNN+CNN 每层提取的特征进行拼接融合作为最终的问题、答案语义特征用于进行问答匹配和答案打分；在每层对问题、答案特征提取时都加入注意力机制，来突出答案和问题相关的特征。相比起用深度神经网络最后一层特征，本模型综合利用每一层的特征，最终的表示特征语义信息更充分；在每一相同层间利用注意力机制有利用及时的捕获问答交互信息。

（3）针对中文利用词向量时需要进行分词、去停用词等复杂的处理，另外虽然现在中文分词工具比较成熟，但受限于医疗特定领域专业名词的存在，有些专业词汇分词识别不准会直接影响下游模型性能，而构建专业领域字典虽然能解决这一问题，但任务量大，并且移植性差，利用字向量不需要分词可避免上述问题。因此，本文在进行问答语义匹配时，除了用词向量进行模型训练的同时，也用字向量进行实验，研究字向量的有效性和可行性。

（4）利用医疗社区的数据，和上述挂号模型以及问答匹配模型，设计了一个医疗问答系统，来解决用户挂号问题，以及问答需求。其问答模块，相对于医疗社区信息检索服务是基于关键词匹配返回和用户问题相关的一堆问题和相应答案文档，本文设计的问答系统，是利用多层融合特征交互的问答匹配模型对问题和候选答案进行深度语义匹配后打分，返回一个确定答案。

关键词：深度学习；问答匹配；挂号系统；注意力机制；自然语言处理

面向医疗语义理解的结构化处理方法的研究与实现_黄璨

随着人工智能热潮的掀起，人工智能在医疗健康领域的应用场景也越来越丰富，人工智能技术影响着医疗行业的发展。在一些检查中，医生双手无法离开检查设备，亟需引入智能化的语音交互医疗产品来协助工作，提升工作效率。在智能化语音助手中语义理解引擎起着核心作用，医疗语义理解的含义是帮助语音助手进行理解医生的意图、提取医生说话内容的关键信息，并对获取的文本信息进行结构化处理，最终生成电子病历。

蓬勃发展的背后，人工智能在医疗领域的应用和推广也面临着诸多问题和挑战。目前针对中文自然语言的医疗文本结构化处理方法采用的方案是存在较多的弊端：灵活性不足、无法实现各种业务的定制化、容易丢失重要病历信息等等。针对以上存在的问题，

本文主要从以下几个方面开展工作：

本文基于科大讯飞有限公司智慧医疗内核部门“面向医疗语义理解引擎”项目，对语音转写文本进行结构化处理的研究。本文通过对语音转写文本数据以及需求进行深入的分析，给出了一个“规则+命名实体识别+知识库+分类”一体化的医疗文本结构化处理方案。

首先，针对目前传统信息提取技术应用于本课题中效果较差的问题，本文给出了一种基于规则和命名实体识别融合的信息抽取处理方法，该方法进行NLP(Natural Language Processing，自然语言处理)文法解析和命名实体识别的提取，并保留提取信息的并集。

其次，针对传统医疗文本结构化中知识图谱的应用只是实体间语义的简单拼接，结构化效果较差的问题，因此本文引入知识图谱校验思路。其方法是在医疗知识图谱构建完成后，对结构化系统中提取的语义信息进行合法性校验，包括值类型、值范围、以及语义关系等，以提高文本结构化的正确率。

然后，为防止文本中有用信息丢失，本文给出了一种基于CNN(Convolutional Neural Networks，卷积神经网络)分类模型对文本进行二分类的方法，并对其中 CNN 模型结构进行了改良。经过实验对比分析，最后选用 jieba 分词和 CNN 组合的方案对文本进行二分类。

最后，通过对前面三个主要部分的研究，设计和实现了面向语义理解的结构化处理方案。通过真实的现场语音转写文本数据验证改进后的系统在结构化效果以及分类效果有较大提升。

关键词：自然语言处理；语义理解；神经网络模型；文本分类；文本结构化

基于中文自然语言处理的糖尿病知识图谱构建

随着人民生活水平的提高以及生活方式的改变，糖尿病的发病率在逐年增加，糖尿病会导致持续高血糖与长期代谢紊乱等问题，从而致使全身组织器官，特别是眼、肾、心血管及神经系统的损害及其功能障碍和衰竭。然而我国基层医生人数不足，专业水平参差不齐，同时糖尿病也分为很多种类，只有正确的认识糖尿病的种类才能够帮助人民群众有效的、有针对性的预防和治疗糖尿病。近年来自然语言处理技术快速发展，可运用此技术从医学文本中抽取医学实体和实体间的关系等知识，使用抽取到的知识可构建医学知识图谱，成功地将无结构化数据转换成结构化数据。医学知识图谱可以辅助医务人员对疾病诊断治疗，同时可更好的向人民普及医学知识，加快推动医学产业发展。当前运用自然语言处理技术构建知识图谱是一个学术研究的热点，同时它也广泛的应用于工业界的各方面。

本文运用中文自然语言处理技术从糖尿病医学文献中抽取知识，构建糖尿病知识图谱。这些医学文献中蕴含大量医学信息，对糖尿病的预防、诊断和治疗都有着十分重要的意义。由于这些医学文献无结构化，若利用人工抽取相关知识将会耗费大量的人力物力，如何高效准确地抽取文献中的知识，是本文研究的重点。

本文通过对医学知识图谱构建流程的研究，将构建基于中文自然语言处理的糖尿病知识图谱，主要分为命名实体识别、关系抽取及知识图谱构建三大部分。命名实体识别部分,本文提出了BERT-Bi LSTM-CRF 命名实体识别模型，该模型在传统 Bi LSTM-CRF 模型基础上，融合了 BERT 字嵌入模型，更好的结合文章上下文，充分考虑了一词多义等问题；关系抽取

分，本文构建了一种新的基于参数共享的关系抽取联合模型，即融合 BERT-Bi LSTM-CRF 和多头选择的联合模型，可同时进行命名实体识别和关系抽取两项子任务，同时在训练过程中两项子任务共享 Bi LSTM 层隐藏状态参数，并将两项子任务的损失函数之和作为最终的损失函数进行优化，增强了两项子任务间的交互性。知识图谱构建部分，构建了基于 Neo4j 图数据库的糖尿病知识图谱，详细的介绍了 Neo4j 图数据库和构建知识图谱的过程，并对糖尿病知识图谱进行分析。该知识图谱的成功构建可进一步应用于医药推荐系统，医学辅助诊疗系统等，对糖尿病患者的预防、诊断、治疗及康复管理都有着重要的帮助。
关键词：自然语言处理；命名实体识别；关系抽取；知识图谱