密 级_____

桂林電子科技大學
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

硕 士 学 位 论 文

（全日制专业学位硕士）

题目 中文网络招聘文本中的技能词抽取研究

（英文） Research on skill word extraction in Chinese online recruitment text

研 究 生 学 号: 1703303056

研 究 生 姓 名: 杨 鹏

指导教师姓名、职称: 文益民 教授

申 请 学 位 类 别: 工程硕士

领 域: 自然语言处理

论 文 答 辩 日 期: 2020 年 5 月 19 日

中国知网 https://www.cnki.net

中国知网 https://www.cnki.net

摘 要

近年来，随着我国高等教育的迅猛发展，大学毕业生也日益增多。尽管就业岗位的数量在不断增加，但我国劳动力市场的供需失配的结构性问题依然非常严重。如今，随着互联网的普及，网络招聘成为企业招聘人才的主流方式。招聘信息中列出的技能词为实时、准确地了解企业对人才的需求提供了可能。本文将技能词抽取任务转化为序列标注问题，借鉴了命名实体识别或者术语抽取的方法。然而，由于中文的语义和上下文情形的复杂性以及手工标注成本昂贵，从招聘文本中自动抽取技能词并非易事。

目前，深度神经网络已成为解决序列标注问题的主流方法。但是，这类方法专注于领域内监督学习，需要大量带标注的数据。对于网络招聘数据而言，由于人工标注既费时又昂贵，只能依靠领域专家手工标注少量语句。其次，这类方法完全依赖于神经网络进行特征提取，忽略了领域内的语料特征，没有充分利用到领域知识。另外，针对缺乏足够标注数据的困难，更好的方法应该采用迁移学习，借助其他领域中有标注的数据帮助提升技能词的抽取性能。而现有的基于深度学习的迁移方法需要源域和目标域具有相同的标签集或相同的标签含义，以及如何将从源域学到的知识迁移到目标领域也是一个挑战。因此，针对上述不足与困难，本文开展了两个研究：

（1）本研究以序列标注的经典模型 Bi-LSTM-CRF(Bidirectional Long Short Term Memory-Conditional Random Field)为基础。为了能够充分的利用领域知识，在它的输入层中加入语料特征，并将输入层的输出与 Bi-LSTM 层的输出进行拼接作为 CRF 层的输入。大量实验的结果表明了本研究的技能词抽取方法的合理性，加入的语料特征能有利于提升技能词抽取的准确率。

（2）针对缺乏足够的标注数据的困难，本研究提出了一种跨领域迁移学习的技能词抽取方法。它首先将源域语料库分解为三个子源域，然后在 Bi-LSTM 层和 CRF 层之间插入一个域自适应层，以帮助将每个源域中学到的知识迁移到目标域。再使用参数迁移方法来训练每个子模型。最后，通过多数表决得出最佳标记序列的预测。大量实验的结果说明了本研究方法的合理性，可以缓解人工标注数据的稀缺性。

本文工作的创新点如下：

1）提出了一种基于深度学习与语料特征相结合的中文网络招聘文本中的技能词抽取的算法；

2）提出了一种基于跨领域迁移学习的中文网络招聘文本中的技能词抽取算法；

3）建立了 IT 类行业的招聘文本语料库。

关键字：网络招聘，迁移学习，深度学习，技能词，参数迁移

中国知网 https://www.cnki.net

Abstract

In recent years, with the rapid development of higher education in China, the number of college graduates is also increasing. Although the number of jobs is increasing, the structural problem of supply-demand mismatch in China's labor market is still very serious. Nowadays, with the popularity of the Internet, online recruitment has become the main way for enterprises to recruit talents. The skill words listed in the recruitment information provide the possibility to understand the enterprise's demand for talents in real time and accurately. In this thesis, skill words extraction is transformed into sequence tagging problem, and the methods of named entity recognition or term extraction are used for reference. However, due to the complexity of Chinese semantics and context and the high cost of manual annotation, it is not easy to automatically extract skill words from recruitment texts.

At present, deep neural network has become the mainstream method to solve the problem of sequence tagging. However, this type of method focuses on supervised learning in the domain and needs a lot of labeled data. On the one hand, for online recruitment data, manual annotation takes time and effort. Therefore, only a few sentences can be manually annotated by domain experts. On the other hand, this method relies on neural network to extract features completely, neglects the corpus features in the field, and cannot make full use of the domain knowledge. In addition, for the difficulty of lacking enough annotation data, a better method should use transfer learning to help improve the recognition performance of skill words by annotation data from other domains. However, the existing transfer methods based on deep learning require that the source domain and the target domain have the same label set or the same label meaning. It’s also a challenge that transfer the knowledge learned from the source domain to the target domain. Therefore, in view of the above shortcomings and difficulties, this thesis carried out two studies:

(1) This study is based on the classical model Bi-LSTM-CRF of sequence annotation. In order to make full use of domain knowledge, corpus features are added to its input layer. The output of the input layer and the output of the Bi-LSTM layer are spliced as the input of the CRF layer. A large number of experimental results show that the method of skill word extraction in this study is reasonable, and the added corpus features can improve the accuracy of skill word extraction.

(2) In order to solve the problem of lack of enough annotation data, this study proposes a cross domain transfer learning method for skill word extraction. Firstly, it decomposes the

中国知网 https://www.cnki.net

source domain corpus into three sub source domains. Secondly, a domain adaption layer is inserted between the Bi-LSTM layer and CRF layer. It can help transfer the knowledge learned from each source domain to the target domain. Then the parameter transfer method is used to train each sub model. Finally, the prediction of marker sequence is obtained by majority vote. A large number of experimental results show the rationality of this research method, which can alleviate the scarcity of manual annotation data.

The innovations in this thesis are as follows:

1) In this study, we propose a skill word extraction algorithm based on the combination of deep learning and corpus features.

2) This study proposes an algorithm of skill word recognition in online recruitment text based on cross domain transfer learning.

3) A corpus of recruitment texts for IT industry is established.

Keywords: Online Recruitment, Transfer Learning, Deep Learning, Skill Words, Parameter Transfer

中国知网 https://www.cnki.net

目录

中国知网 https://www.cnki.net

中国知网 https://www.cnki.net

第一章 引 言

§1.1 研究背景与意义

近年来，尽管就业岗位的数量在不断增加，但世界劳动力市场上供求之间的技能缺口仍未解决。麦肯锡（McKinsey）[1]最近的一份报告讨论了欧洲主要经济体的雇主如何面对越来越多的危机，即找不到合适的人才来填补入门级职位。但是，与此同时，欧盟有 560 万年轻人没有工作。类似的案例发生在美国，印度，泰国和其他国家/地区。在国内，根据《中国劳动力市场技能缺口研究》[1]报告显示，仅 2016 年中国就有超过 1200 万的本科和高职专科毕业生。尽管我国拥有如此庞大的劳动力供应，但许多企业仍然很难招聘到合适的人才。报告中还提到，中国的劳动力供应存在着严重的人才层次矛盾。另外，技能缺口会对经济产生严重影响：失业率高，公司利润持续下降，严重阻碍经济增长，浪费大学教育资源。为了填补技能缺口，迫切需要快速，准确地分析雇主的需求，提高人才培养的针对性。

如今，随着互联网的普及，网络招聘成为企业招聘人才的主流方式。招聘信息中含有企业对所招岗位专业能力需求的具体描述，反映了企业的需求。因此，一些研究人员指出，网络招聘数据是提取专业技能词汇以准确检测企业需求的良好资源。然而，网络招聘数据虽易于获取，但是中文的语义和上下文复杂性以及缺乏标注数据的问题，从非结构化的网络招聘信息中自动抽取技能词并非易事。图 1-1 为网络招聘文本示例，其中“C++”、“Linux”、“推荐系统”等为岗位所要求的专业知识或专业能力。

任职要求：1.熟练运用C++编程，具有良好的编程习惯。2.有较强的进行代码调试和解决技术问题的能力。3.对于主流Deep Learning框架有了解。对于主流Deep Learining框架的内部框架有深入了解的优先。4、熟悉Linux开发环境，熟悉Python/C++语言；5、熟悉自然语言处理常见算法与模型（语言模型、MaxEnt/CRF, pLSA/LDA, w2v, CNN/RNN等）；6、参与过NLP项目（如中文分词、文本分类、文本聚类）；7、对推荐系统、大数据挖掘， deep learning等方向有浓厚的兴趣，有很强的自学能力的优先；8、计算机科学、机器学习、人工智能等专业职位工作经验优先；9、3-5年实际项目开发经验；10、有较强的团队精神和沟通交流能力。11、热爱运动喜欢跑步。

图 1-1 网络招聘文本示例

为简洁起见，在本文中，将岗位对所需人才的专业知识和专业能力的要求称为技

[1] 《中国劳动力市场技能缺口研究》，2016，清华大学、复旦大学

中国知网 https://www.cnki.net

能词。技能词可以看成是特定专业领域内的命名实体[2]或者术语[3]。因此，网络招聘文本技能词的抽取任务可以借鉴命名实体识别或者术语抽取的方法。目前，虽然命名实体识别的相关工作很多，但重点都是在识别正式文本中的人名，地名和机构名，而术语抽取的相关工作主要针对特定领域内的术语识别，缺乏通用性和可移植性。其次，大多数命名实体识别或者术语抽取研究均采用人工规则或传统的机器学习方法，基于一个学科或一个领域进行抽取，需要人工设计规则进行特征提取。近年来，许多相关研究也将命名实体识别或术语抽取任务转化为序列标注问题，而最新技术主要基于深度神经网络进行端到端训练以捕获上下文信息。这类方法可以将输入字符转换为输出标签，而无需显式的人工设计特征提取。但是，基于深度学习的序列标注方法仅专注于域内监督学习，这需要大量带标注的数据。对于网络招聘数据而言，人工标注数据既费时又昂贵，只能依靠领域专家手工标注少量语句。因此，很难获得足够的带标注的语句来训练深度神经网络。如何使用少量标注数据来快速、准确地抽取技能词是非常具有挑战的。由此可见，中文网络招聘文本中的技能词抽取研究仍然是一项颇为艰巨的，也是非常具有价值的任务。

§1.2 网络招聘数据挖掘研究现状

大量的网络招聘数据中蕴含着潜在的价值，近年来已经有一些关于网络招聘数据挖掘的研究。国内对于网络招聘数据挖掘的研究晚于国外，之前国内的研究大多是对网络招聘现状进行分析与阐述，但陆续也开展了很多深入的研究。按照数据挖掘的方向和应用场景不同，主要分为三种：技能需求挖掘、专业课程设计和人岗匹配分析。

§1.2.1 技能需求挖掘

针对网络招聘数据的岗位技能需求挖掘，主要有如下相关工作：在国外， Kim[35]等人通过对数据科学家岗位的招聘信息进行人工分析，提取出数据科学家这个岗位在企业界内所需的专业能力以及学历等各方面的要求。Chao[36]等人从 Monster 网站人工收集了 484 条招聘广告进行人工分类，找出对应职位的岗位职责、技能需求和工作经验等方面的要求。Lee[37]等人也人工收集了 555 条关于 IT 管理岗位的招聘信息，构建了 IT 行业中 3 个大类 50 个子类的岗位需求的技能词典。Cragin[38]等人根据从不同的招聘网站中人工收集到的招聘广告中，分析出求职者应该在工作经验、学历、技能要求等方面加强自我提升。Zhao[39]等人首先使用常用技能词短语、领域专家预定义的各种专业术语来分析网络招聘信息，再使用维基百科进行去重和规范化来提取岗位技能需求。Aniwat[40]等人使用网络抓取技术从泰国招聘网站上收集最新的职位信息，并使用自动关键字提取方法从职位描述中提取专业技能词。De Mauro[41]等人提出了一种基于机器学习算法和专家判断相结合的半自动分析方法，以按每个大数据技能组所需的适当水平的能力来表征每个大数据工作族。在国内，詹川[42]根据已有的关于电子商

中国知网 https://www.cnki.net

务行业的专业术语构建出该专业的技能词词典，对采集的 66925 条电子商务岗位的招聘信息进行关键词抽取，并对抽取出的高频技能关键词进行分类，分析出电子商务行业中不同类型岗位的通用技能需求和特定岗位技能需求。俞琰[43]等人从前途无忧招聘网站中抓取了 10000 条计算机领域和 30000 条非计算机领域的招聘信息，利用依存句法分析从计算机领域的招聘信息中选取候选技能，再利用非计算机领域的招聘信息计算候选技能中每个词的领域相关性并与候选技能的词频、词长等统计信息相结合得出领域相关性的 C-value 值，按值降序排列。最终选取前 N 个候选技能信息作为被抽取的技能词进行人工判定是否正确。

§1.2.2 专业课程设计

网络招聘数据用于专业课程设计，主要有如下相关工作：司莉[44]等人以国外招聘网站为数据源，从不同类型的招聘企业、具体岗位职责、基本职业能力需求和专业技能要求、工作经验要求等方面分析了国外的信息行业对图书情报学专业的人才要求，并对图书情报学科人才培养方案在课程体系设置、基本职业能力与专业实践能力需求等方面提出了建议。吕斌[45]等人、李国秋[46]等人也人工调研了 300 个关于情报学专业的招聘信息，分析了企业对情报职业的需求，以及情报学专业的学生毕业后从事的行业类型、职责和作用等。夏火松和潘筱昕[47]对比了硕、博士论文中提到的专业技术以及招聘网站中关于硕、博士相关招聘信息中所需要的技能要求，分析了我国大数据行业在学术界和企业界的发展现状，总结出我国大数据行业在企业界所需人才的技能需求与高校学术界的科学研究之间的关系。黄崑[48]等人从智联招聘网站上收集了 2615 份有关大数据岗位的招聘信息，分析出这些岗位所招聘的人才需要掌握的知识和能力要求，并对图书馆情报学科人才培养方案提出建议。夏立新[49]等人根据中华教育在线平台中提供的职业类别大全、在线招聘岗位分类、学术论文中技术关键词以及结合中文分词和词性标注，构建出“专业－岗位－知识点”的就业需求关系。

§1.2.3 人岗匹配分析

近年来，随着网络招聘数据的飞速增长。企业招聘人才需要寻求更加智能的方法去解决如何对不同岗位招聘到到合适的求职者这一问题，即人岗匹配问题。目前，针对网络招聘数据的人岗匹配分析的研究，有如下工作：Hoang[1]等人针对人岗失配的问题提出了 SKILL 系统。该系统利用网络招聘数据提取岗位要求中的技能词，并对技能词进行规范化表示。Zhu[50]等人提出了人岗匹配神经网络，在将岗位要求和简历中的个人经历表示为技能词的集合后，该网络能从岗位申请的历史数据中学习到人岗匹配的联合表示，从而能对申请者所掌握的技能与岗位所要求的技能是否匹配进行自动判断。Zhou[51]等人使用了主流的 TF-IDF(Term Frequency–Inverse Document Frequency)文本特征提取方法分析了网络招聘广告中技能词与工作岗位名称之间的关

中国知网 https://www.cnki.net

系，以修正技能词排序带来的不足，使得求职者能更准确的寻找适合自己的工作岗位。Xu[52]等人提出了技能词流行度主题模型，该模型从网络招聘数据中学习，将薪资水平、公司规模与技能要求集成起来，对技能词的流行度进行排序，使得求职者能及时明白什么时候最应该学习什么样的技能。Qin[53]等人提出了一种基于能力感知的人岗匹配神经网络，将招聘岗位中的需求信息与求职者简历中的项目经验进行语义表征自动测量每个项目经验对岗位需求中的技能的贡献度，从而得出求职者与岗位的吻合度。Shen[54]等人提出了一种基于主题模型 LDA(Latent Dirichlet Allocation)拓展的智能学习面试评估的联合学习模型，以工作描述，候选人简历和面试评估共同建模。从历史上成功的求职记录中有效地学习不同求职面试过程的代表观点，挖掘出这几类文本的相关联系，提高面试评估质量。

总体来说，这些研究都取得了一定的进展，但基于人工来获取和手工处理网络招聘数据的方式，存在如下缺陷：1)获取的网络招聘数据样本量偏小;2)采用人工手段处理数据，执行效率过低;3)对数据进行分析时存在主观性。基于外部资源和规则统计的方法存在专业词典构建简单、覆盖面狭窄、词典信息来源更新较慢和规则制定依赖于领域知识，需要人工编制等缺陷。另外，通过网络招聘数据提取企业对人才的技能需求还可以进一步用于大学的招聘市场趋势分析，技能热度预测以及企业竞争力评估等[55]。由此可见，如何自动、快速、准确地识别招聘信息中所包含的技能需求非常重要。

§1.3 论文研究内容

论文以网络招聘数据挖掘为研究课题，针对缺乏标注数据的技能词抽取问题尝试开展了如下的两个研究：

（1）网络招聘文本中具有特定领域知识，为了能够充分的利用领域内的知识，挖掘更多的语料特征来更好的提升技能词的抽取准确率。本研究提出了一种结合语料特征的技能词抽取方法。具体地来说，我们以序列标注的基本模型 Bi-LSTM-CRF 为基础，在输入层中使用字符嵌入特征的同时加入了位置特征、词性特征与上下文特征。并将输入层的输出与 Bi-LSTM 层的输出进行拼接作为 CRF 层的输入来预测最佳标签序列。大量实验表明本研究提出的技能词抽取模型具有最佳的抽取性能，并评估了加入丰富的语料特征，本研究提出的模型是否能够缓解模型对大量标注数据的依赖。

（2）由于缺乏足够的标注数据，更好的方法应该利用域外标注数据，通过迁移学习方法提升域内技能词抽取。所以在本研究中我们提出了一种跨领域迁移学习的专业技能词抽取方法。它首先将源域的语料库分解为三个子源域，在 Bi-LSTM 层和 CRF 层之间插入一个域自适应层，以帮助从每个源域中学到的知识迁移到目标域。然后，使用参数迁移方法来训练每个子模型。最后，通过多数表决获得标记序列的预测。大量实验的结果说明了本研究方法的合理性，跨领域迁移学习的专业技能词抽取方法可

中国知网 https://www.cnki.net

以缓解手工标注数据的稀缺性，对 IT 领域招聘数据的技能词抽取取得的最佳结果，以及验证了对机械制造和服装设计行业的招聘数据中技能词抽取的有效性。

§1.4 论文组织结构

在论文我们将其划分为六个章节，具体组织结构如下：

第一章，引言。在该章节中，首先介绍了本文的研究背景以及意义，并指出了存在的挑战。然后详细地介绍了网络招聘数据挖掘目前的研究现状。最后，对本文的结构布局和研究内容进行了简要的总结。

第二章，命名实体识别与术语抽取相关方法。在该章节中，首先介绍了传统命名实体识别与术语抽取的方法。其次详细阐述了基于深度学习的命名实体识别与术语抽取方法以及采用深度迁移学习的相关工作。最后对深度学习的理论基础以及本研究所采用的基线模型：Bi-LSTM-CRF 模型进行了详细的描述。

第三章，网络招聘信息的采集。在该章节中，首先简单介绍了本研究所用的网络招聘信息的数据来源。其次详细介绍了采集数据所设计的网络爬虫架构，以及数据的去重和存储。

第四章，基于深度学习的中文网络招聘文本中的技能词抽取。在该章节中，提出了一种基于深度学习模型和语料特征相结合的技能词抽取方法，通过在模型的输入层中使用字符嵌入特征的同时加入了位置特征、词性特征与技能词的上下文特征。并将输入层的输出与 Bi-LSTM 层的输出进行拼接作为 CRF 层的输入来预测最佳标签序列。

第五章，基于跨领域迁移学习的技能词抽取算法。在该章中，提出了一种跨领域迁移学习抽取网络招聘信息中技能词的方法。它首先将源域语料库分解为三个子源域，在 Bi-LSTM 层和 CRF 层之间插入一个域自适应层，以帮助每个源域学习到的知识迁移到目标域。然后，使用参数迁移方法来训练每个子模型。最后，通过多数表决得到最佳标记序列的预测。

第六章，总结与展望。对本文的研究工作进行了系统地总结。并且对于网络招聘数据中技能词抽取工作的未来研究方向做了进一步的讨论。

中国知网 https://www.cnki.net

第二章 命名实体识别与术语抽取相关方法

上一章介绍了网络招聘信息中技能词抽取问题研究的背景和意义以及网络招聘数据挖掘的研究现状。本研究将技能词抽取任务转化为序列标注问题，类似于命名实体识别或术语抽取。因此，我们简要回顾与技能词抽取密切相关的一些研究。这些相关工作包括命名实体识别和术语抽取。

命名实体识别旨在识别特殊实体并将其分类为预定义类别的任务，例如产品名称，旅游景点名称，新闻领域的人名或生物医学领域的疾病名称[2]，是自然语言处理中一项基本任务。术语抽取也是自然语言处理中一项基础任务，指从文本中自动发现术语的过程。它可以应用于信息检索、关系抽取、对话生成等复杂任务领域[32]。随着信息技术的不断发展，各领域数据不断扩张，产生大量的领域术语[3]。这些领域术语在该领域中具有很强的特定意义，是构成该领域专业文本的信息主体。与术语识别类似，命名实体识别也是从一段自然语言文本中识别特定类型的实体。因此，命名实体识别也逐渐由识别通用对象转向识别特定领域术语，特定领域术语抽取也逐渐采用命名识别中的方法。另外，从本质上讲，如果仅仅抽取文本中术语的名称或简称，不标注术语的具体表述或评价，术语抽取可以被视为一类命名实体识别[22]。当前，用于命名实体识别和术语抽取的方法可以简单地分为两种：传统方法和基于深度学习的方法。

本章重点介绍应用于命名实体识别与术语抽取任务的主流技术以及深度学习技术的理论基础。之后将详细介绍本文开展的两个研究工作中所使用的基础模型 Bi-LSTM-CRF 框架。最后，对本章内容进行简单的总结。

§2.1 传统命名实体识别与术语抽取方法

§2.1.1 基于规则的命名实体识别与术语抽取方法

传统方法大致可以分为三类：基于规则，无监督学习和基于特征的监督学习方法。最早，命名实体和术语的识别方法是依据语言学家手工构造的规则模板，而规则的设计一般是基于句法、语法、词汇的模式以及特定的领域知识等。例如：可以通过结合中文姓氏词典与词性特征来识别中文人名“张三”[5]。Chen[23]等人根据所需要抽取的术语在语料中的上下文信息和其自身的组成规则进行抽取。

在识别命名实体时，需要根据特定领域制定不同的规则模板，利用最适合该领域的规则模板才能有效的完成命名实体识别任务[4]。例如，在识别“张三”、“老赵”、“小王”等类似的中文人名时，可以利用中文姓氏词典和常用名表再加上词性信息等特征进行识别。而直接将之前针对“张三”等识别人名训练好的识别模型用于识别其他领域的实体时，很难完成有效的识别。在识别特定领域术语时，同样需要根据相应领域的

中国知网 https://www.cnki.net

术语构词和组成特点来制定规则。例如，科技领域的术语多为名词，可以将科技领域术语限定为名词性短语，利用词性特征来获取候选词串，或者通过分析语料中各语句的语法结构来判定语句中的名词短语是否为该领域术语。

在特定领域，人工构造的规则模板可以准确地总结出特定实例的特征，其识别效果比较准确。然而这些规则的构建往往依赖于具体语言、特定领域知识，耗时耗力，并且需要根据不同领域需要制定不同的规则模板，受限于规则制定者所具有的知识。此外，人工制定的规则不能涵盖所有该领域出现的命名实体或者术语的特点，往往需要通过多次不断完善的规则模板才能达到预想的识别效果。

§2.1.2 基于无监督学习的命名实体识别与术语抽取方法

无监督学习方法通常利用语言学知识与统计学或信息论相结合的方法，但不需要训练语料、较少人工干预。无监督学习方法应用于命名实体的研究工作不是很多，主要思路是利用上下文语义相似性，从聚集的词组中提取命名实体。其核心思想在于从巨大语料中统计出词汇资源、词汇模型来推断命名实体的类别。Zhang[8]等人提出了一种借助语料库统计信息（例如逆文档频率和上下文向量）和浅层语法知识（例如名词短语分块）来推断命名实体。而基于无监督学习的术语抽取有许多相关工作。主要思路也是先从语料库中选取候选术语，然后利用统计信息（例如使用互信息、词频和词长度等）计算候选术语成为术语的可能性。例如：Pantel[24]等人提出了一种利用语料统计信息抽取术语的方法，通过计算两个词语或两个词语组成的词串之间的互信息再结合似然比对候选术语进行评分，抽取出评分较高的前 N 个术语。该方法的一般流程是：首先将语料经过预处理得到词向量或字符向量后，对语料中各种信息进行统计，例如，使用词性过滤、词语频率、均值和方差等统计学或信息论中常用的统计方法。再根据统计结果，采用特征值量化的方式形成候选词语集合，然后在此基础上计算出候选词语成为命名实体或者术语的可能性，识别出最终的命名实体或者术语。比较常用的统计量化指标主要有 3 种：（1）基于词权重,具体包括词性、词计数、词频、文档频率、逆文档频率、x1 测度、平均词频、相对词频、词长等；（2）词的文档位置，如：文档前 N 个词、文档后 N 个词、段首 N 个词、段尾 N 个词、标题词、文档特殊位置词等；（3）基于词的关联信息，主要包括互信息、均值、方差、共现度、依存度、中心性测度、聚类系数、TFIDF 值、PageRank 值、Hits 值、短概念缩减、长概念提升、词跨度等[32]。一般衡量成为命名实体或者术语的可能性的方法有：T 检验、卡方检验、对数似然比、点互信息等假设实验方法。总体而言，这些基于统计学方法的基本思想都一样，只是在公式的参数设置或计算有所不同。

基于无监督学习的方法不需要句法以及语义上的信息，也不局限针对某一专门领域，更不依赖任何资源，具有较少人工干预、通用性较强等优点。但是其算法性能比较依赖于语料库的规模、中文分词的准确性以及候选命名实体或术语的词频，存在

中国知网 https://www.cnki.net

无法识别出低频率的候选命名实体或领域术语的缺陷, 针对数据稀疏或者数据量规模较小的语料使用该方法很难获得理想的效果。

§2.1.3 基于特征监督学习的命名实体识别与术语抽取方法

随着技术的不断发展，机器学习算法应用在自然语言处理场景任务逐渐增多。基于特征监督学习的方法，主要是将命名实体识别或者术语抽取任务可以被转化为序列标注任务。通过对序列标注任务可以建立复杂的数学模型。即对于输入序列 $X=\{x_1,x_2,...,x_n\}$，再给出对应的标签序列 $Y=\{y_1,y_2,...,y_n\}$。主要流程是：先根据标注好的数据，研究者应用领域知识与工程技巧设计复杂的特征(如，词性特征、词语构造特征、上下文特征等)来表征每个训练样本，然后应用机器学习算法训练模型，使其对已有的标注语料进行学习，最后利用训练好的模型来识别需要标注的文本数据中的实体或术语。本节主要介绍基于特征监督学习的命名实体识别与术语抽取方法中应用到的常用机器学习算法：如，最大熵模型、隐马尔科夫模型及条件随机场模型，下面进行详细介绍。

（1）最大熵模型（Maximum Entropy，ME）

最大熵模型的选取采用最大熵原理的核心思想，即在满足一定约束条件的情况下使得熵值最大的模型就是最好的模型。换句话说，在所有满足最大熵原理约束条件的模型中条件概率的熵值最大的模型便是最大熵模型，公式表示如下：

$$P(y|x)=\frac{1}{H(x)}\exp(\sum_{i=1}^{j+1}w_i\cdot f_i(x,y)) \qquad (2\text{-}1)$$

其中，$H(x)$ 为规范化函数，公式表示如下：

$$H(x)=\sum_{y\in Y}\exp(\sum_{i=1}^{j+1}w_i\cdot f_i(x,y)), \qquad (2\text{-}2)$$

其中，$f_i(x,y)$ 是特征函数，其函数权重为 w_i。而所有可能的标注结果 y 组成的集合表示为 Y。

在序列标注任务中，$\{x_1,x_2,...,x_n\}$ 用来表示输入的文本序列，对应输出的标签序列则是 $\{y_1,y_2,...,y_n\}$。利用最大熵模型来处理序列标注任务的关键就是如何选取有效的特征函数 f_i 来计算出由输入的文本序列 $\{x_1,x_2,...,x_n\}$ 得到对应的标签序列 $\{y_1,y_2,...,y_n\}$ 的条件概率 $P(y|x)$。而当条件概率 $P(y|x)$ 最大时，所产生的对应标注序列便是由最大熵模型所得到的最佳标注序列。

（2）隐马尔科夫模型(Hidden Markov Models，HMM)

20 世纪 80 年代初，隐马尔科夫模型由 Rabiner 提出，是一个时间序列概率模型，十分适合处理自然语言处理中的序列标注问题[6]。

中国知网 https://www.cnki.net

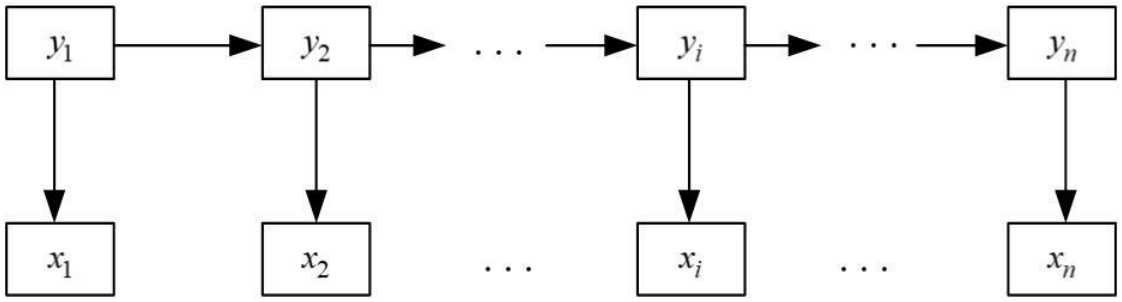


图 2-1 隐马尔科夫模型结构图

在隐马尔科夫模型中存在两组变量，一组是不可观测的隐变量$\{y_1,y_2,...,y_n\}$，另一组是观测变量$\{x_1,x_2,...,x_n\}$。如图 2-1 所示：在t时刻，隐变量y_t决定观察变量x_t，而t-1时刻的隐变量y_{t-1}决定t时刻的隐变量y_t。可用公式表示为：

$$P(x_1,y_1,...,x_n,y_n)=P(y_1)P(x_1\mid y_1)\prod_{i=2}^{n}P(y_i\mid y_{i-1})P(x_i\mid y_i) \qquad (2\text{-}3)$$

此外，隐马尔科夫模型的构建还需要如下信息：状态信息$Q=\{q_1,q_2,...,q_N\}$、观测信息$O=\{o_1,o_2,...,o_M\}$以及参数$\lambda=(A,B,\pi)$，记为$(Q,O,\lambda=(A,B,\pi))$。其中，$A=[a_{ij}]_{N\times N}$代表状态转移矩阵，即模型中各状态之间的转移概率组成矩阵。$B=[b_{ij}]_{N\times M}$则表示为输出观测矩阵，即当前状态模型获得各观测值的概率矩阵。初始状态概率用$\pi=\{\pi_1,\pi_2,...,\pi_N\}$表示，即初始时刻各状态出现的概率。

在处理序列标注问题中，隐马尔科夫模型根据已知观测序列$X=\{x_1,x_2,...,x_n\}$和参数$\lambda=(A,B,\pi)$，计算概率$P(X\mid\lambda)$的概率，找出使得概率$P(X\mid\lambda)$)最大的参数$\lambda=(A,B,\pi)$和匹配状态序列$Y=\{y_1,y_2,...,y_n\}$。但其前提是其输出的观测序列$Y=\{y_1,y_2,...,y_n\}$必须满足独立性假设。这使得在实体或者术语识别过程中，HMM 模型限制了当前状态对位于其前面位置的状态的依赖长度，不能充分的利用语料中的上下文特征。

（3）条件随机场模型(Conditional Random Fields，CRF)

条件随机场模型是在 2001 年由 Lafferty 等人提出的，通过概率模型的构造来标注序列的方法[22]。其主要思路就是通过计算条件概率模型$P(Y\mid X)$，得出观测序列$X=\{x_1,x_2,...,x_n\}$所对应的标记序列$Y=\{y_1,y_2,...,y_n\}$。在序列标注任务中，条件随机场从句子层面考虑，通过计算联合概率进行避免最大熵和隐马尔科夫模型中出现的标识偏置等问题。

在 CRF 模型中，存在两种特征函数，即状态特征函数和转移特征函数。状态特征函数只和当前节点有关，记为：Z_j。

$$Z_j(y_i,X,i),j=1,2,...,J, \qquad (2\text{-}4)$$

而转移特征函数受当前节点和上一节点的影响，记为：t_k

$$t_k(y_{i\text{-}1},y_i,X,i),k=1,2,...,K, \qquad (2\text{-}5)$$

从而，条件随机场模型的条件概率定义为：

中国知网 https://www.cnki.net

$$P(Y \mid X)=\frac{1}{H}\exp\left(\sum_{k}\sum_{i=1}^{n-1}\lambda_k t_k(y_{i-1},y_i,X,i)+\sum_{j}\sum_{i-1}^{n}\mu_j Z_j(y_i,X,i)\right), \quad (2\text{-}6)$$

其中，序列 X 在标注为 y_{i-1} 和 y_i 之间的转移概率表示为 $t_k(y_{i-1},y_i,X,i)$，序列 X 中位置 i 的标记为 y_i 的概率表示为 $Z_j(y_i,X,i)$， H 为规范化函数， λ_k 和 μ_j 为权重参数。

在使用 CRF 模型处理序列标注任务的过程中，只依靠特征函数来考虑当前输出与上一时刻输出之间的影响。因此，识别的精度很大程度上依赖于人工所设定的特征模板。然而，我们还需要耗费大量的时间对语料的特征进行充分的挖掘，如果特征模板的覆盖范围不够好，便会有很大的可能导致识别精度不够高，以及很难识别较长的组合型的命名实体和术语。

§2.2 基于深度学习的命名实体识别与术语抽取方法

近年来，深度神经网络已成为提取命名实体或术语有效潜在特征的主流方法。标准的做法是运用不同的神经网络模型（卷积神经网络，Convolutional Neural Networks, CNN；长期短期记忆，Long Short Term Memory, LSTM）提取不同类型的表示形式（字符级或单词级），将学习到的特征表示馈入顶部的 CRF 层中以进行序列标签预测，从而实现命名实体识别或术语抽取。例如，Collobert[14]等人首先利用 CNN 网络学习单词级特征表示，然后将学习到的特征表示馈入到 CRF 进行序列标签预测。Lample[15]等人利用 Bi-LSTM 网络结构提取字符的上下文潜在特征表示，并利用 CRF 层进行标签解码。Huang[16]等人提出了一个与 LSTM-CRF 类似的模型，但是加入了人工构造的特征（例如，单词是否以大写字母开头、字母的前缀和后缀等）。他们首先通过 Bi-LSTM 网络提取了字符的上下文潜在特征表示，然后将其与人工构造特征相连接，最后将其输入到 CRF 层中。CNN-CRF 模型 LSTM-CRF 模型的区别在于，CNN 网络对提取词的形态信息（例如，词的前缀和后缀）非常有效，相对于中文单字无法分解，更适用于处理英文数据。因为，英文字母潜藏着一些潜在的特征，而英文单词便是由若干个这种细粒度的字母组成。但 CNN 网络在长序列输入上特征提取能力弱，而 LSTM 提供了长距离的依赖，拥有较强的长序列特征提取能力。因此，有研究者为了充分利用两个模型的优点，将两者结合。如，Ma[17]提出了 Bi-directional LSTM-CNNs-CRF 网络架构. 在该架构中，CNN 网络首先用于提取单词的字符级特征表示. 将 CNN 网络提取出的字符级特征表示向与预训练的词嵌入连接起来，然后再馈入 LSTM 网络以提取单词级的上下文潜在特征表示，最后进入 CRF 层进行标签解码。这些基于神经网络的方法不需要特定于任务的特征工程，便可以获得良好的性能。

深度学习也逐渐应用在术语抽取研究工作中，如：闫兴龙[28]等人针对网络安全领域提出了一种基于 Bi-LSTM-CRF 网络结构的新型网络安全专业名称抽取模型，以从非结构化文本中提取与安全相关的概念和实体。赵东玥[29]等人同样采用了 Bi-LSTM

中国知网 https://www.cnki.net

模型对科技文献进行术语抽取。赵洪[30]等人以 Bi-LSTM-CRF 模型为基础框架，融入了理论术语的语料特征（例如，理论术语的尾词特征、术语的构词特征等），构建了基于深度学习的理论术语抽取模型，并提出一种自训练算法，实现模型的弱监督学习。

基于深度学习方法的模型，虽然不需要设计特定于任务的特征工程，但网络模型的训练需要大量的带标注数据。而并非所有领域都具有充分的带标注数据，所以在基于深度学习技术的序列标注模型中如何使用少量标注数据也能取得不错的识别性能的研究是最近的重点[57]。迁移学习针对只有少量标注数据的目标领域，旨在通过利用从源域学到的知识来帮助提高目标任务的性能。根据 Pan[56]等人的观点，迁移学习算法主要分为实例迁移，特征表示迁移，关系知识迁移和参数迁移四种类型。在序列标注任务上，主要的迁移学习方法包括参数迁移和特征表示迁移以及将这两种方法组合也可以用于迁移[64]。

（1）参数迁移方法

参数迁移，通常利用参数共享或联合训练来获得与源域模型的参数接近的目标域模型参数。Yang[57]等人首先研究了表示的不同层次的可迁移性。他们提出了基于 Bi-GRU-CRF 模型的三种不同的参数共享架构，以处理序列标记任务中的跨应用程序，跨语言和跨域迁移。在训练过程中，每次迭代都随机采样源任务或目标任务，并再次按批次采样来自同一采样任务的训练实例，然后执行梯度更新以更新任务特定参数和共享参数。考虑到参数迁移方法要求在源域和目标域中输出空间都相同，Lin[58]等人提出在预训练的源域模型之上引入三个神经适应层：单词适应层，句子适应层和输出适应层。在训练过程中，使用预先训练的源模型的参数与目标任务模型的相应参数进行初始化，然后调整所有层的参数。 Lee[59]等人提出使用源域中的标注数据来训练模型。接下来，使用从源域数据训练出的模型中学习到的参数作为目标任务模型的初始参数，然后使用来自目标域的标注数据对参数进行进一步的微调。

（2）特征表示迁移方法

特征表示迁移方法通常学习特征映射，该映射将源域和目标域同时按照相似的分布投影到公共特征空间上，或者查找域共享和特定于域的特征表示。Peng[60]等人提出了一个多任务领域自适应模型，其中包括共享的表示学习器，域投影层和任务特定分类器。在该模型中，表示学习器层在所有领域和任务之间共享，域投影层将学习的不同领域的表示转换为相同的共享空间。在每个数据集之间进行交替优化来训练该模型。Kulkarni[61]等人提出了除了捕获全局语义之外还捕获单词的领域特定语义并利用领域特定词嵌入来学习不同领域的命名实体识别模型。Xiao[62]等人提出了一种对数双线性语言适应模型，通过学习可推广的和特定于源的单词表示来建模源分布，并通过学习域共享和特定于目标的单词表示来建模目标分布。学习的嵌入特征表示用于将原始数据映射到诱导表示空间中，然后将每个单词的诱导表示向量用作目标任务的增强特征。

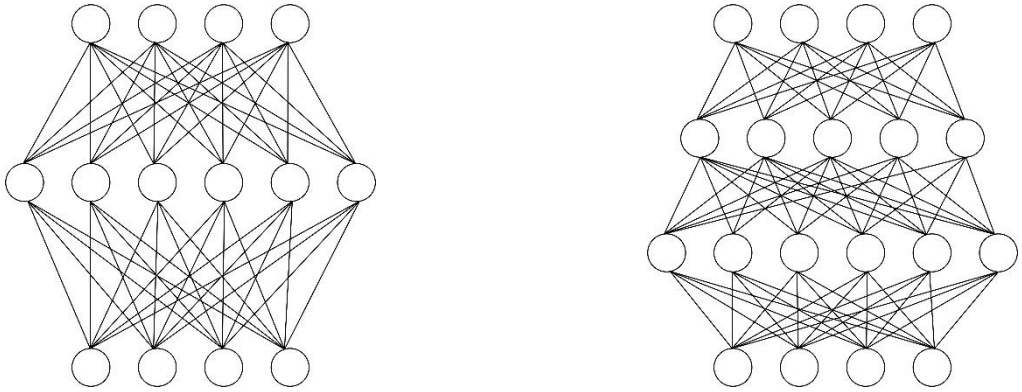
中国知网 https://www.cnki.net

§2.3 深度学习理论基础

本节主要介绍深度学习的基本原理以及不同深度神经网络模型的相关理论概述主要包括：前馈神经网络,又称多层感知器(Multi-Layer Perceptron，MLP)、循环神经网络(Recurrent Neural Network，RNN)以及长短时记忆神经网络(Long Short-term Memory Networks，LSTM) 等。

§2.3.1 前馈神经网络

前馈神经网络又被称之为多层感知机，是之后其他神经网络模型结构的基础[2]。顾名思义，多层感知机便是由若干层的神经网络组成，而通过若干个神经元组成每层的神经网络。



(a) 单隐层前馈神经网络结构　　(b) 多隐层前馈神经网络结构

图 2-2 前馈神经网络结构图

如图 2-2 所示，单层前馈神经网络与多层前馈神经网络具有类似的结构，第一层输入层，作用便是输入特征向量 $X=[x_1,x_2,...,x_n]^T$，最后一层是输出层，作用是输出分类类别 $y_j(j=1,2,...,m)$，而中间层均为隐藏层。在单层神经网络中，输入层的向量乘上相应权重 w_{ji} 直接得到输出层的值。并且，整个网络只按照单方向传播。在数学形式中，单层神经网络可进行如下表示：

$$s_j=\sum_{i=1}^{n}w_{ji}x_i-\theta_j\,, \tag{2-7}$$

$$y_j=f(s_j)=\begin{cases}1, s_j\geq 0\\ 0, s_j<0\end{cases}, \tag{2-8}$$

在数学形式中，多层前馈神经网络可进行如下表示：

$$s_i^{(q)}=\sum_{j=0}^{n_{q-1}}w_{ij}^{(q)}x_j^{(q-1)},(x_0^{(q-1)}=\theta_i^{(q)},w_{i0}^{(q-1)}=-1)\,, \tag{2-9}$$

$$\begin{aligned}&x_i^{(q)}=f(s_i^{(q)})=\begin{cases}1, s_i^{(q)}\geq 0\\ 0, s_i^{(q)}<0\end{cases},\\ &i=1,2,...,n_q;\ j=1,2,...,n_{q-1};\ q=1,2,...,Q\end{aligned}\,, \tag{2-10}$$

中国知网 https://www.cnki.net

此时，多层前馈神经网络通过将输入特征向量投影到高维超平面的形式，能够处理输入数据较为复杂的分类问题。

§2.3.2 循环神经网络

循环神经网络是一种时序网络结构，善于处理和预测序列信息[2]。从本质上来讲，循环神经网络的作用便是处理一个序列当前的输出与之前信息的关系。换句话说，RNN 会记忆当前节点之前的信息，并利用之前的信息影响后一时刻的输出。如图 2-3 所示，这是某一时刻的 RNN 网络结构，当前节点 A 除了接收来自输入层中传输的信息 x_t，还接收来自上一时刻的隐藏状态信息 h_{t-1}。

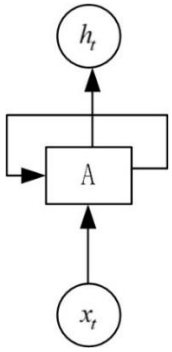


图 2-3 循环神经网络典型结构图

从数学形式，可以表示为：

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h)\,, \tag{2-11}$$

$$y_t = W_{hy}h_t + b_y\,, \tag{2-12}$$

其中，σ 表示非线性激活函数(一般常用的激活函数为：tanh 或 sigmoid 等)，x_t 表示 t 时刻样本的输入，y_t 表示 t 时刻网络的输出，h_{t-1} 为 t-1时刻的隐藏状态，W_{xh} 表示输入节点 x_t 与当前隐藏状态 h_t 之间的权重矩阵，W_{hh} 为上一时刻隐藏状态 h_{t-1} 与当前隐藏状态 h_t 之间的权重矩阵，b 为偏置。

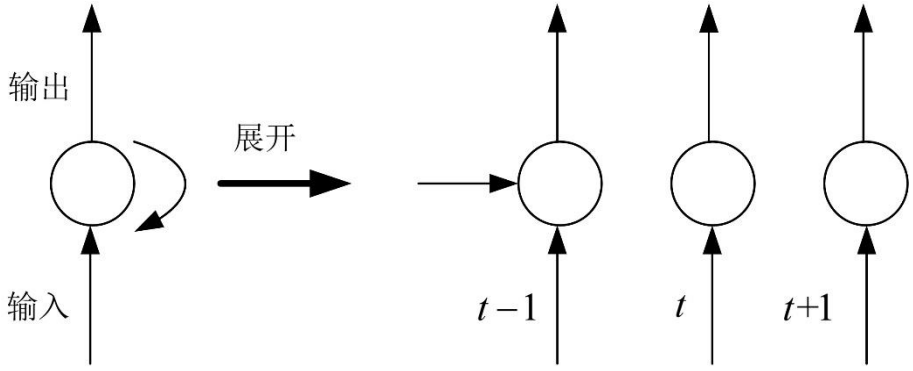


图 2-4 循环神经网络按时间展开的结构图

此外，RNN 可以从任意长度的时序信息中提取出序列特征，映射成固定大小的特征向量。RNN 对长度为 n 的序列展开后，可以视为一个有 n 个隐藏层的前馈神经网

中国知网 https://www.cnki.net

络。图 2-4 所示的便是 RNN 网络按时间展开的结构图。但不同于前馈神经网络，循环神经网络不仅可以单向传播，还可以反向传播。在处理实际问题时，随着隐藏层的数量不断增加，前面隐藏层梯度也逐渐低于后面隐藏层梯度导致分类性能逐渐变差，从而也逐渐引发梯度消失。

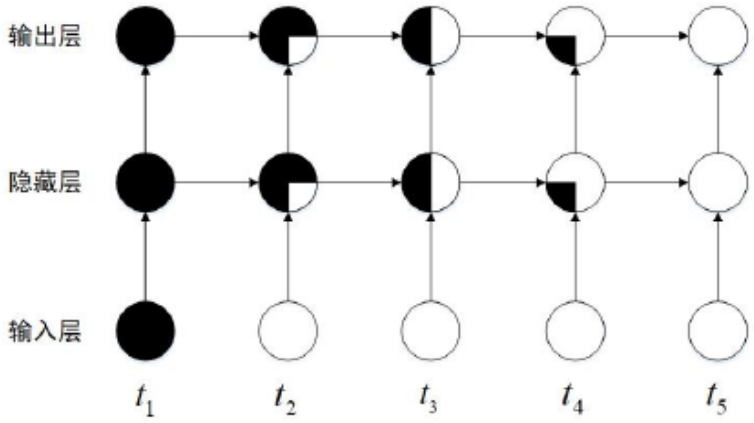


图 2-5 梯度消失示意图

如图 2-5 所示，t_1 时刻，图中的黑色部分代表着模型中的输入信息从输入层到输出层时的留存状态。随着不同时刻的逐层传递，后续隐藏层和输出层所接收到的模型信息在传递过程中逐渐变少。当传递至 t_5 时刻时，t_1 时刻的模型信息已经完全丢失。当反向传播时，模型也无法计算 t_5 时刻的误差不能有效更新相对应的权值。造成梯度消失这种现象的原因在于模型对较长距离信息的序列无法训练，即长期依赖问题。为了处理这种长期依赖问题，长短期记忆网络便应运而生。

§2.3.3 长短时记忆神经网络

长短期记忆网络是循环神经网络的扩展，通过增加记忆单元能够解决循环神经网络存在的长期依赖问题。它通过两种“记忆”方式来改变记忆单元，一种是允许网络将当前时刻的信息进行遗忘，另一种是当有新的模型信息传递到网络时，对当前记忆单元内存储的内容进行更新。因此，如图 2-6 所示，LSTM 在隐藏层的神经元中加入一个细胞状态和三个门，分别是：输入门，遗忘门，输出门。

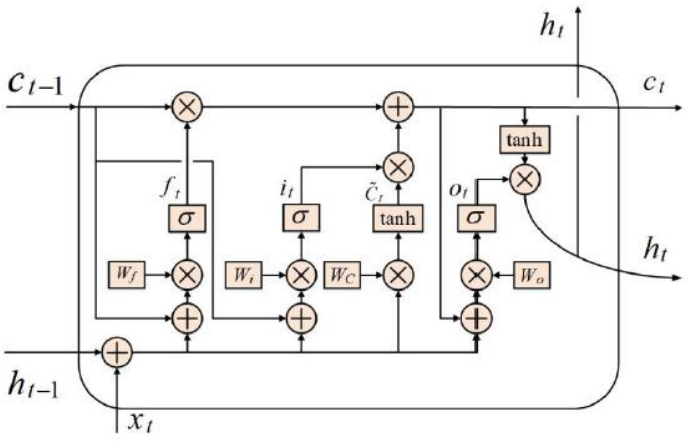


图 2-6 LSTM 记忆单元图

中国知网 https://www.cnki.net

初始输入的信息中是否可以传递到细胞状态中由输入门控制。上一时刻的单元状态中的信息是否保留到当前时刻来控制细胞状态的信息传递是由遗忘门决定，更新后的细胞中是否可以输出的信息则由输出门控制，具体实现是：

$$f_t = \sigma(W_f \cdot (c_{t-1} + h_{t-1} + x_t) + b_f), \quad (2\text{-}13)$$

$$i_t = \sigma(W_i \cdot (c_{t-1} + h_{t-1} + x_t) + b_i), \quad (2\text{-}14)$$

$$\tilde{c}_t = \tanh(W_c \cdot (h_{t-1} + x_t) + b_c), \quad (2\text{-}15)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t, \quad (2\text{-}16)$$

$$o_t = \sigma(W_o \cdot (c_t + h_{t-1} + x_t) + b_o), \quad (2\text{-}17)$$

$$h_t = o_t \tanh(c_t), \quad (2\text{-}18)$$

其中σ表示 sigmoid 激活函数，tanh 为新的细胞状态输出时的激活函数。在t时刻的输入门，遗忘门，输出门和细胞状态分别用i_t， f_t， o_t， c_t表示，新候选向量表示为$\tilde{c}_t$。在t时刻，x_t表示模型的输入向量，对应的隐藏层向量为h_t。在t-1时刻的细胞状态和 LSTM 的输出，则分别用c_{t-1}和h_{t-1}表示。W代表权重矩阵，b代表偏置向量，其下标分别对应各个门。例如，W_i表示输入门的权重，则输入门的偏置表示为b_i。

简而言之，LSTM 网络通过增加记忆单元来保存上文信息，再加上门模块的设置在一定程度上能够避免网络在逐层传递的过程中对单一矩阵进行重复的乘法操作从而能够缓解了 RNN 的长期依赖问题。这也是 LSTM 能够成功应用到时序任务的原因。因此，近年来 LSTM 在语音识别、文本挖掘、对话生成等任务中都有很好的表现。

§2.3.4 Bi-LSTM-CRF 网络模型

在序列标注任务中，通常不仅需要考虑上文信息还需要考虑下文信息。但是，LSTM 只能记录t时刻之前得历史信息，而不能记录t时刻之后的将来信息，在一定程度上限制了 LSTM 模型的性能。而 Bi-LSTM 可以从全局上下文中学习序列的隐藏表示，具体 Bi-LSTM 模型结构如图 2-7 所示：

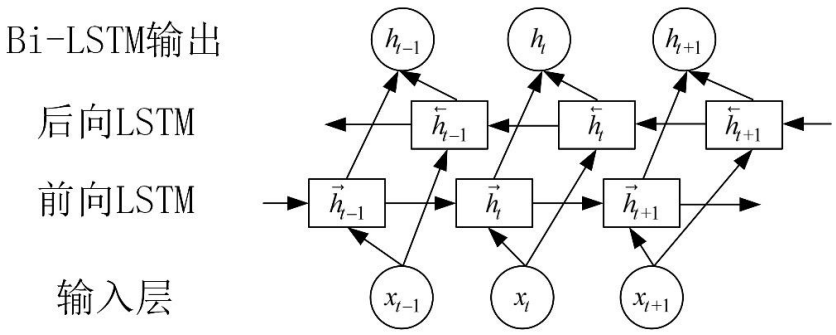


图 2-7 Bi-LSTM 模型结构图

对于输入序列$X=(x_1,x_2,...,x_n)$，将$\overrightarrow{LSTM}$表示为 LSTM 网络从左到右扫描输入序列，$\overrightarrow{LSTM}$学习到的隐藏表示可以表示为$[\overrightarrow{h_1},\overrightarrow{h_2},...,\overrightarrow{h_n}]$。类似地，$\overleftarrow{LSTM}$表示 LSTM 网络从右到左扫描序列，$[\overleftarrow{h_1},\overleftarrow{h_2},...,\overleftarrow{h_n}]$表示由$\overleftarrow{LSTM}$学习到的隐藏表示。再通过拼接$\overrightarrow{LSTM}$

中国知网 https://www.cnki.net

和 $\overleftarrow{LSTM}$ 上下文表示得到：$h_i=[\overrightarrow{h_i},\overleftarrow{h_i}]$。最终，Bi-LSTM 层的输出为 $h=[h_1,h_2,...,h_n]$。

从 Bi-LSTM 层输出的向量可以直接用作特征，采用 softmax 函数输出每个字符最可能的分类标记。但是，在序列标注任务中，每个输入字符的标签都涉及上下文语义关系，相邻标签通常具有很强的依赖性。而由 softmax 函数输出的字符分类标签只是依据的是其当前字符可能成为某个标签的状态概率，没有从全局角度考虑序列最优问题，来为每个输入字符输出最可能的标签。例如，用 BIO 定义字符序列的标签集表示，“B”表示该字符是技能词的开头，“I”表示字符位于技能词的中间位置，“O”表示该字符不属于技能词的一部分，“I”标签通常在“B”或“I”之后，但不可能在“O”之后。CRF 层的作用便是在训练过程中自动从训练数据中学习这些规则，约束最终的预测标签以确保它们有效。如图 2-8 所示，是 Bi-LSTM-CRF 模型的详细结构：

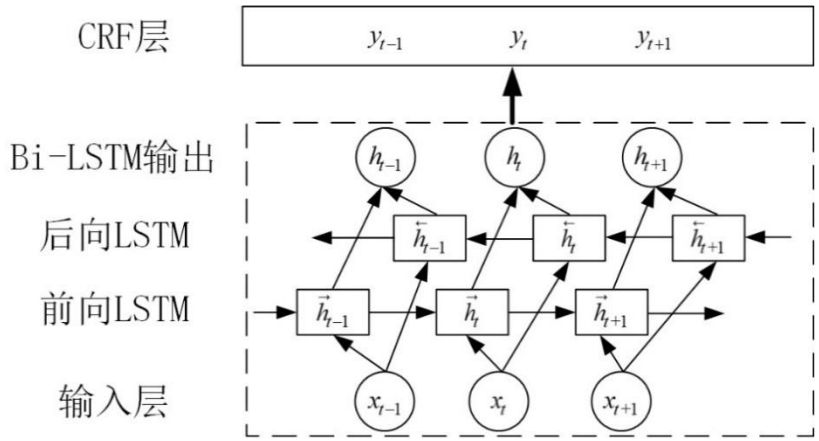


图 2-8 Bi-LSTM-CRF 模型框架

具体的，CRF 层的输入是 Bi-LSTM 层的输出，CRF 层的输出是输入序列对应的标签序列 $Y=(y_1,y_2,...,y_n)$。模型将 Bi-LSTM 的输出映射到得分矩阵 F，E 是状态转移矩阵，代表标签与标签之间的转移得分。一个输入序列 $X=(x_1,x_2,...,x_n)$，预测出的对应标注序列表示为 $Y=(y_1,y_2,...,y_n)$。则对应的预测输出分数为：

$$s(X,\mathrm{Y})=\sum_{i=0}^{n}E_{y_i,y_{i+1}}+\sum_{i=1}^{n}F_{i,y_i} \qquad (2\text{-}19)$$

其中，从标签 i 转移到标签 j 的概率为 $E_{i,j}$。句子中第 i 个字符被标记为第 j 个标签的概率为 $F_{i,j}$。模型的最佳标签序列则通过对 $s(X,Y)$ 求解出最大值而得。

§2.4 本章小结

本章对命名实体识别与术语抽取的主流技术进行了阐述。分别从传统命名实体识别与术语抽取的方法到基于深度学习的命名实体识别与术语抽取方法以及深度迁移学习方法都进行详细的介绍，这些方法是对我们后面的研究提供了很多帮助。最后对深度学习的理论基础和本研究所采用的基线模型：Bi-LSTM-CRF 模型进行了详细的介绍。

中国知网 https://www.cnki.net

第三章 网络招聘数据采集

§3.1 数据简介

根据艾瑞咨询发布的《2019 年中国网络招聘行业发展报告》[2]中提到，中国网络招聘行业发展现状从综合模式走向垂直细分，目前仍以综合模式为网络招聘主体。智联招聘(https://www.zhaopin.com/)是中国最早一批发展的综合模式的网络招聘平台，也是国内最大的在线招聘网站之一，它拥有 400 万个合作公司，每天在各个领域发布大量的招聘职位。

图 3-1 智联招聘网站中某一招聘职位的具体招聘信息

如图 3-1 所示，这是智联招聘网站中某一招聘职位的具体招聘信息，其中包括职位名称，发布时间，岗位职责要求，职位链接，职位类别，招聘人数，学历要求，经验要求，薪资和福利待遇，工作地点，公司所在地和招聘公司名称。本文要从智联招聘网站上获取这些招聘信息，尤其是招聘信息中的岗位职责要求。将这些信息存储在本地文件中用于后续实验分析。为了有效的获取这些信息，我们设计了一种具有布隆

[2] 艾瑞咨询，2019 年中国网络招聘行业发展报告，http://report.iresearch.cn/report/201907/3409.shtml.

中国知网 https://www.cnki.net

过滤器的计算机自动网络爬虫程序，系统地检索招聘网页并将其存储在数据库中，然后从检索的网页中提取特定信息。我们从 2017 年 10 月开始收集数据，一直持续至今。截至 2018 年 12 月 31 日，已获得超过 1300 万条无重复的职位数据。其中，获得 IT 行业类别数据 1,364,874 个职位，机械制造行业类别包括 617,753 个职位，服装设计行业类别包括 370,685 个职位等。我们收集的数据涵盖了中国 336 个城市，13 个行业类别和 930 个职位类别。

§3.2 数据采集

本节将主要介绍针对智联招聘网站设计的招聘信息爬虫程序的架构设计以及数据的存储，该爬虫程序基于 Python 编程语言，使用了多线程、分布式、布隆过滤器和异步队列等技术对智联招聘网站进行数据采集。

§3.2.1 爬虫的架构设计

由于智联招聘网站每日的招聘信息数量较大，并且招聘信息在不断更新。需要设计高效的爬虫架构，才能够尽可能的获取招聘网站中所有的招聘信息。本研究所设计的网络爬虫程序共分为 3 个模块，分别为：爬虫任务控制器、子爬虫实例、布隆过滤器。该爬虫还通过增加多个分布式子爬虫实例来提高爬取效率，具体架构设计如图 3-2 所示：

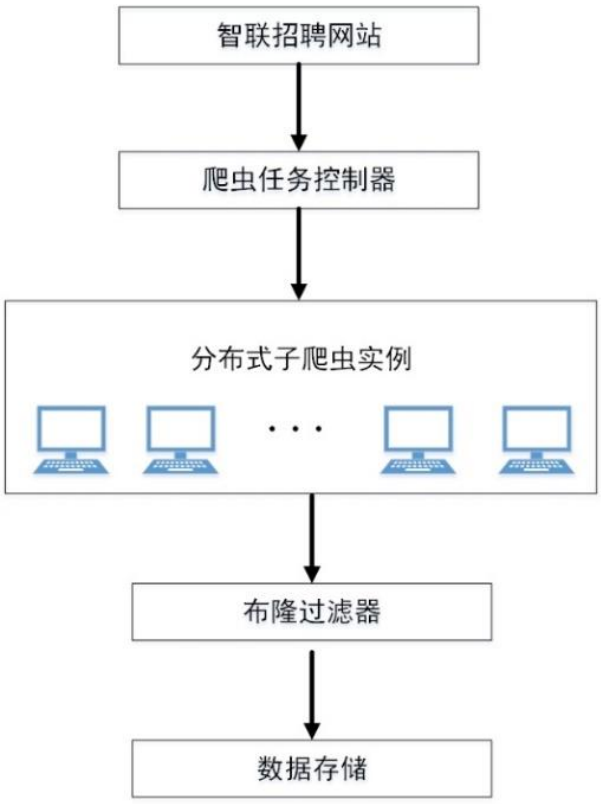


图 3-2 爬虫架构设计

中国知网 https://www.cnki.net

（1）爬虫任务控制器

在智联招聘检索页面中，用户可以按照工作地点、行业类别、职位类别等筛选条件，或直接输入关键字进行精确搜索。由于智联招聘网站设置了反爬虫机制，通过模拟用户登录方式对检索页面搜索最多能够返回 1000 条搜索结果，若想爬取所有的招聘信息，则需要尽可能组合所有的筛选条件。不过，智联招聘的前端通过 HTTP（超文本传输协议）中的 GET 请求方法将筛选条件参数提交到服务器后才返回请求搜索的结果，所以我们可以通过组合筛选条件参数方式生成 URL（统一资源定位符）。因此，经过调试检索页的前端，我们得到工作地点、行业、职位类别等筛选条件的请求参数列表逐一进行组合，得到 1.4 万个条件组合并生成出对应的搜索结果的 URL。之后，直接访问这个 URL 即可爬取网站中发布的招聘信息。

这个生成好的 URL 列表将储存在爬虫任务控制器模块中，我们将 URL 作为“任务”发送给子爬虫，子爬虫会直接访问 URL 并爬取搜索结果。由于该爬虫需要关注招聘网站的发布、删除等动态变化情况，所以该 URL 列表被设计成循环队列的形式，将分配的 URL“任务”从队头出栈，再从队尾入栈，使爬虫能够 24 小时不间断的爬取招聘网站中的数据。如下图 3-3 所示：

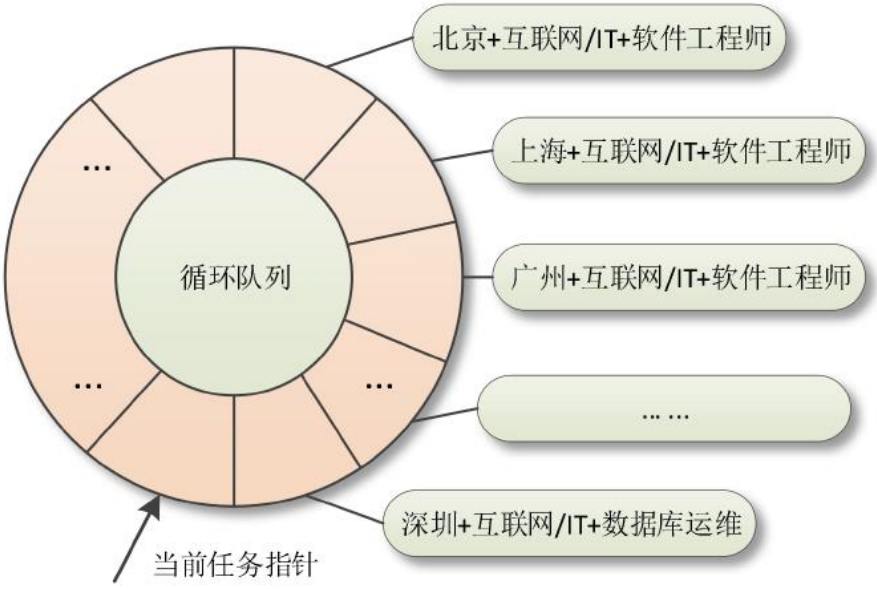


图 3-3 循环队列

（2）子爬虫实例

子爬虫是互联网上若干运行着爬虫程序的计算机，它们是分布式结构，并且不断循环执行以下任务：子爬虫首先向爬虫任务控制器请求一个 URL“任务”，然后访问这个 URL，解析返回的网页，并提取相关字符串，将其序列化为特定数据格式。最后通过 HTTP 的 POST 请求方法将提取完成的字符串发送给后一级的布隆过滤器。

由于单机运行爬虫受限于网络速度，爬取效率较低，为了解决这个问题，该爬虫

中国知网 https://www.cnki.net

将子爬虫实例设计成的分布式结构，整个爬虫系统的爬取速度与子爬虫实例的数量是线性增关系。子爬虫实例的伪代码如下所示：

```
While True:
    Step 1:
        Get A Task From Crawler Controller
    Step 2:
        Request URL From Server
        Parse Response To XML
    Step 3:
        String Matching
    Step 4:
        Send String To Bloom Filter
```

（3）布隆过滤器

为了能够尽可能的爬取整个智联招聘网站每天所发布的招聘信息，而智联招聘网站中发布的招聘信息是动态增加的。因此，本研究所提出的网络爬虫 24 小时不间断的反复爬取任务控制器中所生成的 URL“任务”。但是，通过获取的数据进对比发现，存在相同的招聘信息在当月内重复发布的现象，所以我们需要在爬取数据时加入去重机制。布隆过滤器模块便是用于接收子爬虫实例发送的信息，然后利用布隆过滤器相关算法，对招聘信息进行去重，最后将不重复的信息存储至数据库。

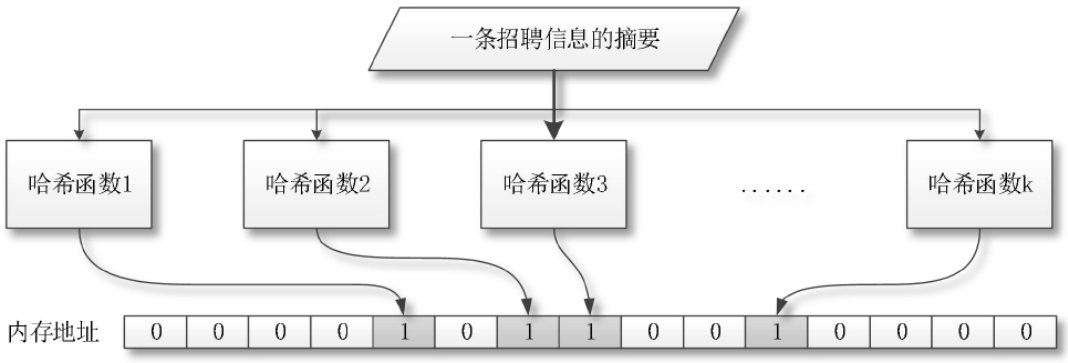


图 3-4 布隆过滤器模块

如图 3-4 所示，布隆过滤器模块在收到子爬虫实例发来的一条招聘信息时，会提取这条招聘信息的职位名称、公司名称，与当前系统日期进行字符串连接，然后进行 SHA256（Secure Hash Algorithm，256bits）安全散列计算特征值，得到 HEX 格式的摘要信息，最后将该摘要信息通过布隆过滤器，如果布隆过滤器返回“Not Exist”标识，则将这条招聘信息插入数据库，并将这条招聘信息的最后爬取时间更新至当前日期，若返回“Exist”标识，则释放该条数据。这个过程的伪代码如下：

中国知网 https://www.cnki.net

```
Begin
    Info ← Response From subCrawler
    Job_Name, Company_Name ← Parse(Info)
    mixedString ← Job_Name + Company_Name + DateTime(Now())
  HexDigest ← HEX(SHA256(mixedString))
  isExist ← BloomFilter(HexDigest)
  if isExist == Not Existed:
        Update_Database()
End
```

§3.2.2 数据存储

最终爬取到的招聘信息将会储存在 MySQL 数据库中，每条招聘信息包含以下信息：职位名称、公司名称、职位类别、行业、工作城市、薪资范围、岗位福利、招聘人数、岗位职责、首次爬取日期、最后爬取日期。数据库表结构设置如下：

表 3-1 数据库表结构设置

字段名	数据类型	含义
ID	integer	记录编号（主键）
Job_Name	varchar	职位名称
Company_Name	varchar	公司名称
Job_Type	varchar	职位类别
Industry	varchar	行业
City	varchar	工作地点
Salary	varchar	薪资范围
Welfare	varchar	福利待遇
Recruitment_Count	varchar	招聘人数
Job_Detail	Text	岗位职责
FirstTime	datetime	首次爬取日期
LastTime	datetime	最后爬取日期

为了避免数据库的读写延迟对爬虫效率的影响，经过布隆过滤器去重后的数据并未直接写入 MySQL 数据库中，而是经过基于 Redis（Remote Dictionary Server）构建的异步队列，其作用是将数据库与爬虫的其他模块相互隔离，下图是该过程的示意图：

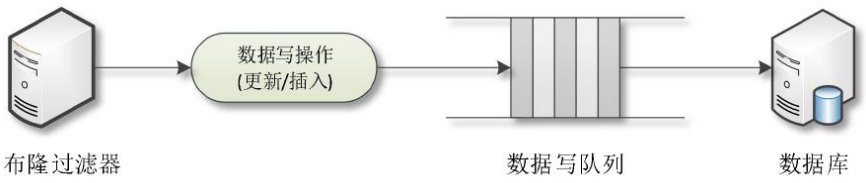


图 3-5 异步队列构建

最终，我们从 2017 年 10 月开始收集数据，一直持续至今。截至 2018 年 12 月 31

中国知网 https://www.cnki.net

日，已获得超过 1300 万条无重复的职位数据。图 3-6 是我们 MYSQL 数据库中存储的部分爬取到的招聘信息。

job_name	hash_code	applyType	city_	city_n	position_url	source_url	industry	salary	welfare	j_number	workingl	company_nun	company_t	company_size	job_type	edu_level	empl_type
律师助理（安庆）	7ba371cfc7	1	671	安庆	https://jobs.zl	https://fe-api.z	200300	4K-6K	[[每年多次调薪, 绩效	CC5966182	不限	CZ596618220	其它	100-499人	律师/法务/合规	本科	全职
专职律师	00db543bff	1	671	安庆	https://jobs.zl	https://fe-api.z	200300	10K-15K	[[节日福利, 员工旅游	CC5966182	不限	CZ596618220	其它	100-499人	律师/法务/合规	本科	全职
授薪律师（安庆）	8e888172f7	1	671	安庆	https://jobs.zl	https://fe-api.z	200300	10K-15K	[[每年多次调薪, 绩效	CC5966182	不限	CZ596618220	其它	100-499人	律师/法务/合规	本科	全职
实习律师（安庆）	3da2d0eb2	1	671	安庆	https://jobs.zl	https://fe-api.z	200300	5K-7K	[[每年多次调薪, 绩效	CC5966182	不限	CZ596618220	其它	100-499人	律师/法务/合规	本科	全职
执业律师	c3416c606e	1	671	安庆	https://jobs.zl	https://fe-api.z	200300	8K-10K	[[五险一金, 员工旅游	CZ8466481	不限	CZ846648170	民营	20人以下	律师/法务/合规	本科	全职
文档/资料管理员	b28c18158f	1	671	宿松县	https://jobs.zl	https://fe-api.z	210500	4.5K-6K	[[带薪年假, 弹性工作	CC4603939	不限	CZ460393980	民营	100-499人	行政/后勤/文秘	学历不限	全职
文档/资料管理员	8c6d11cffd	1	671	安庆	https://jobs.zl	https://fe-api.z	210500	4.5K-6K	[[带薪年假, 弹性工作	CC4603939	不限	CZ460393980	民营	100-499人	律师/法务/合规	学历不限	全职
资料员	5fc1b4c87b	1	671	安庆-迎	https://jobs.zl	https://fe-api.z	140000,1	3K-4K	[[五险一金, 绩效奖金	CC8065912	1-3年	CZ806591220	民营	100-499人	行政/后勤/文秘	大专	全职
资料员文员	a8147570511	1	671	安庆-迎	https://jobs.zl	https://fe-api.z	140000	2K-4K	[[]]	CC8638750	1-3年	CZ863875080	民营	20-99人	行政/后勤/文秘	大专	全职
Java储备岗（无经	954a91df9c	1	671	安庆-迎	https://jobs.zl	https://fe-api.z	160400	4K-6K	[[]]	CC8411058	无经验	CZ841105850	合资	20-99人	软件/互联网开发/	学历不限	实习
嵌入式软件工程师	967cb67d8	1	671	安庆	https://jobs.zl	https://fe-api.z	121500	8K-10K	[[周末双休, 五险一金	CZ4159406	1-3年	CZ415940620	国企	100-499人	软件/互联网开发/	本科	全职
软件工程师	2afd625517	1	671	安庆	https://jobs.zl	https://fe-api.z	121500	8K-10K	[[住房补贴, 五险一金	CZ4159406	1-3年	CZ415940620	国企	100-499人	软件/互联网开发/	本科	全职
软件开发工程师	f69e3ef6c2f	1	671	安庆	https://jobs.zl	https://fe-api.z	160400,1	4K-6K	[[五险一金, 绩效奖金	CC2281491	1-3年	CZ228149120	民营	20-99人	软件/互联网开发/	本科	全职
软件开发助理工程	e03758d0cc	1	671	安庆	https://jobs.zl	https://fe-api.z	160400,1	3.5K-5K	[[五险一金, 绩效奖金	CC2281491	不限	CZ228149120	民营	20-99人	软件/互联网开发/	大专	全职
软件开发高级工程	c7537f411d	1	671	安庆	https://jobs.zl	https://fe-api.z	160400,1	6K-8K	[[五险一金, 加班补贴	CC2281491	不限	CZ228149120	民营	20-99人	软件/互联网开发/	学历不限	全职
Java开发工程师	fc1ac97602	1	671	安庆	https://jobs.zl	https://fe-api.z	160000,1	4K-4.5K	[[]]	CC2049324	1-3年	CZ204932420	民营	100-499人	软件/互联网开发/	本科	全职
项目实施工程师	0d20965b1	1	671	安庆	https://jobs.zl	https://fe-api.z	160400,1	4K-5K	[[五险一金, 绩效奖金	CC2281491	不限	CZ228149120	民营	20-99人	IT运维/技术支持	大专	全职
电气工程师	52dd7c013	1	671	安庆	https://jobs.zl	https://fe-api.z	140100	8K-10K	[[五险一金, 带薪年假	CC2920705	3-5年	CZ292070530	民营	100-499人	电子/电器/半导体	本科	全职
电气工程师	f42c4c5c26	1	671	安庆	https://jobs.zl	https://fe-api.z	160500	4K-8K	[[全勤奖, 加班补助,	CC3817528	3-5年	CZ381752820	民营	500-999人	电子/电器/半导体	大专	全职
变电电气一次主设	c87c1d0aac	1	671	安庆	https://jobs.zl	https://fe-api.z	130100,1	2K-4K	[[五险一金, 餐补, 带	CC2746976	无经验	CZ274697630	民营	20-99人	电子/电器/半导体	本科	校园
电气工程师(安徽	1aff2dc9f8c	1	671	安庆	https://jobs.zl	https://fe-api.z	140200,2	4K-6K	[[五险一金, 包吃, 弹	CC8084258	1-3年	CZ808425820	合资	20-99人	电子/电器/半导体	中专	全职
电气主管	68574027b	1	671	安庆	https://jobs.zl	https://fe-api.z	120500	6K-8K	[[五险一金, 年底双薪	CC6741474	5-10年	CZ674147430	上市公司	1000-9999人	电子/电器/半导体	本科	全职
电气工程师	89be91a3d	1	671	安庆	https://jobs.zl	https://fe-api.z	129900	5K-8K	[[带薪年假, 交通补贴	CZ8974703	3-5年	CZ897470340	民营	20-99人	电子/电器/半导体	本科	全职
电气工程师（储备	73e3800d6	1	671	安庆	https://jobs.zl	https://fe-api.z	129900	3K-4K	[[带薪年假, 绩效奖金	CZ8974703	1-3年	CZ897470340	民营	20-99人	电子/电器/半导体	大专	全职
电气工程师	07a91a063	1	671	安庆	https://jobs.zl	https://fe-api.z	129900	6K-8K	[[绩效奖金, 加班补贴	CC3813233	3-5年	CZ381323330	股份制企业	20-99人	电子/电器/半导体	大专	全职

图 3-6 网络爬虫获得的部分招聘信息

§3.3 本章小结

本章主要介绍了本研究所使用的网络招聘信息的数据来源，以及介绍了基于智联招聘网站的数据获取方案，并对爬虫架构设计与数据存储进行了详细的阐述。最后，对已获取到的全部网络招聘信息进行统计。

中国知网 https://www.cnki.net

第四章 基于深度学习的中文网络招聘文本中的技能词抽取

§4.1 问题描述及分析与挑战

通过对网络招聘文本中技能词的分析，可以看出技能词抽取与命名实体识别和术语抽取相比，具有其自身的特殊性与复杂性，具体表现在：1)部分技能词类似于术语，用于表达各专业领域的特殊概念，仅仅在该领域特定上下文中才表示为技能词。比如，“操作系统”、“自然语言处理”等；2)部分技能词具有与命名实体相似的特点，是语料中具有很强特定意义的词语，并且不受上下文影响，在任何场景下出现都表示为技能词。比如，“Java”、“Linux”、“算法”等；3)技能词不同于一般常见的命名实体（人名、地名、机构名等），没有明确的关于技能词的定义，不能清晰的界定技能词的边界，并且技能词形式多变，表现在长度、组成模式等方面。例如:“C#”，“数据库/表结构索引设计”，还有类似于“Deep Learning 框架”和“Neo4j”等中英文混合、数字英文混合词语等形式；4)大量技能词在表述时多采用英译词或英文缩写。比如，“J2EE”和“JavaEE”等；5)已有词表或者技能词图谱不足以涵盖全部技能词，而且随着技术的不断进步，新的技能词会不断出现。比如，“联邦学习”、“胶囊网络”等；6)技能词存在嵌套形式，比如“Linux 操作系统”，其中“Linux”和“操作系统”又分别可以单独作为技能词出现。此外，不同于正式文本，招聘信息通常很不规范，很容易出现拼写错误的技能词。如，将技能词“Excel”误写成“Excle”、“SPSS”错拼为“SPASS”等。目前，虽然命名实体识别的相关工作很多，但重点都是识别正式文本中常见的命名实体，而术语抽取的相关工作主要针对特定领域内的术语识别，缺乏通用性和可移植性。由此可见，从招聘文本中抽取技能词仍然是一项颇为艰巨的任务。

本研究选取 IT 类网络招聘数据作为研究对象，分析了技能词的特点，对从招聘网站上获取的招聘语料进行预处理后根据相应的标注规则进行人工标注，充分挖掘招聘语料的各类型特征，分析了各类语料特征对识别结果所起的作用。本研究的主要贡献包括 4 个方面:1) 据我们所知，本研究是首次使用基于深度学习的序列标注模型与语料特征相结合的方法抽取网络招聘中的技能词的研究；2) 与经典的序列标注模型相比，本研究提出的模型引入了更多的语料特征并将输入层的输出与 Bi-LSTM 输出连接在一起作为 CRF 层的输入，模型的性能得到了极大提高；3) 通过实验评估了本研究所提出的模型中加入的语料特征是否能够缓解模型对大量标注数据的依赖；4) 进行了大量实验。实验结果表明，本研究提出的技能词模型具有最佳的抽取性能。

中国知网 https://www.cnki.net

§4.2 基于深度学习的技能词抽取模型构建

在本节中，我们将详细描述本研究提出的技能词抽取模型。如图 4-1 所示，本研究提出的模型由四个层次的模块组成，即输入层、Bi-LSTM 层、特征拼接层和 CRF 层。在输入层中，将每个输入语句转换为一系列字符特征向量，再与输入语句中各个字符的位置特征（Seg）、输入语句分词后的词性特征（Pos）和技能词的上下文特征（Con）进行拼接，将其输入到 Bi-LSTM 层中，以将上下文信息顺序编码为固定长度的隐藏向量，接着在特征拼接层中将输入层的输出与 Bi-LSTM 层的输出连接在一起，作为 CRF 层的输入。最后由 CRF 层预测出最佳标签序列作为整个网络的输出。接下来，我们详细介绍这四个组成部分。

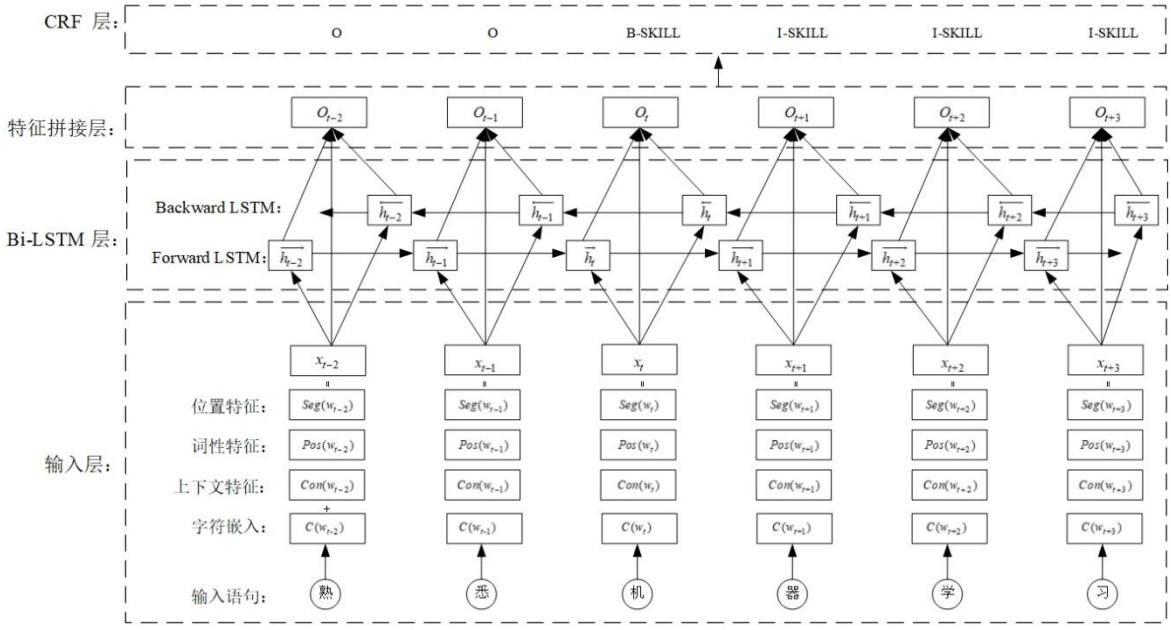


图 4-1 基于深度学习的技能词抽取模型架构

§4.2.1 输入层

在输入层中，我们分为两个步骤。首先是将输入语句转换为字符级密集向量序列。在这一步骤中，先生成一个包含语料库中所有字符的字典，再用一个嵌入矩阵 $M \in R^{D \times V}$ 用于将每个字符映射到一个密集矢量中，其中 D 是嵌入向量维度，而 V 是字典中所有字符的总量。输入句子表示为 $S=[w_1, w_2, ..., w_n]$，其中 n 是句子的长度，$w_i \in R^V$ 是第 i 个字符在字典中的one-hot表示。句子的字符嵌入向量表示为，其中 $c_i = M_{w_i} \in R^D$。

第二步加入的是网络招聘信息中各类语料特征，与字符嵌入向量进行拼接。语料特征主要由以下三种特征构成：位置特征（Seg）、词性特征（Pos）和上下文特征（Con）。

位置特征（Seg）是指对输入句子进行 jieba[3]分词后，每个字符与所在词语的相对位置。如：“操作系统”是分词后得到的词，那么“操”的位置特征标记为“0”，“作”的位

[3] jieba 分词工具：https://github.com/fxsjy/jieba

中国知网 https://www.cnki.net

置特征标记为“1”，“系”的位置特征标记为“2”，“统”的位置特征标记为“3”。词性特征（Pos）是指将输入语句进行 jieba 分词后将每个字符的词性标记为所在词语所对应的词性。如：“具备”的词性为“动词”，则“具”、“备”的词性都记为“动词”。根据术语特性可知，有些词依赖于领域特定上下文，用于表达各专业领域的特殊概念，只在本领域流通。而本研究所抽取的技能词大多都属于特定领域的专业知识和专业能力，因此可以考虑在技能词抽取时加入词语的位置特征。同样，通过对技能词的分析可知，虽然技能词的构成模式有很多种，但是大部分是名词性短语，可见词性对于技能词抽取是另一个重要的特征。

上下文特征（Con）是根据技能词的上下文特点构造的特征。本节接下来将对其详细介绍。首先，通过分析招聘语料库，我们随机抽取了 1000 多条网络招聘文本后发现：包含技能词的文本通常为动宾结构，且技能词大多数为“名词/名词性短语”。如“熟悉关系型数据库”。技能词在句子中的位置主要位于动词，形容词/形容词短语或“和”、“或”、以及“、”等之后。例如：“了解 Python”、“常用的深度学习模型”、“掌握文本挖掘、实体抽取、词性标注等技术）。表 4-1 统计了网络招聘语料中技能词出现的位置。通过分析技能词出现的下文，我们发现：其下文使用了较多的习惯语，如：“掌握 XX 能力”和“具有 XX 经验”等。

表 4-1 网络招聘语料中技能词出现的位置

语料数量	技能词数量	动词后	形容词/ 形容词短语后	“和”、“或”、“、” 等并列形式	其他位置
1006	8126	2739	1979	2831	577

因此，在标注上下文特征时，首先将输入语句进行jieba分词，提取出每个词的词性．然后依据如下规则进行标注：1)若“动词”之后出现“名词”，则将该“动词”词语组成的字符都标注为“1”，该“名词”词语组成的字符都标注为“0”。如，“掌握”之后出现“名词”，则“掌”和“握”都标注为“1”；2)若“动词”之后出现其他词性词语，则将该“动词”和其他词性词语组成的字符都标注为“0”；3)若“形容词/形容词短语”之后出现“名词”，则将该“形容词/形容词短语”词语组成的字符都标注为“2”，该“名词”词语组成的字符都标注为“0”；如，“常用的”之后出现“名词”，则“常”、“用”和“的”都标注为“2”。4) 若“形容词/形容词短语”之后出现其他词性词语，则将该“形容词/形容词短语”和其他词性词语组成的字符都标注为“0”；5)若以并列形式，连续出现两个或两个以上的“名词”，则将并列形式的连接字符，如“、”与“和”等都标注为“3”；6)若以“动词 +名词+名词”形式出现，则将后一个“名词”词语组成的字符都标注为“4”，该“动词”词语组成的字符都标注为“1”，第一个“名词”词语组成的字符都标注为“0”。如“具有XX能力”，则“具”、“有”都标注为“1”，“能”、“力”都标注为“4”；7)除此以外，输入语句中的其他字符都标注为“0”。关于如何给每个字符标注上下文特征，举一个例子，如输入语句

中国知网 https://www.cnki.net

为“具备数据库和数据结构基础。”将其标注为“具/1 备/1 数/O 据/O 库/O 和/3 数/O 据/O 结/O 构/O 基/4 础/4 。/O”。具体技能词的上下文特征词如表4-2所示。

表 4-2 技能词的上下文特征词

特征类型	特征词
动词后	熟悉、具有、具备、使用、精通、熟练、理解、编写、绘制、挖掘、进行、实现、设计、掌握、通晓、处理、解决、学会…
形容词/形容词短语后	常用的、主流、扎实的、相关、优秀的、常见的、丰富的、实际的、复杂的、较好的、基本的、良好的、常识性、通用的…
“和”、“或”、“、”等	或、\、并、“、”、和、与、及、等…
下文特征	优先、工具、、经验、实现、基础、能力、经历、成果、效果、技术、细节、理论、操作、系统、工作、用法、原理、方向…

最后，输入层的输出由各输入语句序列中每个节点的字符特征向量、位置特征向量（Seg）、词性特征向量（Pos）和上下文特征向量（Con）四组特征向量构成，即 $X_i=[c_i,seg_i,pos_i,con_i]$。例如：输入语句为“具备数据库和数据结构基础”，词性特征向量表示为：“0,0,1,1,1,2,1,1,1,1,1,1”，其中“0”代表动词，“1”代表名词，“2”代表连词。位置特征向量表示为：“0,1,0,1,2,0,0,1,2,3,0,1”，其中“0”代表所在词语的第一个字符，“1”代表所在词语的第二个字符，…。上下文特征表示为： “1,1,0,0,0,3,0,0,0,0,4,4”。之后，将这四组特征连接成向量作为Bi-LSTM层的输入。

§4.2.2 Bi-LSTM 层

LSTM 是一种特殊类型的循环神经网络，它可以捕获长距离序列信息，并且在序列数据建模方面功能强大。与标准 RNN 的区别在于，LSTM 在隐藏层的神经元中加入细胞状态和输入门，遗忘门，输出门。细胞状态的更新时需要同时使用输入门和遗忘门结果。具体实现是：

$$i_t=\sigma\left(W_{xi}X_t+W_{hi}h_{t-1}+W_{ci}c_{t-1}+b_i\right) \quad (4\text{-}1)$$

$$f_t=\sigma\left(W_{xf}X_t+W_{hf}h_{t-1}+W_{cf}c_{t-1}+b_f\right) \quad (4\text{-}2)$$

$$c_t=f_t\otimes c_{t-1}+i_t\otimes\tanh\left(W_{xc}X_t+W_{hc}h_{t-1}+b_c\right) \quad (4\text{-}3)$$

$$o_t=\sigma\left(W_{xo}X_t+W_{ho}h_{t-1}+W_{co}c_t+b_o\right) \quad (4\text{-}4)$$

$$h_t=o_t\otimes\tanh\left(c_t\right) \quad (4\text{-}5)$$

其中，σ是 logistic sigmoid 激活函数，$\otimes$是元素的乘积。在t时刻，i， f ，o和c分别代表：输入门、遗忘门、输出门和细胞状态。输入门，输出门和遗忘门由 sigmoid 激活函数实现，细胞状态由三个门控制。权重矩阵W 下标表示每个门之间的连接,b是偏置。例如,W_{xi}是输入节点 X_t 与输入门之间权重矩阵、W_{hi} 表示$t-1$时刻隐藏层状态 h_{t-1}与输入门之间权重矩阵、W_{ci} 表示$t-1$时刻细胞状态c_{t-1}与输入门之间权重矩阵。

但是，LSTM 只能看到t时刻之前得历史信息，而不能看到t时刻之后的将来信息。而 Bi-LSTM 可以从全局上下文中学习字符的隐藏特征表示。对于从输入层输出的包含n个字符的序列$[X_1,X_2,...,X_n]$。将$\overrightarrow{LSTM}$表示为使用LSTM网络从左到右扫描语句，

中国知网 https://www.cnki.net

则 $\overrightarrow{LSTM}$ 学习的隐藏表示可以表示为$[\overrightarrow{h_1},\overrightarrow{h_2},...,\overrightarrow{h_n}]$。类似地，$\overleftarrow{LSTM}$ 表示 LSTM 网络从右到左扫描句子，$[\overleftarrow{h_1},\overleftarrow{h_2},...,\overleftarrow{h_n}]$表示由 $\overleftarrow{LSTM}$ 学习到的隐藏表示。第i个字符隐藏表示是通过将拼接其 $\overrightarrow{LSTM}$ 和 $\overleftarrow{LSTM}$ 上下文表示得 $h_i=[\overrightarrow{h_i},\overleftarrow{h_i}]$。Bi-LSTM 层的输出为 $h=[h_1,h_2,...,h_n]$，其中 $h_i\in R^{2F}$，F 是 LSTM 网络中隐藏特征表示的维数。Bi-LSTM 输出的是每个字符的上下文特征拼接后的向量。

§4.2.3 特征拼接层

特征拼接层将 Bi-LSTM 的输出与前面输入层中的字符嵌入向量、位置特征（Seg）、词性特征（Pos）和下文特征（Con）拼接在一起，以期更好地提升模型的识别准确率。拼接后，第i个字符的特征表示为 $A_i=[h_i,x_i]$。其中，X_i 为输入层中第 i 个字符的输出，h_i 为 Bi-LSTM 层中第 i 个字符的输出。此外，为了避免简单的线性组合，增强神经网络模型的非线性因素，本研究提出在特征拼接层输出之前，将 Bi-LSTM 层与输入层的输出拼接后的向量 A_i 映射成 L 维向量。其中，L 是技能词标签集中标签的数量。即 $O_i=\tanh(w*A_i+b),O_i\in R^L$。其中，tanh 为激活函数，$w$ 是映射的权重，b 是偏置，将 $\varphi=\{w,b\}$ 记为特征拼接层中的参数集。最终，特征拼接层的输出为 $O=[O_1,O_2,...,O_n]$。

§4.2.4 CRF 层

从特征拼接层输出的向量可以直接用作特征，采用 softmax 函数输出每个字符最可能的分类标签。但是，在序列标注任务中，每个输入字符的标签都涉及上下文语义关系，相邻标签通常具有很强的依赖性。而由 softmax 函数输出的字符分类标签只是依据其当前字符可能成为某个标签的状态概率，没有从全局角度考虑序列最优为每个输入字符输出最可能的标签。例如，用 BIO 定义字符序列的标签集表示，“B”表示该字符是技能词的开头，“I”表示字符位于技能词的中间位置，“O”表示该字符不属于技能词的一部分，“I”标签通常在“B”或“I”之后，但不可能在“O”之后。CRF 层作用便是在训练过程中自动从训练数据中学习这些规则，约束最终的预测标签以确保它们有效。

将 $y=[y_1,y_2,...,y_n]$ 表示为句子 S 的标签序列，其中 $y_i\in R^L$ 是第 i 个字符标签的 one-hot 表示。给定输入 O 的标签序列 y 的条件概率计算如下[20]：

$$p(y\mid O,\theta)=\frac{\prod_{i=1}^{n}\Psi(O_i,y_i,y_{i-1})}{\sum_{y'\in Y(S)}\prod_{i=1}^{n}\Psi(O_i,y_i',y_{i-1}')} \tag{4-6}$$

其中，$Y(S)$ 是句子 S 的所有可能标签序列的集合，θ 表示参数集，而 $\Psi(O_i,y_i,y_{i-1})$ 是势函数，如下所示：

$$\Psi(O_i,y_i,y_{i-1})=\mathrm{e}^{\left(y_i^T E^T O_i+y_{i-1}^T F y_i\right)} \tag{4-7}$$

其中，$E\in R^{L\times L}$，$F\in R^{L\times L}$ 是 CRF 层的参数，即 $\theta=\{E,F\}$。该方法的损失函数为：

$$Loss=-\sum_{S\in corpus}\log\left(p(y_S\mid O_S,\theta)\right) \tag{4-8}$$

中国知网 https://www.cnki.net

其中 *corpus* 是训练数据集中的所有语句。

§4.2.5 训练过程

本研究使用学习率为 0.001 的小批量自适应矩估计（Adam）优化算法[70]以端到端的方式训练技能词抽取模型。在训练期间，我们使用每批次 20 条语句训练 100 个周期。换句话说，在每个周期中，语料库中 1006 条语料被随机分为 51 个批次进行批次训练，每个批次不超过 20 个句子。对于每一批次，我们采用预先训练的字符嵌入而非随机初始化的嵌入作为输入字符的嵌入特征向量。在 CRF 层中，我们使用动态编程来计算公式（4-6）的结果并预测标签序列。最后，反向传播时根据 CRF 层预测出的标签序列与真实标签序列之间的误差，依次更新 CRF 层中的参数集 θ、特征拼接层中的参数集 φ 以及 Bi-LSTM 层中的所有的权重矩阵和网络参数并保存模型。

§4.3 实验设置与结果及分析

在本节中，为了验证我们提出的技能词抽取模型的有效性，将本研究的模型与主流方法进行比较。我们将在实验中详细描述数据集、标注策略、预训练的字符向量、模型参数设置、结果分析。

§4.3.1 数据集与评价指标

在此实验中，我们选择从智联招聘网站中爬虫获取的 IT 行业类别的岗位招聘数据作为语料库。将每条职位招聘数据中的“岗位职责要求”视为一条语句。由于时间和手工标注成本的限制，我们仅标注了 1006 条“岗位职责要求”作为实验语料。在标注语料时，我们尝试选择句法结构标准的句子作为研究语料，并将语句中的专业课程名称，专业知识点和相关专业工具标记为技能词，例如“C 语言程序设计”，“堆排序”，“SpringMVC 框架”等。因此，一条语句将被标记为：“熟/ O 悉/ O 机/ B-SKILL 器/ I-SKILL 学/ I-SKILL 习/ I-SKILL 与/ O 自/ B-SKILL 然/ I-SKILL 语/ I-SKILL 言/ I-SKILL 处/ I-SKILL 理/ I-SKILL。/O”。

由于没有明确的标准如何将数据集划分为训练/验证/测试集，因此我们在实验中将数据集进行两轮交叉验证。首先将整个数据集划分成 5 份，选择其中的 80%作为训练数据，而将其余 20%用作测试数据。然后，再次将训练数据划分成十份，选择其中的 90%作为最终训练数据，并将其余的 10%用作验证数据。换句话说，将全部数据的 72％作为训练数据，将 8％作为验证数据，将 20％作为测试数据。每次实验的训练周期为 100 个周期，通过验证集找出训练集在这 100 个周期内最佳网络模型参数后再使用测试集进行测试，以获得本次验证的结果。表 4-3 中分别展示了用于训练，验证和测试数据集的句子数，技能词数量。值得注意的是，由于每个句子中包含的技能词数量不同，训练/验证/测试数据集中的技能词数量会随着每次划分而动态变化，因此，

中国知网 https://www.cnki.net

技能词数量的范围也在表中。

实验采用机器学习中常用的准确率（Precision）、召回率（Recall）指标评价技能词抽取模型的性能，并采用 F1 值指标评价其综合性能。

表 4-3 数据集的统计信息

数据集	数据类型	技能词类型	技能词数量	语句数量
招聘数据	训练集	SKILL	[5836,6586]	725
	验证集	SKILL	[593,817]	80
	测试集	SKILL	[1577,1903]	201

§4.3.2 模型参数设置

为了考虑批量训练样本量大小（batch）和学习率对本研究提出的模型的影响，我们通过如下实验进行超参数选择。首先，根据已有的参考文献中批量训练样本量大小（batch）和学习率的使用，确定出 batch 大小和学习率的选择范围，分别为[20、50、100]和[0.001、0.002、0.005]。其次，在交叉验证的实验方案基础上，我们尝试使用不同批量训练样本量大小（batch）和学习率的参数组合，结果如图 4-2 所示。

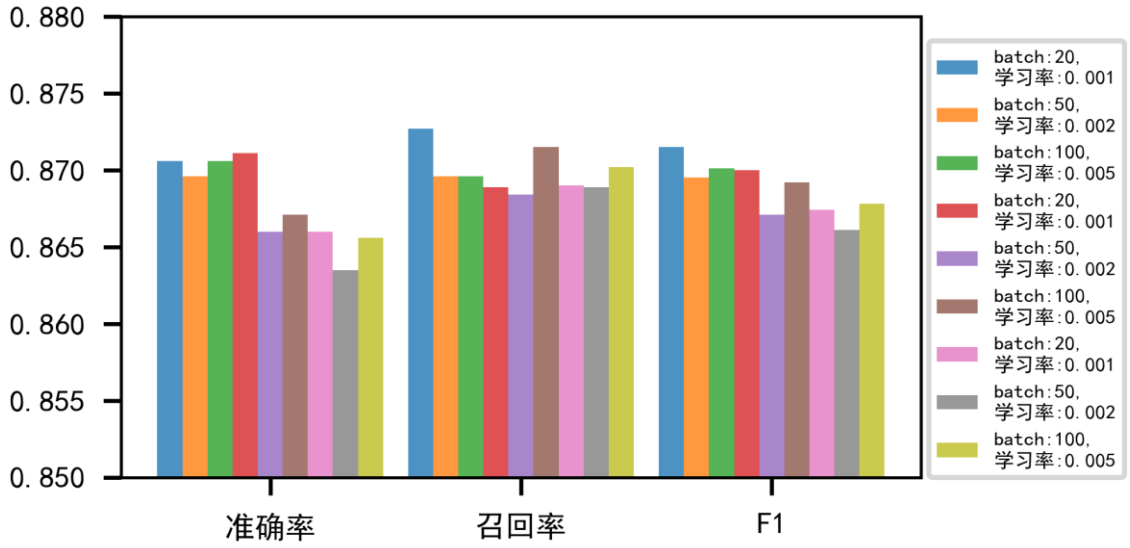


图 4-2 本研究模型采用不同批次大小和学习率时准确率、召回率、F1 的对比

从图 4-2 可以发现当批量训练样本量大小（batch）设置为 20，学习率设置为 0.001 时，该模型可获得最佳总体效果。因此，我们认为最合适的批次大小为 20，最佳学习率为 0.001。

模型的其他超参数的初始值设置：字符嵌入特征向量设置为 100。本研究使用的是 word2vec 中的 CBOW 模型对中文维基百科语料预训练[67]的字符向量，epoch 次数固定为 100，隐藏层维度固定为 100，其他网络参数包括特征拼接层中的参数集 φ 和 Bi-LSTM 层中所有权重矩阵的初始值，以及 CRF 层的参数集 θ 在[-1，1]范围内随机地均匀初始化。

中国知网 https://www.cnki.net

§4.3.3 实验结果分析

实验一：为了验证本研究提出的技能词抽取模型中加入的各类语料特征对技能词抽取有效性的影响，进行了如下实验，结果如表 4-4 所示。本研究选择基于字符级的 Bi-LSTM-CRF 模型作为基线对比实验，此时只将字符嵌入特征输入网络，并不输入字符的语料特征。然后，分别在 Bi-LSTM-CRF 模型的输入层中加入不同类型的语料特征，如 Model_1 代表在 Bi-LSTM-CRF 的输入层中加入字符的位置特征（Seg），但并没有将输出层的输出与 Bi-LSTM 层的输出进行拼接，Model_4 代表在 Model_1 基础上加入技能词的上下文特征（Con），即在 Bi-LSTM-CRF 的输入层中加入所有语料特征。Model_8 代表在 Bi-LSTM-CRF 的输入层中加入所有语料特征，并且将输出层的输出与 Bi-LSTM 层的输出进行拼接，即本研究最终的技能词抽取模型。

从表 4 可以看出，在 Bi-LSTM-CRF 模型的输入层中分别加入位置特征（Seg），词性特征（Pos）和技能词的上下文特征（Con）时，相比于 Bi-LSTM-CRF 模型的 F1 值分别提升了 0.44%、0.35%和 7.66%。其中，加入上下文特征获得的提升最大，这是因为：由于招聘语料的句法结构比较单一，技能词的上下文特征较为固定，充分挖掘技能词上下文特征能较好的反映技能词在语料中出现的位置，能有效地抽取出“…具备数据库开发能力…”、“…常用的 Java、C、python 等编程语言…”这类具有明显句法结构的技能词，从而使训练出的模型更具有泛化能力。而字符的位置特征（Seg）和词性特征（Pos）的加入，虽然提高了 F1 值，但提升的幅度不是很大。其中可能的原因是：词语的位置特征（Seg）是通过对句子进行 jieba 分词，再提取每个字符与所在词语的相对位置而得到的。中文分词结果的不准确性，影响了字符位置特征的提取，从而对技能词的抽取带来了一定程度的干扰。对于词性特征（Pos），则可能是因为 jieba 分词无法标注出英文字符的词性，以及技能词的构成在词法特征上规律性不强。

从表 4-4 中还可以看出：如果同时将字符的位置特征（Seg）和词性特征（Pos）加入到 Bi-LSTM-CRF 模型的输入层中。相比于 Bi-LSTM-CRF，模型的 F1 值提升了 0.5%。如果同时将字符的位置特征（Seg）和技能词的上下文特征（Con）加入到 Bi-LSTM-CRF 模型的输入层中。相比于 Bi-LSTM-CRF，模型的 F1 值提升了 7.83%。如果同时将词性特征（Pos）和技能词的上下文特征（Con）加入到 Bi-LSTM-CRF 模型的输入层中。相比于 Bi-LSTM-CRF，模型的 F1 值提升了 7.78%。如果同时将字符的位置特征（Seg），词性特征（Pos）和技能词的上下文特征（Con）都加入到 Bi-LSTM-CRF 模型的输入层中，模型的 F1 值能得到更进一步的提升。相比于 Bi-LSTM-CRF，模型的 F1 值从 0.7892 提高到了 0.8706，提升幅度达到 8.14%。说明在模型中加入的语料特征越多，越能有利于模型的技能词抽取。

另外，我们在 Bi-LSTM-CRF 的输入层中加入所有语料特征的同时，并将输入层的输出和 Bi-LSTM 层的输出拼接，即本研究所最终提出的技能词抽取模型。相比于

中国知网 https://www.cnki.net

只加入语料特征的情况，又能更进一步的提升了模型的 F1 值，提升幅度达到 8.23%。最终，我们可以得出结论，本研究的技能词抽取模型能够有效的进行技能词抽取，并且抽取性能得到了极大提高，加入各类语料特征也都有利于技能词抽取性能的提升。

表 4-4 各种语料特征加入后技能词抽取的性能

模型编号	模型特征	Precision	Recall	F1	提升
Model_0	Bi-LSTM-CRF	0.7890	0.7899	0.7892	--
Model_1	Model_0+Seg	0.7919	0.7958	0.7936	+0.44%
Model_2	Model_0+Pos	0.7905	0.7951	0.7927	+0.35%
Model_3	Model_0+Con	0.8662	0.8655	0.8658	+7.66%
Model_4	Model_0+Seg+Con	0.8676	0.8675	0.8675	+7.83%
Model_5	Model_0+Pos+Con	0.8676	0.8666	0.8670	+7.78%
Model_6	Model_0+Seg+Pos	0.7914	0.7973	0.7942	+0.5%
Model_7	Model_0+Seg+Pos+Con	**0.8716**	0.8698	0.8706	+8.14%
Model_8	技能词抽取模型	0.8706	**0.8727**	**0.8715**	**+8.23%**

实验二：为了验证加入各类语料特征在不同规模训练集下也是否都有利于模型抽取性能的提升，以及评估引入丰富的语料特征是否能够缓解模型对大量标注数据需求的依赖。在实验一的基础上，本组实验进一步从训练集中抽取 25%、50%和 75%的样本，同时保持测试集不变，进行实验，实验结果如表 4-5 所示。

样本抽取的具体方法如下：在每次使用两轮交叉验证方法实现对数据的划分后，再将训练数据平均分成 N 份，每次选择其中若干份作为最终训练集，并重复的进行多次实验。25%的训练集的抽取方案为：将训练集划分为 4 份，每次选择 1 份作为最终训练集，实验重复 4 次；50%的训练集抽取方案为：将训练集划分成 2 份，每次选择 1 份作为最终训练集，实验重复 2 次；75%的训练集抽取方案为：将训练集划分成 4 份，每次选择 3 份作为最终训练集，实验重复 4 次。本组实验在不同训练集比例下进行，同样也是选择 Bi-LSTM-CRF 模型作为基线对比实验。

如表 4-5 所示，在不同规模的训练集下，本研究的模型抽取性能相比于 Bi-LSTM-CRF 模型仍然有较大的提升，在 25%、50%和 75%训练集比例下，F1 值分别由 0.7267、0.7646 和 0.7807 提高到了 0.8336、0.8517 和 0.8646。另外，在不同比例的训练集下，加入的各类型语料特征也依然有利于模型抽取性能的提升，并且可以得出与实验一中使用全部训练集同样的结论，即在模型中加入的语料特征越多，越能有利于模型的技能词抽取。例如：在 25%、50%和 75%训练集比例下，仅加入词性特征（Pos），相比于 Bi-LSTM-CRF 模型的 F1 值分别提升了 0.74%、0.18%和 0.48%。若同时将字符的位置特征（Seg）和词性特征（Pos）加入到 Bi-LSTM-CRF 模型的输入层中，相比于 Bi-LSTM-CRF 模型的 F1 值分别提升了 1.32%、0.44%、0.46%。而同时将字符的位置特征（Seg），词性特征（Pos）和技能词的上下文特征（Con）都加入到 Bi-LSTM-CRF 模型的输入层中，模型的 F1 值能得到更进一步的提升。相比于 Bi-LSTM-CRF 模型，F1 值分别从 0.7267、0.7646 和 0.7807 提高到了 0.8267、0.8491 和 0.8623，提升幅度

中国知网 https://www.cnki.net

分别达到 10.0%、8.45%和 8.16%。

表 4-5 不同训练集比例下加入各类语料特征后的技能词抽取性能

训练集比例	25%	50%	75%
模型	P，R，F1，提升	P，R，F1，提升	P，R，F1，提升
Model_0	0.7146, 0.7402, 0.7267, --	0.7549, 0.7750, 0.7646, --	0.7769, 0.7848, 0.7807, --
Model_1	0.7234, 0.7512, 0.7367, +1.00%	0.7601, 0.7783, 0.7688, +0.42%	0.7841, 0.7882, 0.7859, +0.52%
Model_2	0.7195, 0.7500, 0.7341, +0.74%	0.7573, 0.7761, 0.7664, +0.18%	0.7831, 0.7883, 0.7855, +0.48%
Model_3	0.8124, 0.8244, 0.8172, +9.05%	0.8413, 0.8456, 0.8433, +7.87%	0.8511, 0.8562, 0.8535, +7.28%
Model_4	0.8155, 0.8293, 0.8221, +9.54%	0.8451, 0.8529, 0.8489, +8.43%	0.8616, 0.8551, 0.8583, +7.76%
Model_5	0.8158, 0.8227, 0.8191, +9.24%	0.8443, 0.8481, 0.8461, +8.15%	0.8563, 0.8573, 0.8567, +7.60%
Model_6	0.7247, 0.7563, 0.7399, +1.32%	0.7610, 0.7775, 0.7690, +0.44%	0.7818, 0.7891, 0.7853, +0.46%
Model_7	0.8195, 0.8341, 0.8267, +10.0%	0.8459, 0.8525, 0.8491, +8.45%	0.8635, 0.8613, 0.8623, +8.16%
Model_8	**0.8291, 0.8382, 0.8336, +10.7%**	**0.8498, 0.8537, 0.8517, +8.71%**	**0.8644, 0.8648, 0.8646, +8.39%**

从表 4-5 中还可以看出：在 Bi-LSTM-CRF 模型的输入层中加入丰富的语料特征确实减轻了可用标注数据的不足。例如，训练集比例为 50%时，Bi-LSTM-CRF 模型的输入层中仅使用字符嵌入特征情况下的 F1 值为 76.46%，而训练集比例为 25%时，在 Bi-LSTM-CRF 模型的输入层中加入位置特征（Seg）情况下的 F1 值为 73.67%。训练集比例为 75%时，Bi-LSTM-CRF 模型的输入层中仅使用字符嵌入特征情况下的 F1 值为 78.07%，而训练集比例为 50%时，在 Bi-LSTM-CRF 模型的输入层中加入位置特征（Seg）和词性特征（Pos）情况下的 F1 值为 76.90%。训练集比例为 100%时，Bi-LSTM-CRF 模型的输入层中仅使用字符嵌入特征情况下的 F1 值为 78.92%，而训练集比例仅为 25%时，在 Bi-LSTM-CRF 模型的输入层中加入技能词的上下文特征（Con）情况下的 F1 值便可达到 81.72%。因此，我们可以得出结论，加入丰富的语料特征本研究模型能够缓解模型对大量标注数据的依赖。

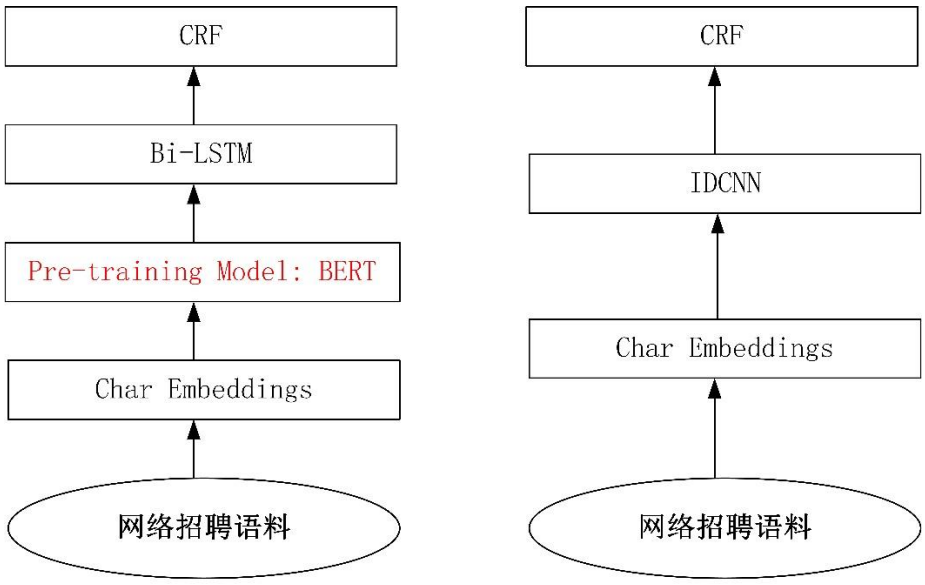
实验三：为了说明本研究提出的技能词抽取模型的有效性，我们选取了目前主流的序列标注模型 BERT-Bi-LSTM-CRF 和 IDCNN-CRF 模型进行对比。

方法一，BERT-Bi-LSTM-CRF：BERT (Bidirectional Encoder Representations from Transformers)是由 Devlin[71]等人提出的，一种以 Transformers 为主要框架的双向编码表征模型。BERT-Bi-LSTM-CRF 是在预训练的 BERT 模型的顶部添加了用于序列标

中国知网 https://www.cnki.net

记的 Bi-LSTM 层和 CRF 层，并且通过招聘语料数据对预训练的 BERT 模型的参数进行了调整。

方法二，IDCNN-CRF：IDCNN-CRF(Iterated Dilated Convolutional Neural Network-Conditional Random Field)模型是由 Emma[72]等人提出的，类似于 Bi-LSTM-CRF 模型，采用深度学习模型进行特征提取，再放入 CRF 层解码出标注结果。但不同于 Bi-LSTM 网络，该模型使用的是四个结构相同的膨胀卷积（DCNN）来提取语句特征，称之为 IDCNN。



(a)：BERT-Bi-LSTM-CRF模型　　(b)：IDCNN-CRF模型

图 4-3 BERT-Bi-LSTM-CRF 和 IDCNN-CRF 网络模型

数据集的划分同样采用交叉验证的实验方案，也同样选择 Bi-LSTM-CRF 模型作为基线对比模型，实验结果如表 4-6 所示。方法 1 中 BERT-Bi-LSTM-CRF 模型的超参数设置如下，初始学习率为$1e^{-4}$、epoch 次数为 50、隐藏层维度为 100 和批量训练样本量（batch）为 32。方法 2 中 IDCNN-CRF 模型的超参数设置如下，字符嵌入维度为 100、初始学习率为 0.001、epoch 次数为 100、卷积核大小为 1*3 和批量训练样本量（batch）为 20。

表 4-6 不同网络模型抽取性能对比

模型	Precision	Recall	F1	提升
Bi-LSTM-CRF	0.7890	0.7899	0.7892	--
BERT-Bi-LSTM-CRF	0.8156	0.8184	0.8170	+2.78%
IDCNN-CRF	0.8048	0.7967	0.8009	+1.17%
技能词抽取模型	**0.8706**	**0.8727**	**0.8715**	+8.23%

从表 6 中可以看出，本研究的技能词抽取模型的 F1 值远优于对比方法 1：BERT-Bi-LSTM-CRF 模型和方法 2：IDCNN-CRF 模型的 F1 值。再考虑到 BERT 模型使用

中国知网 https://www.cnki.net

的训练语料库的大小较大，而本研究的技能词抽取模型使用的训练语料库较小。因此，我们可以得出结论，本研究的技能词抽取模型可以有效的抽取技能词，并且加入的语料特征更有利于模型的抽取性能。

§4.4 本章小结

在本章节中，首先详细地描述了所研究的问题情形与挑战。在第二节中介绍基于深度学习的技能词抽取模型。在第三节中介绍了数据划分和实验参数设置，并且对实验结果进行描述与分析。在这项工作中，提出了一种基于深度学习的技能词抽取算法，将技能词抽取任务转换为序列标注问题，并在模型中引入了更多的语料特征并将输入层的输出与 Bi-LSTM 输出连接在一起作为 CRF 层的输入。评估了加入丰富的语料特征，本研究提出的模型是否能够缓解模型对大量标注数据的依赖。进行了大量实验，实验结果表明，本研究提出的技能词抽取模型具有最佳的抽取性能。该研究已撰写成论文《基于深度学习的中文网络招聘文本中的技能词抽取》，且已被桂林电子科技大学学报期刊录用。

中国知网 https://www.cnki.net

第五章 基于跨领域迁移学习的技能词抽取算法

§5.1 问题描述及的挑战与分析

上一章节中，通过对网络招聘文本中的技能词分析，看出技能词抽取与命名实体识别和术语抽取问题相似，都可以转换为序列标注任务。当前，虽然基于深度神经网络的序列标注模型可以通过端到端的训练方式以捕获上下文信息，将输入字符转换为输出标签，而无需显式的手工设计进行特征提取。但是，这类方法仅专注于域内监督学习，这需要大量带标注的训练数据。对于网络招聘数据而言，由于人工标注既费时又昂贵，只能依靠专家手工标注少数句子。因此，很难获得足够多的带标注数据来训练深度神经网络。然而，更好的方法应该利用领域外有标注的数据，通过迁移学习方法改善域内技能词抽取性能。

现有的关于序列标注模型的深度迁移学习方法主要包括：特征表示迁移和参数迁移。特征表示迁移方法通常利用深度神经网络来学习源域和目标域之间的紧密特征映射，而参数迁移方法通常利用参数共享或联合训练来使目标域模型参数接近源域模型的参数[64]。但是，这些方法需要源域和目标域具有相同的标签集或相同的标签含义，即强的域相似性。本研究选择了中国语言处理特别兴趣小组（SIGHAN）的中文命名实体识别语料库[4]作为源域。 SIGHAN 是关于新闻领域的，它包含三种类型的实体(人名、地名和机构名)，并具有七个标签（B-Person，I-Person，B-Location，I-Location，B-Organization，I-Organization 和 Other）。但是，在本研究中，网络招聘数据仅标注三个标签（B-SKILL，I-SKILL，Other），并且标签集的含义也与 SIGHAN 中的不同。因此，如何将从 SIGHAN 获得的知识迁移到目标领域也是一个挑战。

为了应对上述挑战，本研究提出了一种跨领域迁移学习的技能实体抽取方法(Cross-Domain Transfer Learning Recognize Professional Skill Entity, CDTL-PSE)，该方法从带标注的域外数据和带标注招聘数据中学习知识来提升技能词抽取性能。据我们所知，本研究是首次尝试从现实世界的中文招聘数据中自动提取技能词。这项研究的主要目标是：（1）提供用于抽取技能词的(CDTL-PSE)的跨域迁移学习方法。所提出的方法首先将 SIGHAN 语料库分解为三个源域，并利用域自适应层来帮助目标任务在每个源域上迁移学习。然后，使用参数迁移方法来训练每个子模型。最后，通过多数表决获得序列标签预测。（2）提供 CDTL-PSE 的合理性，即将 SIGHAN 语料库分解为三个源域，并且集成学习可以提高标签序列预测的准确性，域自适应层对于跨域迁移学习至关重要。（3）评估 CDTL-PSE 是否可以缓解可用标注数据的稀缺性并可以有效识别技能词。

[4] http://www.aclweb.org/mirror/ijcnlp08/sinhan6/chinesebakeo.htm

中国知网 https://www.cnki.net

§5.2 跨领域迁移学习的技能词识别(CDTL-PSE)算法

在本节中，我们将介绍所提出的 CDTL-PSE 方法的细节，并深入讨论其合理性。CDTL-PSE 方法使用 2.3 小节中介绍的 Bi-LSTM-CRF 模型作为基础模型。

§5.2.1 CDTL-PSE 框架概述

图 5-1 展示了基础模型 Bi-LSTM-CRF 的模型结构，它主要由嵌入层（Embedding layer），Bi-LSTM 层和 CRF 层组成。在嵌入层中，将每个输入语句转换为一系列嵌入向量，然后将其输入到 Bi-LSTM 层中，以将上下文信息顺序编码为固定长度的隐藏向量，并最终由 CRF 层预测最佳标签序列。

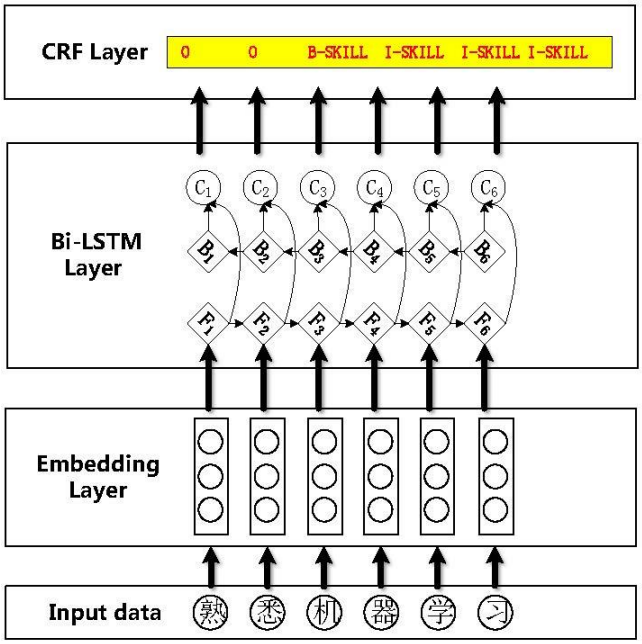


图 5-1 Bi-LSTM 模型图.

第一层是嵌入层，在这一层中，由一个嵌入矩阵 $M \in R^{D \times V}$ 用于将每个字符映射到一个密集矢量中，其中 D 是嵌入向量维度，而 V 是字符量。输入句子表示为 $S=[w_1,w_2,...,w_n]$，其中 n 是句子的长度，$w_i \in R^V$ 是第 i 个字符的one-hot表示。句子的字符嵌入向量表示为，其中 $c_i = M_{w_i} \in R^D$。

第二层是 Bi-LSTM 网络，将 $\overrightarrow{LSTM}$ 表示为 LSTM 网络从左到右扫描输入文本，则 $\overrightarrow{LSTM}$ 学会的隐藏表示可以表示为 $[\overrightarrow{h_1},\overrightarrow{h_2},...,\overrightarrow{h_n}]$。类似地，$\overleftarrow{LSTM}$ 表示 LSTM 网络从右到左扫描序列，$[\overleftarrow{h_1},\overleftarrow{h_2},...,\overleftarrow{h_n}]$ 表示由 $\overleftarrow{LSTM}$ 学习到的隐藏表示。第 i 个字符隐藏表示是通过将拼接其 $\overrightarrow{LSTM}$ 和 $\overleftarrow{LSTM}$ 上下文表示得到 $h_i=[\overrightarrow{h_i},\overleftarrow{h_i}]$。最终，Bi-LSTM 层的输出为 $h=[h_1,h_2,...,h_n]$，其中 $h_i \in R^{2K}$，K 是 LSTM 网络中隐藏状态的维数。

最后一层是CRF层。将 $y=[y_1,y_2,...,y_n]$ 表示为句子 S 的标签序列，其中 $y_i \in R^L$ 是第 i 个字符标签的one-hot表示。CRF层的输入是Bi-LSTM层的输出，CRF层的输出是标签序列 y。给定输入 h 的标签序列 y 的条件概率计算如下[20]：

中国知网 https://www.cnki.net

$$p(y|h,\theta)=\frac{\prod_{i=1}^{n}\Psi(h_i,y_i,y_{i-1})}{\sum_{y'\in Y(S)}\prod_{i=1}^{n}\Psi(h_i,y'_i,y'_{i-1})} \tag{5-1}$$

其中，$Y(S)$是句子S的所有可能标签序列的集合，θ是参数集，而$\Psi(h_i,y_i,y_{i-1})$是势函数，如下所示：

$$\Psi(h_i,y_i,y_{i-1})=\mathrm{e}^{\left(y_i^T E^T h_i+y_{i-1}^T F y_i\right)} \tag{5-2}$$

其中，$E\in R^{2K\times L}$，$F\in R^{L\times L}$是CRF层的参数，即$\theta=\{E,F\}$。该方法的损失函数为：

$$Loss=-\sum_{S\in corpus}\log\left(p(y_S|h_S,\theta)\right) \tag{5-3}$$

其中*corpus*是训练数据集中的所有语句。

图 5-2 概述了 CDTL-PSE 框架，其中包括四层，即：嵌入层（Char Emb）、Bi-LSTM 层、域自适应层（Domain Adaptation）和 CRF 层。 在本研究中，SIGHAN 语料库根据其不同类型的实体识别任务分解为三个源域，并且这三个任务的标签映射到相同的标签集（B-Entity，I-Entity，Other）。目标任务是从网络招聘文本中抽取技能词。嵌入层（Char Emb）和 Bi-LSTM 层在源域和目标域之间共享，用于将学习到的知识从每个源域迁移到目标域，但不同的任务构建了特定于任务的域自适应层（Domain Adaptation）和 CRF 层。因此，CDTL-PSE 的框架由三个子网组成。

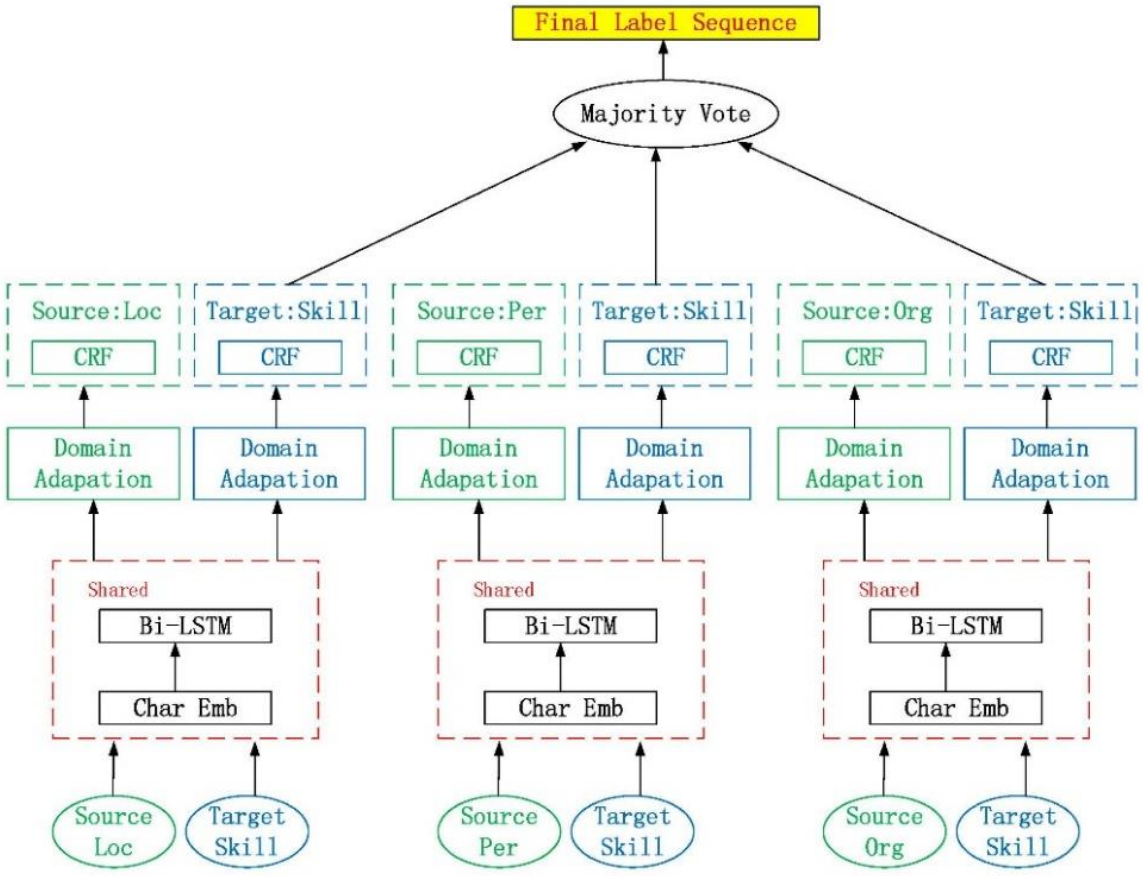


图 5-2 CDTL-PSE 框架图.

中国知网 https://www.cnki.net

在图 5-2 中，对于目标任务，从下到上，每个输入语句都被转换为一系列嵌入向量，馈送到 Bi-LSTM 层中以将上下文信息顺序编码为固定长度的隐藏向量。 然后，将由共享 Bi-LSTM 层生成的字符的隐藏表示形式输入到目标领域特定的域自适应层中，以学习领域特定的知识。域自适应层的输出直接馈送到目标域特定的 CRF 层，以预测标记序列。最后，通过对这三个子网的结果进行多数表决获得标签序列预测。

与将在第 5.3.3 节中进行比较的方法 B，C 和 D 相比，CDTL-PSE 采用与方法 B 或 C 相同的迁移学习策略，并且与方法 D 不同之处在于使用不同的域自适应层。CDTL-PSE 的一个特点是，SIGHAN 语料库根据其不同的实体识别任务分解为三个源域，并使用集成学习策略来获得最终的标签序列。据我们所知，本研究提出的 CDTL-PSE 方法是同时使用 SIGHAN 语料库中的三种实体进行迁移学习的首次尝试。

§5.2.2 源域分解

对于中文命名实体识别，语料资源的稀缺是不言而喻的，并且很难在现有语料资源中找到与招聘数据相关或相似的，且标签丰富的源域。经过仔细选择，我们将标准 SIGHAN Bakeoff 2006 新闻语料库作为源域。在 Yang[68]等人的工作中，指出参数迁移要求源任务和目标任务之间的相似性很强。如果源任务和目标任务之间的潜在相似性不太突出，则可迁移性将降低，并且通过迁移学习获得的改进将很弱。通过分析招聘语料库，我们发现技能词在句子中的位置主要位于动词，形容词/形容词短语，以及“和”，“或”，“、”等之后。例如：“了解自然语言处理”，“常用的机器学习算法”，“掌握文本挖掘、实体抽取、词性标注等技术”。

为了分析 SIGHAN 与招聘语料之间的相似性，我们从 SIGHAN 语料库中随机抽取了 2000 个句子。经过统计，我们发现在这些随机选择的语句中有 880 个人名实体，963 个机构名实体和 1822 个地名实体。表 5-1 中展示了 SIGHAN 抽样语料和招聘语料中不同位置的实体的出现频次的统计，从中可以看出，对于出现在动词后或形容词/形容词短语之后位置的实体中，地名实体和机构名实体出现的情形比人名实体出现的情况与技能词更相似，而在“和”，“或”，“、”等之后出现的频次，人名和机构名出现的情形比地名实体出现的情况更与技能词相似。因此，我们可以得出结论，在人名，地名和机构名这三种情况之间存在差异。

表 5-1 相应语料库中不同位置的实体的出现情况统计

实体类型	实体数目	动词后	形容词/形容词短语后	名词/名词短语后	“和”、“或”、“、”等并列形式	其他位置
技能词	3930	1309	764	2	1512	343
人名	884	136	27	204	73	444
地名	1951	649	106	57	87	1052
机构名	984	278	54	36	128	488

因此，如果将整个 SIGHAN 语料库直接用作源域，则通常会在同一句子中出现

中国知网 https://www.cnki.net

多种类型的实体，这会削弱源域和目标域之间的相似性。通过前面的分析，我们可以推测，如果将 SIGHAN 分解为三个源域，则每个源域都将与目标域具有一定的相似性。因此，SIGHAN 语料库被分解为三个源域，并且每个源域都使用相同的实体标签。基于 Bi-LSTM-CRF 模型，我们并行利用不同的源域通过网络参数共享迁移到目标域。

§5.2.3 域自适应层

由于源域和目标域具有不同的语言样式，并且包含大量特定于域的术语，这些术语在各个域之间没有相同的语义，因此仅分解 SIGHAN 语料库是不够的。例如，“机器学习”是目标域中的技能词。字符“机”应标记为“B-SKILL”。但是，“机器学习”不是这三个源域中的实体，并且将被标记为“Other”。此外，如果直接采用网络参数迁移，则由于源域的数据量远大于目标域，很难学习目标域中的特定领域知识。因此，我们需要引入一个域自适应层，以从目标域中重新学习上下文信息。它将基于共享的特征表示中学习目标域的领域知识。在本研究中，我们介绍了两种用于域自适应的方法：增加额外的 Bi-LSTM 层和特征扩充。域自适应层将先前共享的 Bi-LSTM 层中输出的隐藏状态作为其输入，并从共享网络中捕获的上下文信息中的输出进行更改，为最终 CRF 层生成一系列新的隐藏状态。

（1） 增加额外的Bi-LSTM层

在最终 CRF 层之前，我们在域自适应层中使用额外的 Bi-LSTM 网络来重新学习目标域中的特定域知识。域自适应层中使用的 Bi-LSTM 网络与基础模型中 Bi-LSTM 层的网络结构完全相同。Lin 等人[58]使用 Bi-LSTM 网络作为输出适应层来捕获上下文信息的输出变化，从而解决了在目标域中使用上下文信息执行重新分类和重新识别的问题。

（2） 特征扩充方法

我们还将在最终 CRF 层之前的域自适应层中使用特征扩充方法，使模型既可以从源域中学习，又能保留目标域特定的知识。特征扩充方法由 Daumé III[69]提出，主要思想是复制和扩充原始特征。扩充后特征包括域共同特征和域特定特征。共同的特征是从共享的 Bi-LSTM 网络中学到的特征表示。特定于域的特征是源/目标域使用它们自己独立的 Bi-LSTM 网络学习到的隐藏表示。

如果原始的输入空间是 $X \in R^n$，特征扩充后的输入空间则为 $N \in R^{3n}$。$\Psi_s: X \to N$，$\Psi_t: X \to N$ 分别表示源域和目标域的原始输入空间到特征扩充后的输入空间的映射函数，它们的定义分别为 $\Psi_s(X) = (X, X, 0)$，$\Psi_t(X) = (X, 0, X)$。其中，$0 = (0,0,...,0)$ 是 0 向量。映射后向量的第一项代表源域和目标域的共同特征，第二项代表源域的特定特征，第三项代表目标域的特定特征。

中国知网 https://www.cnki.net

§5.2.4 多数投票

由于不同源任务之间存在差异的原因，具有不同源任务的实体识别可能会对目标任务的抽取性能产生不同的影响。在使用三个不同的源域迁移到目标域以获取三组不同的标记序列后，我们使用多数表决方法从这三组标记序列中获取最终的标记序列。

§5.2.5 训练过程

CDTL-PSE的框架包括三个用于迁移学习的子网，三个不同的源域被并行迁移到目标域。CDTL-PSE框架中的每个子网都通过小批量自适应矩估计（Adam）优化算法[70]以端对端的方式交替训练或微调。在每个源域到目标域的迁移学习过程中，嵌入层和Bi-LSTM网络层是共享的，而特定域自适应层和CRF层则用于每个特定任务。每个子网都经过100个周期的训练。

当使用微调策略时，我们首先在每个子网上训练源任务的参数。然后，我们迁移其Bi-LSTM网络层的参数，以在相应的目标任务模型中初始化Bi-LSTM网络层，并通过目标任务的训练数据对相应的目标任务模型进行重新训练。使用交替训练策略时，在每个周期，我们首先使用相应的源域数据来训练源任务模型，然后使用目标域数据来训练目标任务模型，如此进行循环交替。

在每个周期中，源域和目标域的训练数据都被分为若干批次进行批量训练，每批次的大小不超过20个句子。对于每批次，我们采用预训练的字符嵌入，而不是随机初始化的嵌入作为输入字符特征。为了避免过度拟合，我们在将字符特征表示输入到Bi-LSTM层之前使用以概率为0.5的dropout训练。在CRF层中，我们使用动态编程来计算（5-1）公式中的规范化并预测标签序列。

§5.3 实验设置及结果分析

为了评估提出的CDTL-PSE方法的抽取性能，我们将CDTL-PSE与其他主流方法进行了比较。我们将在实验中描述数据集，标注原理，预训练的嵌入，基线模型，参数设置和结果的详细信息。

§5.3.1 实验数据设置

（1）目标域数据集

在此实验中，我们同样选择通过从智联招聘网站中获取的IT行业类别的职位招聘数据作为语料库。将每条职位招聘数据中的“岗位职责要求”视为一条语句。由于时间和手工标注成本的限制，我们仅标注了447条“岗位职责要求”作为实验语料。在标注语料时，我们尝试选择句法结构标准的句子作为研究语料，并将语句中的专业课程名称，专业知识点和相关专业工具标记为技能词，例如“C语言程序设计”，“堆排序”，“SpringMVC框架”等。因此，一条语句将被标记为：“熟/ O悉/ O机/ B-SKILL器/ I-SKILL

中国知网 https://www.cnki.net

学/ I-SKILL习/ I-SKILL与/ O自/ B-SKILL然/ I-SKILL语/ I-SKILL言/ I-SKILL处/ I-SKILL理/ I-SKILL。/O”。

表 5-2 目标域数据集的统计信息

数据集	数据类型	实体类型	技能词数量	语句数量
目标域	训练集	技能词	[2677, 2886]	321
	验证集	技能词	[333, 534]	36
	测试集	技能词	[652, 843]	90

由于没有明确的标准如何将数据集划分为训练/验证/测试集，因此我们在实验中将数据集进行两轮交叉验证。首先将整个数据集划分成5份，选择其中的80%作为训练数据，而将其余20%用作测试数据。然后，再次将训练数据划分成十份，选择其中的90%作为最终训练数据，并将其余的10%用作验证数据。换句话说，将全部数据的72％作为训练数据，将8％作为验证数据，将20％作为测试数据。每次实验的训练周期为100个周期，通过验证集找出训练集在这100个周期内最佳网络模型参数后再使用测试集进行测试，以获得本次验证的结果。表5-2中分别展示了用于训练，验证和测试数据集的句子数，实体数量和实体类型。值得注意的是，由于每个句子中包含的实体数量不同，训练/验证/测试数据集中的实体数量会随着每次划分而动态变化，因此，实体数量的范围也在表中。

（2）源域数据集

本研究以SIGHAN语料库作为源域数据。它的训练数据集中包含23,182个句子，测试数据集中包含4,636 465个句子。我们从其训练集中随机选择10％，即2318句作为验证数据集。表5-3详细列出了源域中数据集的统计信息。

表 5-3 源域数据集的统计信息

数据集	数据类型	实体类型	实体数量	语句数量
源域	训练集	人名、地名、机构名	8144,16571,9277	20864
	验证集	人名、地名、机构名	884,1951,984	2318
	测试集	人名、地名、机构名	1864,3658,2185	4636

实验采用机器学习中常用的准确率（Precision）、召回率（Recall）指标评价技能词抽取模型的性能，并采用F1值指标评价其综合性能。

§5.3.2 实验参数设置

（1）预训练字符向量

我们使用预训练的字符嵌入，而不是随机初始化。我们在word2vec中使用CBOW模型对有大量未标记的语料的20151201中文维基百科预训练[67]了维度为100的字符嵌入，源域和目标域任务都使用相同的嵌入。

（2）超参数设置和初始化

本研究使用的超参数包括初始学习率，训练时期，批量训练样本大小，输入嵌入的dropout率以及LSTM隐藏层向量的维数。这些超参数的范围是根据文献中的先前工

中国知网 https://www.cnki.net

作确定的，例如初始学习率选择为： 0.005、0.001或0.002，训练周期为100，批量训练样本大小为：20或50，用于输入嵌入的dropout率为：0.2或0.5。最终，直接将训练周期，输入嵌入的dropout率和LSTM隐藏层向量的尺寸分别设置为100、0.5和100，而通过交叉将初始学习率和批量训练样本大小分别设置为0.001和20（最常见的最佳参数）。其他网络参数包括网络中每个层的初始权重，并且CRF层的状态转换矩阵在[-1，1]范围内随机地均匀初始化。

§5.3.3 对比策略

为了说明本研究所提出的CDTL-PSE方法的合理性和有效性。本研究使用以下几种主流方法进行比较。

方法 A：Non-transfer（不使用迁移学习）

Huang[16]等人提出了多种基于长短期记忆（LSTM）的序列标注模型。本研究仅将Bi-LSTM-CRF模型应用于目标域数据集并用作基线方法。

方法 B：Joint-training（联合训练）

Yang[57]等人提出了一种基于深度递归网络的迁移学习方法，该方法在源任务和目标任务之间共享隐藏的特征表示和部分模型参数。在研究中，仅使用具有不同标签集的跨域迁移模型。在此模型中，每个任务学习一个单独的CRF层，而每个任务共享隐藏的特征表示。我们将其训练过程简化为交替训练。

方法 C：Fine tuning（微调）

Lee[59]等人提出在源数据集上训练模型的参数并迁移所有或部分参数以初始化模型，以便在目标数据集上训练。类似地，在本研究中，仅迁移Bi-LSTM层的参数 。

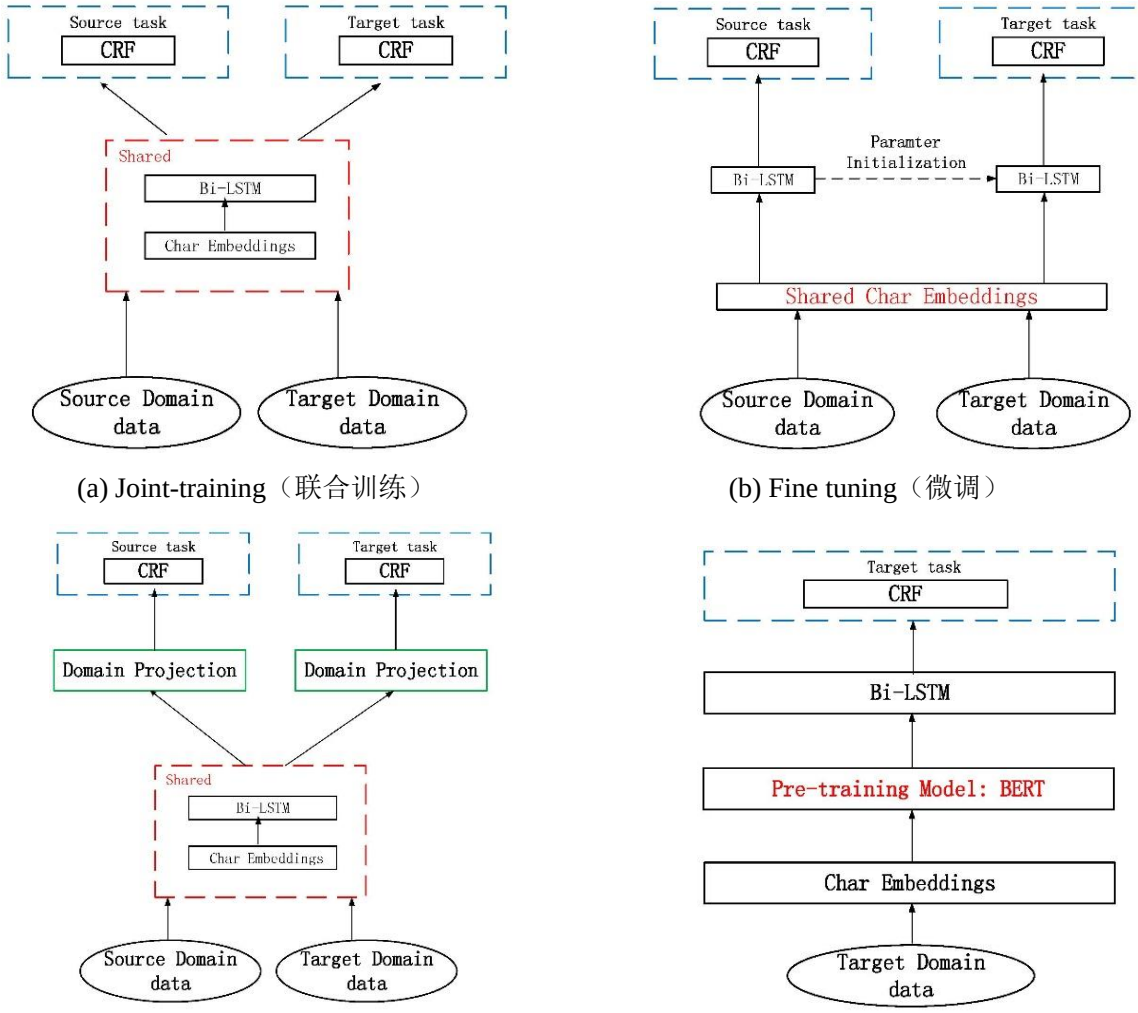
方法 D：Domain Projection（域投影）

该方法由Peng[60]等人提出。它也是一种基于Bi-LSTM-CRF模型的迁移学习方法，旨在通过增加域投影层来产生不同域之间的共享特征表示。在本研究中，我们仅在多领域多任务情况下使用该方法，其中领域投影层和CRF层是任务特定的，而Bi-LSTM层是共享的，并且源任务和目标任务的模型是交替训练。

方法 E：BERT-Bi-LSTM-CRF

BERT (Bidirectional Encoder Representations from Transformers)是由Devlin[71]等人提出的，一种以Transformers为主要框架的双向编码表征模型。BERT-Bi-LSTM-CRF是在预训练的BERT模型的顶部添加了用于序列标记的Bi-LSTM层和CRF层，并且通过招聘语料数据对预训练的BERT模型的参数进行了调整。

中国知网 https://www.cnki.net



(a) Joint-training（联合训练）

(b) Fine tuning（微调）

(c) Domain Projection（域投影）

(d) BERT-Bi-LSTM-CRF

图 5-3 不同对比方法网络结构图

在本研究中，基于上述方法，考虑了不同情况下的识别性能比较。

(i)：在将SIGHAN语料库分解为三个不同的源域之后，每个源域的标签集都映射到一个统一的标签集（B-Entity，I-Entity，Other）。使用方法B的迁移策略将三个源域一个接一个地迁移到目标域（分别标记为Per_Alter、Loc_Alter和Org_Alter）。

(ii)：带有原始标签集的整个SIGHAN语料库都用作源域。使用方法B的迁移策略将源域迁移到目标域（标记为Diff_Alter）。

(iii)：整个SIGHAN语料库都用作源域，但是其标签集被映射到一个统一的标签集（B-Entity，I-Entity，Other）。使用方法B的迁移策略将源域迁移到目标域（标记为Same_Alter）。

(iv)：标签序列预测是通过对Per_Alter、Loc_Alter和Org_Alter的结果进行多数表决而获得的（标记为Alter_Ensemble）。

(v)：与(i)的情况类似，但是将迁移学习方法替换为方法C（分别标记为Per_Tune、Loc_Tune和Org_Tune）。

中国知网 https://www.cnki.net

(vi)：与(ii)的情况类似，但是将迁移学习方法替换为方法C（标记为Diff_Tune）。

(vii)：与(iii)的情况类似，但是将迁移学习方法替换为方法C（标记为Same_Tune）。

(viii)：标签序列预测是通过对Per_Tune、Loc_Tune和Org_Tune的结果进行多数表决而获得的（标记为Tune_Ensemble）。

(ix)：与(i)的情况类似，但是将迁移学习方法替换为方法D（分别标记为Per_Domain、Loc_Domain 和Org_Domain）。

(ix)：与(ii)的情况类似，但是将迁移学习方法替换为方法D（标记为Diff_Domain）。

(x)：与(iii)的情况类似，但是将迁移学习方法替换为方法D（标记为Same_Domain）。

(xi)：标签序列预测是通过对Per_Domain、Loc_Domain 和Org_Domain的结果进行多数表决而获得的（标记为Domain_Ensemble）。

(xii)：使用交替训练的CDTL-PSE，其中使用特征扩充或Bi-LSTM层进行域自适应（标记为CDTL-PSE：Alter（特征扩充）或CDTL-PSE：Alter（Bi-LSTM）。

(xiii)：使用微调的CDTL-PSE，其中使用特征扩充或Bi-LSTM层进行域自适应（标记为CDTL-PSE：Tune（特征扩充）或CDTL-PSE：Tune（Bi-LSTM）。

(xiv)：方法E用于目标域（标记为BERT-Bi-LSTM-CRF）。

本研究将以上实验结果分成四个部分，以说明为什么SIGHAN语料库应分解为三个源域，集成学习可以提高标签序列预测的准确性，以及域自适应层对于跨域迁移学习至关重要。此外，本研究以不同比例的标记率来运行所提出的CDTL-PSE方法来评估该方法是否可以缓解可用标注数据的不足，以及是否可以有效地识别技能词。

§5.3.4 实验结果与分析

（1）实验一

为了说明为什么将SIGHAN语料库划分为三个源域，我们比较了(i)-(xii)情况和基线模型情况下的识别性能。从表5-4的结果可以看出，在不同的迁移技术（方法B，方法C和方法D）下，1)在将SIGHAN语料分解为三个不同的源域的情况下的F1值比基准模型更好或略差。最有趣的是，在仅包含地名类型实体的SIGHAN语料库的情况下，F1值始终比基准模型的情况下的F1值差。2)整个带有原始标签集的SIGHAN语料库的F1值小于将SIGHAN语料库分解为三个不同的来源域的F1值。3)与带有原始标签集的整个SIGHAN语料库的情况相比，仅将整个SIGHAN语料库的标签集映射到一个统一的标签集（B-Entity，I-Entity，Other）不会带来识别精度的提升。4)在将SIGHAN语料分解为三个不同的源域的情况下，将结果集成在一起将显着提高识别性能。

根据这些观察结果，可以进一步发现：1)具有不同实体类型的三个不同源域之间存在差异；2)每个源域仅包含一种类型的实体，与整个SIGHAN语料库相比，与目标

中国知网 https://www.cnki.net

域更相似。因此，应将SIGHAN语料库分解为三个不同的源域，将SIGHAN分解情况下的结果进行集成将有效地提高识别性能。

表 5-4 将不同源任务迁移到目标任务的结果（%），括号中为标准差

方法	Precision	Recall	F1	提升
No-Transfer	80.04 (3.12)	82.11 (2.63)	81.03 (2.41)	--
Per_Alter	79.78 (3.43)	82.59 (2.64)	81.15 (2.93)	+0.12
Loc_Alter	79.76 (2.84)	82.00 (2.76)	80.84 (2.50)	-0.19
Org_Alter	80.07 (3.33)	82.28 (2.26)	81.13 (2.50)	+0.10
Diff_Alter	79.81 (3.22)	81.42 (2.58)	80.58 (2.57)	-0.45
Same_Alter	79.26 (3.54)	81.72 (3.29)	80.44 (3.04)	-0.59
Alter_Ensemble	**80.67** (2.77)	**82.38** (2.70)	**81.51** (2.61)	**+0.48**
Per_Tune	80.13 (2.37)	82.09 (2.42)	81.09 (2.18)	+0.06
Loc_Tune	79.85 (2.30)	81.58 (3.02)	80.68 (2.19)	-0.35
Org_Tune	80.00 (2.42)	82.02 (2.30)	80.96 (2.00)	-0.07
Diff_Tune	79.27 (3.43)	81.79 (1.77)	80.49 (2.44)	-0.54
Same_Tune	79.38 (2.24)	81.73 (2.12)	80.52 (1.96)	-0.51
Tune_Ensemble	**80.37** (2.23)	**82.13** (2.55)	**81.22** (1.97)	**+0.19**
Per_Domain	80.30 (2.43)	82.35 (2.57)	81.29 (2.07)	+0.26
Loc_Domain	79.71 (2.88)	82.35 (2.33)	80.99 (2.26)	-0.04
Org_Domain	80.25 (2.73)	82.26 (2.06)	81.22 (2.14)	+0.19
Diff_Domain	79.71 (4.17)	81.76 (3.31)	80.69 (3.46)	-0.34
Same_Domain	79.61 (3.04)	81.88 (2.36)	80.70 (2.32)	-0.33
Domain_Ensemble	**80.82** (3.34)	**82.41** (3.19)	**81.58** (2.96)	**+0.55**

为了进一步说明源域分解和集成学习的必要性，本研究进行了95%置信度水平的配对t检验，以检查比较方法之间的差异（就平均F1值而言）是否在统计学上显着。表5-5中的结果表明，将整个SIGHAN语料库直接使用不适合用作目标任务的源域，而表5-6中的结果表明，集成学习方法确实可以提高识别性能。

表 5-5 目标数据集上使用不同迁移学习方法的 F1 值

迁移学习方法	DiffLabel	SameLabel	Per	Loc	Org
方法 B	80.58●■★	80.44●■★	81.15	80.84	81.13
方法 C	80.49●■★	80.52●■★	81.09	80.68	80.96
方法 D	80.69●■★	80.70●■★	81.29	80.99	81.22

注：DiffLabel，SameLabel表示两种情况，其中直接使用原始源域而不更改标签集和更改标签集。Per，Loc和Org表示通过分解原始源域而获得的三个新源域。●，■和★分别表示使用Per，Loc和Org的源域比直接使用相应的源域要好得多。

中国知网 https://www.cnki.net

表 5-6 目标数据集上使用不同迁移学习方法与集成学习的 F1 值

迁移学习方法	Per	Loc	Org	Ensemble
方法 B	81.15●	80.84●	81.13●	81.51
方法 C	81.09●	80.68●	80.96●	81.22
方法 D	81.29●	80.99●	81.22●	81.58

注：●表示使用集成学习比仅使用单一类型的源域要好得多。

（2）实验二

为了说明域自适应层对于跨域迁移至关重要，我们将（iv）、（viii）、（xii）、（xiii）和（xiv）情况下的结果与基线方法进行了比较。（iv）和（xiii）情况下的结果可用于分析使用迁移学习方法B，域自适应层是否必不可少，而（viii）和（xiv）情况下的结果可用于分析使用迁移学习方法C，域自适应层是否必不可少。在（xii）的情况下，所使用的方法与本研究所提出的CDTL-PSE方法非常相似，仅域自适应层有所不同，因此，结果(xii)、(xiii)和(xiv)的情况可以用来分析哪种域自适应方法最适合目标任务。

表 5-7 不同迁移学习方法的结果（%），括号中为标准差

方法	Precision	Recall	F1	提升
No-Transfer	80.04 (3.12)	82.11 (2.63)	81.03 (2.41)	--
Alter_Ensemble	80.67 (2.77)	82.38 (2.70)	81.51 (2.61)	+0.48
CDTL-PSE：Alter（特征扩充）	80.71 (2.73)	82.51 (2.45)	81.59 (2.41)	+0.56
CDTL-PS：Alter（Bi-LSTM）	**81.39** (3.61)	**82.58** (2.29)	**81.94** (2.49)	**+0.91**
Tune_Ensemble	80.37 (2.23)	82.13 (2.55)	81.22 (1.97)	+0.19
CDTL-PSE：Tune（特征扩充）	80.26 (2.42)	82.41 (2.64)	81.30 (2.09)	+0.27
CDTL-PSE：Tune（Bi-LSTM）	81.06 (2.53)	82.19 (1.88)	81.61 (1.93)	+0.58
Domain_Ensemble	80.82 (3.34)	82.41 (3.19)	81.58 (2.96)	+0.55
BERT-Bi-LSTM-CRF	80.03 (3.09)	83.84 (2.12)	81.86 (2.14)	+0.83

从表5-7的结果可以看出，在不同的迁移学习方法B和C下，1)两种类型的域自适应层的添加可以进一步提高集成学习情况下的识别精度。在(xiii)的情况下，与集成学习的结果相比，CDTL-PSE：Alter（特征扩充）和CDTL-PSE：Alter（Bi LSTM）的F1值分别提高了0.08%和0.43%。同样，在（xiv）的情况下，与集成学习的结果相比，CDTL-PSE：Tune（特征扩充）和CDTL-PSE：Tune（Bi-LSTM）的F1值提高了0.08%和0.39%。表5-8再次验证了这一点，表5-8中给出了配对t检验在置信度为95%时的结果；2)不同类型的域自适应层对识别精度的提升程度不同。根据这些结果，最好采用新的Bi-LSTM层进行域自适应。这是因为Bi-LSTM层比其他类型的域自适应方法可以极大地适应目标任务。

中国知网 https://www.cnki.net

表 5-8 同一数据集上不同方法的的 F1 值

迁移学习方法	Only Ensemble	特征扩充	Add Bi-LSTM
方法 B	81.51●■	81.59	81.94
方法 C	81.22●■	81.30	81.61

注：●表示通过域自适应层(特征扩充)明显优于仅使用集成学习，而■表示通过域自适应层(Bi-LSTM)显着优于仅使用集成学习。

从表5-7中可以看到，方法E的F1值与具有Bi-LSTM的域自适应层的CDTL-PSE模型的F1值相似，这非常有趣。考虑到BERT方法使用的训练语料库较大，而CDTL-PSE所使用的训练语料库较小。因此，我们可以得出结论，本研究的CDTL-PSE方法可以从域自适应策略中有效地受益。

（3）实验三

为了说明本研究所提出的CDTL-PSE方法可以缓解可用标注数据的稀缺性，我们以不同的标记率运行所提出的CDTL-PSE方法。本节中所有实验采用固定的超参数。我们对目标域数据集执行了5折交叉验证，其中80％用作训练数据，而20％用作测试数据。然后选择10％的训练数据集作为验证数据。从剩余的训练数据中，我们选择10％，25％，50％，75％和100％进行训练，并分别重复10、4、4、4和1次。结果的平均F1值如图5-4所示。

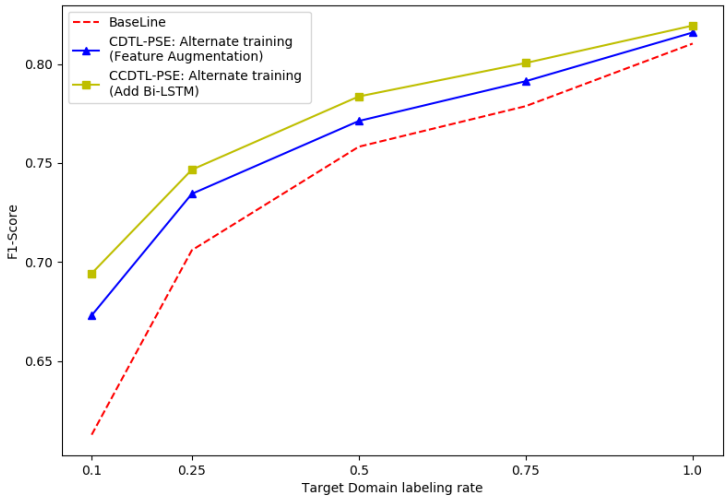


图 5-4 在不同规模的目标域训练数据下，提出的 CDTL-PSE 方法和基线方法的 F1 值

从图5-4的结果可以看出，1)在不同的标记率情况下，本研究提出的CDTL-PSE方法比基线方法一直得到提升，但是，随着标记率的提高，提升程度将降低。例如，标记率为10％时的提升为8.5％，而标记率为100％时的提升为0.91％。2)所提出的CDTL-PSE方法确实减轻了可用标注数据的不足。例如，标记率为25％的不迁移情况下的F1值为70.59％，而标记率为10％的CDTL-PSE情况下的F1值为69.42％。标记率为50％的

中国知网 https://www.cnki.net

不迁移情况下的F1值为75.82％，而标记率为25％的CDTL-PSE的F1值为74.66％。标记率为75％的不迁移情况下的F1值为77.87％，而CDTL-PSE标记率为50％的F评分为78.36％。

（4）实验四

为了验证提出的CDTL-PSE方法可以有效地抽取技能词，我们使用由带标注的IT行业语料库训练的CDTL-PSE模型从机械制造或服装设计行业的招聘文本中抽取技能词。分别使用机械制造行业的6,400个职位和服装设计行业的5,000个职位信息进行验证。 图5-5、图5-6、图5-7和图5-8展示了在每个领域中CDTL-PSE模型识别出的频次最多的前50个专业技能实体。

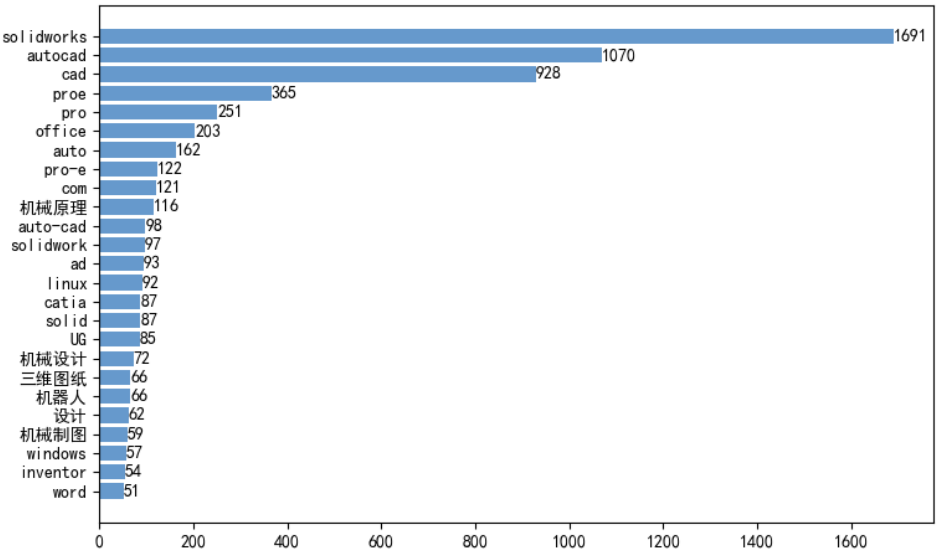


图 5-5 CDTL-PSE 识别出机械制造行业的前 25 个技能词

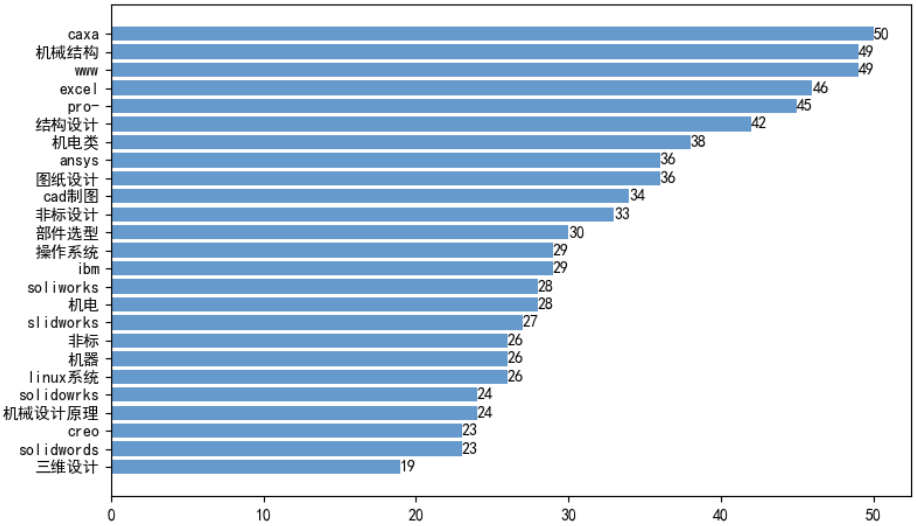


图 5-6 CDTL-PSE 识别出机械制造行业的前 50 个技能词

从机械制造行业的职位信息中识别出1,802个不同的技能词。我们邀请机器制造行业的专家来判断CDTL-PSE算法识别出的前50个技能词是否正确。根据专家判断，前50个技能词中有5个不能完全算作技能词：“com”、“设计”、“www”、“IBM”和“机

中国知网 https://www.cnki.net

器”。我们分析了为什么会将这五个词识别为技能词。由于“机器学习”是计算机行业中的技能词，而“机器”一词在机器制造业中也很常见，因此很容易被误认为是技能词。

其次，“设计”经常出现在“具有”、“熟悉”和“从事”的动词后面，例如“主要从事相关设计开发”、“具有设计能力”和“熟悉设计软件”，因此也容易被误认为是技能词。“IBM”是制造商的品牌名称，并且在职位要求中也经常出现该词，例如“学会使用IBM自主研发的产品”。此外，公司网站经常出现在岗位职责要求中，格式如下：“了解更多信息，请访问http://www.xxx.com/”。此格式类似于IT行业语料库中某些句子的结构，例如：“了解前端基础技术，如HTML / CSS / JavaScript等”。因此，“IBM”，“ www”和“ com”很容易被误认为是技能词。

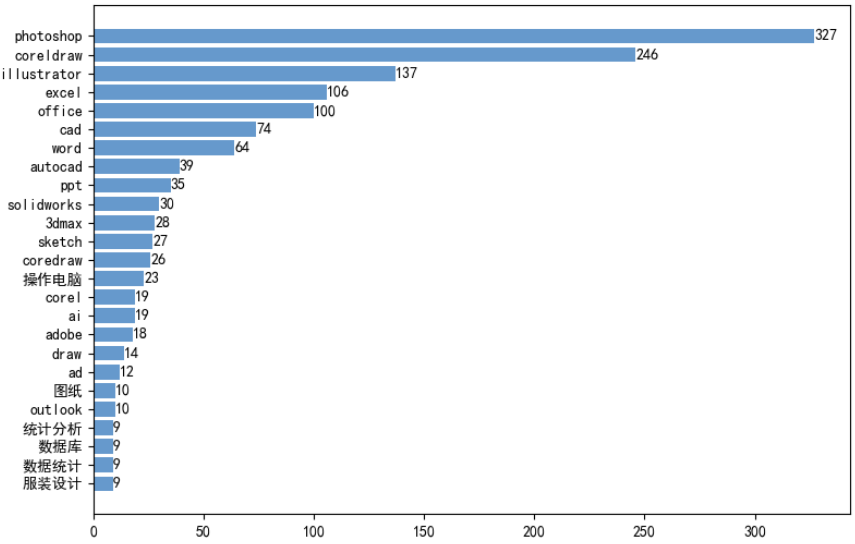


图 5-7 CDTL-PSE 识别出服装设计行业的前 25 个技能词

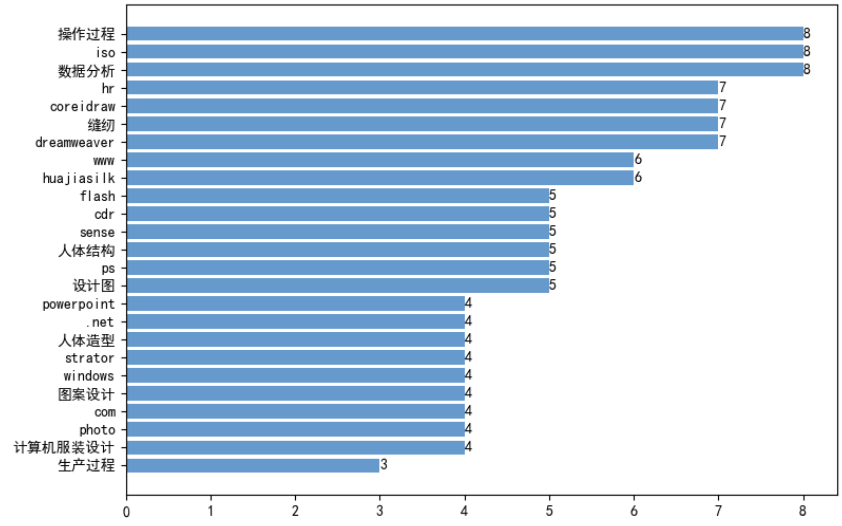


图 5-8 CDTL-PSE 识别出服装设计行业的前 50 个技能词

中国知网 https://www.cnki.net

在服装设计行业的5,000个职位中，共有603个不同的技能词被识别。我们也邀请了服装设计行业的专家来确定CDTL-PSE算法识别出的前50个技能词是否正确。根据专家判断，识别出的前50个技能词中有8个不能完全算作技能词：“hr”、“huajiasilk”、“www”、“sense”、“生产流程”、“操作过程”、“com”和“ISO”。 我们也分析了为什么将这八个词识别为技能词。

在服装设计行业的语料库中，有些句子与IT行业语料库中的句子相似，例如：“精通整个操作过程”、“热爱时装行业，具有审美 ，具有sense”、“具有以上条件可直接联系hr”、“熟悉ISO质量管理体系"，“通过访问www.huajiasilk.com了解更多” 和“熟悉服装的生产流程”等等。 因此，“hr”、“huajiasilk” 、“www”、“sense”、“生产流程”、“操作过程”、“com”和“ISO”容易被误认为是技能实体。

§5.4 本章小结

在本章中，首先简单地描述了本研究的问题情形与挑战，然后，对跨领域迁移学习的技能词抽取算法进行了详细地介绍，最后，阐述了实验所使用的数据集来源与交叉验证时数据集的划分以及实验参数与对比方法的设置，并且对实验结果进行描述与分析。在本研究中，我们提出了一种跨领域迁移学习的技能词抽取（CDTL-PSE）方法来识别网络招聘文本中的技能词。它首先将 SIGHAN 语料库分解为三个源域，并且利用在 Bi-LSTM 层和 CRF 层之间插入的域自适应层来帮助从每个源域学习到的知识迁移到目标域。然后，使用参数迁移方法来训练每个子模型。最后，通过多数表决获得最佳标签序列的预测。大量实验的结果说明了 CDTL-PSE 的合理性，CDTL-PSE 可以缓解手工标注数据的稀缺性。

本研究已撰写成英文稿件 Cross-domain Transfer Learning for Recognizing Professional Skill Entities from Chinese Recruitment Texts 于 2019 年 12 月投稿至 Information Processing and Management 期刊。

中国知网 https://www.cnki.net

第六章 总结与展望

技能词抽取是分析大规模招聘数据与劳动力市场供求关系的基础。目前针对中文技能词提取的研究工作非常少。本文将技能词抽取任务转化为序列标注问题，借鉴命名实体识别或术语抽取中的方法。但当前主流的命名实体识别或术语抽取方法专注于领域内监督学习，需要大量带标注的数据。所以针对中文网络招聘文本中的技能词抽取，本文贡献如下：

（1）提出了一种基于深度学习的中文网络招聘文本中的技能词抽取方法，将序列标注模型与语料特征相结合，在本研究中，贡献如下：1）据我们所知，本研究是首次使用基于深度学习的序列标注模型与语料特征相结合的方法抽取网络招聘中的技能词的研究；2）与经典的序列标注模型相比，本研究的模型引入了更多的语料特征并将输入层的输出与 Bi-LSTM 输出连接在一起作为 CRF 层的输入，模型的性能得到了极大提高；3）评估了加入丰富的语料特征，本研究提出的模型是否能够缓解模型对大量标注数据的依赖；4）进行了大量实验。实验结果表明，本研究的技能词抽取模型取得最佳的抽取性能。

（2）提出了基于跨领域迁移学习的技能词抽取（CDTL-PSE）的方法。在本研究中，贡献如下：1）提供 CDTL-PSE 的合理性，即将 SIGHAN 语料库分解为三个源域，并且集成学习可以提高标签序列预测的准确性，域自适应层对于跨域迁移学习至关重要；2）评估 CDTL-PSE 是否可以缓解可用标注数据的稀缺性并可以有效识别技能词。

尽管我们的初步尝试取得了一定的研究成果，但由于时间和研究水平的制约，本文也存在些不足之处。因此，在未来还可以再进行如下深度的研究：

1）我们通过爬虫程序获取大量的网络招聘数据，如何利用这些大量未标记的语料来减少对源域的依赖性；

2）目前，我们通过预训练的字符嵌入作为输入字符的形式化表示，但字符嵌入的预训练存在训练语料的语义影响，如何利用动态词嵌入来更好地提取技能词；

3）利用模型自动抽取出的技能词，如何提取出有效的结论，为学校、师生提供指导性的帮助。

中国知网 https://www.cnki.net

参考文献

[1] Javed F, Hoang P, Mahoney T, et al. Large-scale occupational skills normalization for online recruitment[J]. AI Magazine, 2018, 39(1): 5-14.

[2] Li J, Sun A, Han J, et al. A Survey on Deep Learning for Named Entity Recognition[J]. arXiv preprint arXiv:1812.09449, 2018.

[3] 冯鸾鸾，李军辉，李培峰, 等. 面向国防科技领域的技术和术语识别方法研究[J]. 计算机科学, 2019, 46(12):231-236.

[4] 孙镇，王惠临．命名实体识别研究进展综述[J]．现代图书情报技术，2010(6): 42-47.

[5] 孙茂松，黄昌宁，高海燕，等. 中文姓名的自动辨识[J]．中文信息学报, 1995, 9(2) : 16 -27.

[6] 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006，27(2):87-94.

[7] Quimbaya A. P, Múnera A. S, Rivera R A G, et al. Named entity recognition over electronic health records through a combined dictionary-based approach[J]. Procedia Computer Science, 2016, 100(100): 55-61.

[8] Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts[J]. Journal of Biomedical Informatics, 2013, 46(6): 1088-1098.

[9] Zhang J, Shen D, Zhou G, et al. Enhancing HMM-based biomedical named entity recognition by studying special phenomena[J]. Journal of Biomedical Informatics, 2004, 37(6): 411-422.

[10] Li L., Mao T., Huang D., et al. Hybrid models for Chinese named entity recognition[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, ACL, 2006: 72-78.

[11] Duan H, Zheng Y. A study on features of the CRFs-based Chinese Named Entity Recognition[J]. International Journal of Advanced Intelligence, 2011, 3(2): 287-294.

[12] Kim J H, Woodland P C. A rule-based named entity recognition system for speech input[C]// Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP), IEEE, 2000: 528-531.

[13] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.

[14] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12:2493–2537.

[15] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), ACL, 2016:260–270.

[16] Huang Z, Xu W, Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[17] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), ACL, 2016.

[18] 郑彦斌,夏志超,郭智,等. 东盟十国新闻文本的命名实体识别[J]. 科学技术与工程, 2018, 18(35): 162-168.

[19] Peng N, Dredze M. Improving Named Entity Recognition for Chinese Social Media with Word

中国知网 https://www.cnki.net

Segmentation Representation Learning[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), ACL, 2016.

[20] Wu F, Liu J, Wu C, et al. Neural Chinese Named Entity Recognition via CNN-LSTM-CRF and joint training with Word Segmentation[C]// Proceedings of the 2019 Conference of the World Wide Web(WWW), ACM, 2019: 3342-3348.

[21] Dong C H, Zhang J, Zong C Q, et al. Character-based LSTM-CRF with Radical-level Features for Chinese Named Entity Recognition[C]// Proceedings of the 5th CCF Conference on Natural Language Processing and Chinese Computing (NLPCC), Springer, 2016:239-250.

[22] 陈锋, 翟羽佳, 王芳. 基于条件随机场的学术期刊中理论的自动识别方法[J]. 图书情报工作, 2016, 60(2): 122-128.

[23] Chen Y J, Zhou C L, Shi X D. Automatic extraction of Chinese terms[C]// Proceedings of the 2005 International Conference on Natural Language Processing and Knowledge Engineering (ICNLP), IEEE, 2005: 281-286.

[24] Pantel P, Lin D. A Statistical Corpus-Based Term Extractor[C]// Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, Springer, 2001:36-46.

[25] 蒋婷, 孙建军. 基于 SVR 模型的中文领域术语自动抽取研究——面向图书情报领域[J]. 情报理论与实践, 2016, 39(1), 24-31.

[26] 王昊, 王密平, 苏新宁. 面向本体学习的中文专利术语抽取研究[J]. 情报学报, 2016, 35(6):573-585.

[27] 何宇, 吕学强, 徐丽萍. 新能源汽车领域中文术语抽取方法[J]. 现代图书情报技术, 2015, 31(10):88-94.

[28] 闫兴龙, 刘奕群, 方奇. 基于网络资源与用户行为信息的领域术语提取[J]. 软件学报, 2013, 24(09): 113-124.

[29] 赵东玥, 杜永萍, 石崇德. 基于 BLSTM 的科技文献术语抽取方法[J]. 情报工程, 2018, 4(1):67-74.

[30] 赵洪, 王芳. 理论术语抽取的深度学习模型及自训练算法研究[J]. 情报学报, 2018, 37(09):67-82.

[31] 车万翔，刘挺,李生.实体关系自动抽取[J]. 中文信息学报，2005, 19(2): 1-6.

[32] 赵京胜, 朱巧明, 周国栋, 等. 自动关键词抽取研究综述[J]. 软件学报, 2017, 28(9):2431-244.

[33] Wowczko I A. Skills and vacancy analysis with data mining techniques Informatics[J]. Multidisciplinary Digital Publishing Institute, 2015, 2(4): 31-49.

[34] Kim Y, Addom B K, Stanton J M. Education for eScience professionals: Integrating data curation and cyberinfrastructure[J]. International journal of digital curation, 2011, 6(1): 125-138.

[35] Kim J Y, Lee C K. An empirical analysis of requirements for data scientists using online job postings[J]. International Journal of Software Engineering and Its Application, 2016, 10(4):161-172.

[36] Chao C A, Shih S C. Organizational and End-User Information Systems Job Market: An Analysis of Job Types and Skill Requirements[J]. Information Technology, Learning & Performance Journal, 2005, 23(2).

[37] Lee S M, Lee C K. IT managers' requisite skills[J]. Communications of the ACM, 2006, 49(4): 111-114.

[38] Cragin M H, Palmer C L, Varvel Jr V E, et al. Analyzing Data Curation Job Descriptions (Poster

中国知网 https://www.cnki.net

and Abstract) [J]. 2009.

[39] Zhao M, Javed F, Jacob F. SKILL: A system for skill identification and normalization[C]// Proceedings of the 27th Conference on Innovative Applications of Artificial Intelligence (AAAI), AAAI Press, 2015: 4012-4017.

[40] Phaphuangwittayakul A, Saranwong S, Panyakaew S, et al. Analysis Of Skill Demand In Thai Labor Market From Online Jobs Recruitments Websites[C]// Proceedings of the 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, 2018: 1-5.

[41] De Mauro A, Greco M, Grimaldi M, et al. Human resources for Big Data professions: A systematic classification of job roles and required skill sets[J]. Information Processing and Management, 2018, 54(5): 807-817.

[42] 詹川. 基于文本挖掘的专业人才技能需求分析[J]. 图书馆论坛, 2017, 5(1): 116-123.

[43] 俞琰, 陈磊, 姜金德. 网络招聘文本技能信息自动抽取研究[J]. 图书情报工作, 2019, 63(13): 105-113.

[44] 司莉, 贾欢. 欧美信息职业对图书情报学人才需求的调查与分析[J]. 图书馆论坛, 2015, 3: 102-108.

[45] 吕斌, 张通, 周钰. 面向组织的具有通用性的情报职业及情报从业人员——基于组织招聘网页信息挖掘的分析之一[J]. 图书情报工作, 2009, 53(04): 19-23.

[46] 李国秋, 桑培铭. 情报过程——情报职业的核心：问题域及方法论——基于组织招聘网页信息挖掘的分析之二[J]. 图书情报工作, 2009, 53(04): 24.

[47] 夏火松, 潘筱听. 基于 Python 挖掘的大数据学术研究与人才需求的关系研究[J]. 信息资源管理学报, 2017,7(1): 4-12.

[48] 黄崑, 王凯飞, 王珊珊, 等. 数据类岗位招聘需求调查及对图情学科人才培养的启示[J]. 图书情报知识, 2016,6(1):42-53.

[49] 夏立新, 楚林, 王忠义, 等. 基于网络文本挖掘的就业知识需求关系构建[J]. 图书情报知识, 2016, 1: 94-100.

[50] Zhu C, Zhu H, Xiong H, et al. Person-job fit: Adapting the right talent for the right job with joint representation learning[J]. ACM Transactions on Management Information Systems (TMIS), 2018, 9(3): 1-17.

[51] Zhou W, Zhu Y, Javed F, et al. Quantifying skill relevance to job titles[C]// Proceedings of the IEEE International Conference on Big Data (Big Data), IEEE, 2016: 1532-1541.

[52] Xu T, Zhu H, Zhu C, et al. Measuring the popularity of job skills in recruitment market: A multi-criteria approach[C]// Proceedings of the Thirty-Second Conference on Artificial Intelligence (AAAI), AAAI Press, 2018:2572-2579.

[53] Qin C, Zhu H, Xu T, et al. Enhancing Person-Job Fit for Talent Recruitment: An Ability-aware Neural Network Approach[C]// Proceedings of the 41st International ACM SIGIR Conference (SIGIR-2018), ACM, 2018:25-34.

[54] Shen D Z, Zhu H S, Zhu C, et al. A Joint Learning Approach to Intelligent Job Interview Assessment[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, 2018:3542-35.

[55] Zhu C, Zhu H, Xiong H, et al. Recruitment market trend analysis with sequential latent variable models[C]// Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM ,2016: 383-392.

[56] Pan S. J., Yang Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge & Data

中国知网 https://www.cnki.net

Engineering, 2010, 22(10):1345-1359.

[57] Yang Z, Salakhutdinov R, Cohen W W. Transfer learning for sequence tagging with hierarchical recurrent networks[C]// Proceedings of 5th International Conference on Learning Representations (ICLR), IEEE, 2017.

[58] Lin B Y, Lu W. Neural adaptation layers for cross-domain named entity recognition[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2018:2012-2022.

[59] Lee J Y, Dernoncourt F, Szolovits P. Transfer learning for named-entity recognition with neural networks[J]. arXiv preprint arXiv:1705.06273, 2017.

[60] Peng N, Dredze M. Multi-task multi-domain representation learning for sequence tagging [J]. arXiv preprint arXiv:1608.02689, 2016.

[61] Kulkarni V, Mehdad Y, Chevalier T. Domain adaptation for named entity recognition in online media with word embeddings[J]. arXiv preprint arXiv:1612.00148, 2016.

[62] Xiao M, Guo Y. Domain adaptation for sequence labeling tasks with a probabilistic language adaptation model[C]// Proceedings of the International Conference on Machine Learning (ICML), IMLS, 2013: 293-301.

[63] Xu J, He H, Sun X, et al. Cross-domain and semi-supervised named entity recognition in Chinese social media: a unified model[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(11): 2142-2152.

[64] Wang Z, Qu Y, Chen L, et al. Label-aware double transfer learning for cross-specialty medical named entity recognition[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), ACL, 2018:1-15.

[65] Pan J, Hu X, Li P, et al. Domain adaptation via multi-layer transfer learning[J]. Neurocomputing, 2016, 190: 10-24.

[66] 庄福振, 罗平, 何清,等. 迁移学习研究进展[J]. 软件学报, 2015, 26(1):26-39.

[67] Dong C, Wu H, Zhang J, et al. Multichannel LSTM-CRF for named entity recognition in Chinese social media[C]// Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data(CCL), Springer, 2017: 197-208.

[68] Yang Z, Salakhutdinov R, Cohen W. Multi-task cross-lingual sequence tagging from scratch[J]. arXiv preprint arXiv:1603.06270, 2016.

[69] Daumé III H. Frustratingly easy domain adaptation[C]// Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), ACL, 2007.

[70] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]// Proceedings of the 3rd International Conference on Learning Representations (ICLR), IEEE, 2015.

[71] Devlin, J, Chang, M, Lee, k, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), ACL, 2019: 4171-4186.

[72] Emma, S, Patrick, V, David, B, et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2017: 2670-2680.

中国知网 https://www.cnki.net

致　谢

三年的研究生生涯即将结束，思及此，不禁感慨万千。回首往昔，怀着对未来的憧憬与家人的期望来到这里，留下了无数的甜美与欢笑，也付出了奋斗和辛劳的“汗水”。值此论文即将完成之际，对曾经给予我帮助的、陪伴的人表示感谢。

首先感谢我的父母，尚且不论养育之恩，无以为报。仅在我读书期间，就给予了我莫大的鼓励与支持。在尚不富裕的家庭中，父母从未让我因经济感到发愁，含辛茹苦的工作供我读书。读研之前，无论是我选择工作还是继续深造，始终都在背后默默地支持我选择的一切。真心的和父母说一声：您们，辛苦了！

然后，我要感谢我的研究生导师文益民教授。俗话说，师傅领进门，修行在个人！但，这三年来，从刚开始研究方向的选取到论文的撰写、投稿，文老师始终悉心的给与了全程指导，无数次的讨论，不厌其烦的讲解，倾注了老师大量的心血。老师身上严谨的科学态度、精益求精的工作作风，让我以后受益无穷。此外，在研究生期间文老师给我提供多次外出学习交流的机会，了解了国内外最新的研究动向，开阔了我的视野，也结识了不少新朋友。衷心祝愿，文老师在今后的工作生活中，工作顺利，身体健康！

感谢实验室的同门以及师兄弟：文坚、田野、秦一休、伊海洋、秦珂珂、魏继鹏师兄和朱望舒师姐，刘帅、韩晗同门，刘长杰、俸亚特、员喆、王利兵师弟以及周琪、茅倩师妹。三年的研究生涯，肯定不是一帆风顺，留下欢笑的同时，必然存在痛苦。感谢你们让我在枯燥、痛苦的时刻又感觉到了无限的活力与欢乐，和你们一起度过时间我一生难忘。当我在遇到问题时各位师兄都不遗余力的帮助我，和我一起讨论，给我指点迷津，让我能从不一样的角度来看待问题，受益匪浅。在查阅资料与文献时，各位师弟师妹也给与我很大的帮助！

感谢一路陪伴我的朋友们：严德东、李鑫旺、王文闯、尚孝聪、徐白杨、陈立旺。感谢你们对我的支持与鼓励，以及对我的包容。在无聊之际能够陪我谈天畅地的聊天，在我心情郁闷时，能不断的开导我，用你们自己的生活经历，鼓励我继续坚持下去，有你们在这三年我的生活添加了无数的乐趣，祝愿你们工作顺利、万事如意！

2020 年，新冠肺炎骤然降临，全国上下共同抗击疫情，向奋战在抗击新冠肺炎的一线医护人员致以最高的敬意，感谢他们为我们带来安全的生活与科研环境。

最后，感谢在百忙之中审阅我论文的专家、教授。

中国知网　https://www.cnki.net

作者在攻读硕士期间的主要研究成果

1. 学术论文

[1] 文益民, 杨鹏, 文博奚. 基于深度学习的中文网络招聘文本中的技能词抽取[J], 桂林电子科技大学学报, 2021, 40(4).

[2] Yimin Wen, Peng Yang, Boxi Wen, Meng Wang. Cross-domain Transfer Learning for Recognizing Professional Skill Entities from Chinese Recruitment Texts，已投稿至 Information Processing and Management (CCF B 类期刊)

2. 参与科研课题

[1]基于 Web 挖掘的工程硕士培养及政策支持研究. 教育部人文社会科学研究项目(No. 17JDGC022)

[2]数据流半监督分类中的多源迁移学习的研究与应用.广西自然科学基金(No. 2018GXNSFDA138006)

[3]基于多任务学习的复杂概念漂移数据流分类研究.国家自然科学基金项目(No. 61363029)

[4]数据流半监督分类中的半监督迁移学习研究.国家自然科学基金项目(No. 61866007)

3. 获奖情况

2017 年获得研究生学业奖学金二等奖

2018 年获得研究生学业奖学金二等奖

2019 年获得研究生学业奖学金三等奖

中国知网 https://www.cnki.net