[文章编号] 1003－4684(2025)04-0056-07

基于多任务学习的技能抽取

钱 奕[1]，陈飞飏[2]

(1 湖北交通职业技术学院，湖北 武汉 430073；2 香港中文大学(深圳)，广东 深圳 518172)

[摘 要] 从文本中抽取技能是一项重要的自然语言处理任务。例如，企业可以从招聘启事中抽取出技能从而了解工作市场上所需热门技能及其变化趋势。技能抽取通常需要建立机器学习模型来进行处理。然而，建模过程需要人工标注大量数据和设计特征，并且在单一文本语料上学习到的机器学习模型在应用到其它文本上时效果不佳。因此通常需要在其他文本语料上重新建模，费时费力。为此，本文提出了一种基于多任务学习的技能抽取框架 MTSE。实验结果表明，MTSE 能有效利用不同文本的数据和特征，显著提高在不同文本上的技能抽取的建模效率和任务表现。

[关键词] 技能抽取；迁移学习；多任务学习；自然语言处理

[中图分类号] TP391.13 [文献标识码] A

技能抽取是一项自然语言处理(NLP)任务，从与人们工作和学习相关的文本中抽取技能表达。这些文本包括招聘启事、个人简历和课程介绍等。例如，可以从一份机器学习工程师的招聘启事中抽取出“机器学习”“深度学习”和“Python”等所需的技能表达。

随着在线招聘网站和在线学习网站的快速发展，技能抽取变得越来越重要，并逐渐应用于各种生产场景中：企业可以通过从大量招聘启事中抽取技能，了解工作市场上所需的热门技能及其变化趋势。在线招聘网站通过分析招聘启事和个人简历，分别抽取出工作所需的技能和个人的技能，从而进行工作匹配。在线学习网站可以通过分析在线课程内容，抽取出课程所涵盖的技能，进而推荐课程和学习路径。

技能抽取的主流方法是使用机器学习模型来进行文本处理。通常，这些模型采用监督学习方法，通过标注数据和设计特征来进行模型训练。为了确保质量，数据标注通常需要由领域专家进行人工完成。由于技能表达的复杂性(技能表达灵活且与上下文相关)，领域专家为了确定一个技能表达是否正确，通常需要阅读整篇技能文本，导致整个数据标注过程费时费力。另外，将一个文本上训练得到的机器学习模型应用到其他文本上时，由于数据分布的差异，原模型的技能抽取效果可能变差。例如，使用在招聘启事数据上训练的机器学习模型直接应用于个人简历数据进行技能抽取，可能效果就不佳。因此，在不同文本上建立模型时，通常需要重新标注数据和设计特征，这个过程需要大量的时间和人力投入。

为了解决上述问题，需要一种快速且有效的在不同文本上建立技能抽取模型的方法。在研究中发现不同文本的数据分布虽然不一致，但技能抽取任务之间存在相关性。不同模型之间的标注数据是可以共享的，并且一些特征可以在不同模型之间共享，例如，技能表达的上下文嵌入。基于这个观察，本文提出了一个基于多任务学习的技能抽取框架(Multi Task Skill Extraction，MTSE)。MTSE 利用迁移学习中的多任务学习，实现了标注数据和特征在不同文本的技能抽取模型之间的共享。实验结果表明，MTSE 不仅显著提高了在不同文本上建立技能抽取模型的效率，而且各个模型的技能抽取效果也有所提升。

1 相关工作

1.1 命名实体识别

技能抽取与 NLP 中的命名实体识别(named entity recognition，NER)任务[1,2]是相关的。NER 是指从非结构化的输入文本中抽取具有特定意义或指代性的实体名称：人名、地名、机构名称等。例如，输入句子“美国国会正式确认拜登当选总统”，NER 任务会提取出：机构名——“美国国会”；人名——“拜登”和专有名词——“总统”。NER 任务一般被

[收稿日期] 2023－03－20

[第一作者] 钱 奕(1973－)，女，湖北武汉人，湖北交通职业技术学院讲师，研究方向为计算机应用。

定义为序列标注问题,即认为输入文本是一个由词组成的序列,任务是标注出词的实体标签。所以NER 任务一般采用监督学习方式建立机器学习模型来解决。目前主流的 NER 模型主要采用深度学习的方法。它首先采用神经网络的序列模型(例如,LSTM)来生成原文本的文本表征,再输入给条件随机场模型(conditional random field,CRF)进行标签标注[3-6]。

除了标准的人名、地名、机构名称和专有名称外,NER 任务可以按照不同业务需求从文本中识别出更多类别的实体,比如产品名称、型号、价格等。因此,可以将技能视为文本中的一种特殊实体,而技能抽取可以被认为是一种特殊的 NER 任务。但是,本文并没有直接使用传统 NER 模型来处理技能抽取任务。主要原因是:1)相比标准实体,技能表达更为复杂,传统模型的技能抽取准确度较低;2)包含技能的文本通常具有特定的文本结构,而传统模型无法直接利用文本结构信息。因此,MTSE 采用了两步法来抽取技能。首先,标记出文中的技能表达候选项;然后,将这些候选项输入分类器,判断是否为正确的技能,并输出正确的技能结果。

1.2　技能抽取

近年来,随着在线招聘网站和在线学习网站的迅速发展,技能抽取逐渐成为一个研究的热点。学术界和工业界都发表了许多关于技能抽取的文章[7]。大部分研究采用了监督学习的方式来建立技能抽取模型。

Sayfullina 等[8]提出了一种基于监督学习的技能抽取判断方法。该方法首先定义和建立了一个技能表达的字典,字典包括常见的技能表达。然后,根据字典中的技能表达条目,在输入文本中进行字符匹配,从而得到技能表达的候选项。由于这些技能表达候选项的准确度不高,该方法进一步训练了一个技能分类器来判断它们的正确性 。作者使用了多种机器学习分类算法(如 CNN 和 LSTM 等)和模型特征来建立不同的技能分类器。实验结果表明,基于 LSTM 训练的模型具有最佳的分类效果。

Tamburri 等[9]利用监督学习建立了一个句子分类器,用于判断输入的句子中是否包含技能表达。该方法首先通过领域专家对训练数据集进行标注(1 表示句子含有技能;0 表示句子不含有技能)。除了使用人工标注数据,该方法还使用了预训练语言模型 BERT [10] 来建立分类器。预训练语言模型有助于分类器理解输入句子的语义,进而判断句子是否包含技能。Chernova 等[11]在芬兰语料中的技能抽取任务中也进行了类似的工作。

Bhola 等[12]将技能抽取定义为一个基于文档的大规模多标签分类(extreme multi-label classification)任务,用于建立模型。例如给定一个招聘启事,该模型能够预测出与该招聘启事相关的多个工作技能。该大规模多标签分类模型也是基于 BERT 建立的。

Zhang 等[13]提出了一种从招聘启事中分别抽取技能和知识的机器学习模型。该模型定义了招聘启事中技能和知识的概念,并使用了多任务学习的方式分别抽取技能和知识,以提高模型的性能。该模型的想法与本文提出的 MTSE 框架有一些相似之处,但 MTSE 在多任务学习中采用了不同的任务设置。MTSE 目的是针对不同的技能文本建立多个技能抽取模型,并从不同的文本中抽取技能。

1.3　迁移学习和多任务学习

监督式机器学习模型,特别是神经网络模型,通常需要大量训练数据来发现数据本身的特征。然而很多任务中会经常遇到训练数据不存在或者不足的问题。如果采用人工标注方法来收集数据,整个过程费时费力。针对训练数据缺乏的问题,迁移学习[14]被提出,如图 1 所示。它的主要思想是将已有任务的模型中的知识迁移到新任务的模型中去,使得新模型的训练不用收集大量标注数据从头学习,从而加快了学习效率。

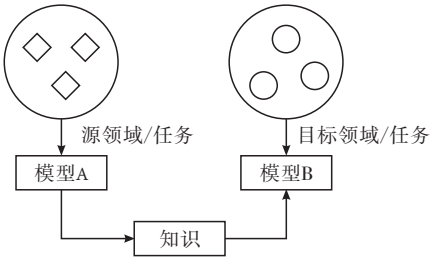

图 1　迁移学习

多任务学习属于迁移学习中的一种重要类型,是指源域和目标域的多个任务同时进行学习[15-18]。其核心思想是通过引入辅助任务(即增加目标函数)来提供先验假设,从而提升整体模型的效果。例如,社交网络的信息流推荐系统可以有多个任务,包括预测用户是否会阅读推荐文章,预测用户是否会点赞,预测用户是否会转发等。这些任务具有不同的目标函数,但彼此相关,可以通过多任务学习同时进行训练。在机器学习任务中应用多任务学习有如下好处:1)降低模型的训练和部署成本,无需对每个任务训练特定的模型;2)提高模型的泛化能力。对于数据较少的任务,如预测文章转发模型,由于用户转发文章的操作较少,标注数据较少,模型容易过拟合。如果把预测文章转发和预测文章点击这两个相

关并且常同时发生的任务进行联合学习，可以缓解模型的过拟合问题，提高模型的泛化能力。同时，对于数据稀疏的任务，多任务学习也会帮助推荐系统中的“冷启动”问题。多任务学习通常被认为是为了同时解决相关或类似任务的数据稀疏问题。多任务学习的核心问题是如何选择相关的任务。Mou 等[19]发现使用神经网络来进行 NLP 多任务学习，如果任务之间的语义相关，则模型效果更好。

多任务学习有两种主要学习方法：硬参数共享和软参数共享。硬参数共享方法如图 2 所示：

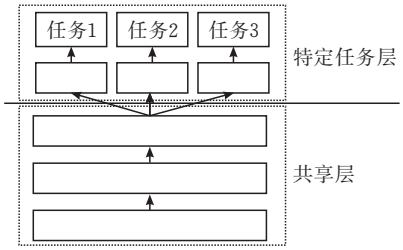

图 2 多任务学习的硬共享参数

在该类模型中，所有任务共享一些底层参数，而在特定任务层使用任务特定的参数。这种方法能减少过拟合的风险，相当于对模型进行了正则处理。Collobert 和 Weston[20]提出的进行多任务 NLP 学习的神经网络，其采用了硬共享参数的架构。

硬共享参数方法要求所有任务共享同样的隐藏空间，这样在参数优化时可能会出现冲突，因为所有任务在共享层使用同样的参数。这种方法限制了模型的表现能力，并且使得处理弱相关任务变得困难。为了解决这个问题，软参数共享方法被提出。软参数共享方法如图 3 所示。

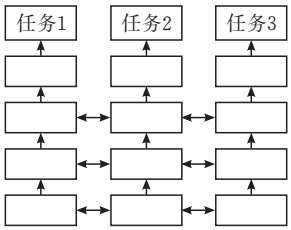

图 3 多任务学习的软共享参数

在软参数共享方法中，每个模型都有自己的参数。每个模型之间的参数会被正则化，使得模型之间的参数比较类似。Duong 等[21]建立少资源语言的依赖关系分析模型，就利用了正则化的思想。他们先建立了多资源语言的依赖关系模型，得到模型参数（包括词嵌入，模型参数等），然后将他们作为初始参数应用到少资源语言的模型中（少资源语言模型的架构和 POS 标记集合相同）再建立一个有正则化的目标函数，使得少资源模型的模型参数通过不同大小的训练数据集来进行微调。

软参数共享方法的不足之处在于它没有假设任务相关性，需要给每个任务训练一个模型，需要的参数较多。这样就要求更多的训练数据，不利于训练大规模模型。本文提出的 MTSE 的架构是采用了硬参数共享的方式。

2 技能抽取框架 MTSE

MTSE 采用了基于多任务学习的神经网络学习框架。因为在不同文本上的技能抽取任务是相关的，多任务学习可以通过多个相关任务训练来共享数据和特征，减轻对特定任务的过度拟合，以正则化来提升最终模型的性能。在技能抽取任务中，一个重要的共享特征是技能表达的上下文，在很多情况下，人们仅通过上下文就可以判断技能是否正确。所以，MTSE 使用了技能上下文嵌入作为整个框架的共享特征层。除了考虑不同任务的相关性，MTSE 也考虑到了不同任务之间的差异性。它在共享特征层的基础上，针对每个任务的特点建立了单独的神经网络模型。MTSE 即利用了多任务学习的优点使数据和特征在不同文本的技能抽取模型上共享，也考虑到了各个任务本身的特点，从而显著提高了建模效率和模型表现。

2.1 问题定义

在介绍 MTSE 的架构之前，需要定义技能抽取的相关概念。

1）技能　国际标准化组织（ISCO 和 ESCO）对技能的定义[22]是指人们应用知识并且知道如何完成任务和解决问题的能力。

2）技能文本　指人们在工作和学习中使用的与技能相关的文本，例如，招聘启事，个人简历和课程介绍。

3）技能抽取　指从技能文本中抽取技能表达的 NLP 任务。其主流方法是建立机器学习模型从技能文本中抽取技能。模型会生成一个映射函数 f：$T \to [0,1]^S$，这里，$S = |\mathbb{S}|$，$\mathbb{S}$是技能表达的集合。即给定技能文本 $t \in T$，映射函数 f 针对每个技能 $s \in \mathbb{S}$输出一个概率值，即

$$f(t) = P(r_i = 1 \mid t) \tag{1}$$

这里 $i \in \{1,\cdots,S\}$，r_i 是对第 i 个技能的标注。

2.2 模型架构

MTSE 的基本架构如图 4 所示。图 4 中显示了 MTSE 被应用于招聘启事，个人简历和课程介绍这三种技能文本的技能抽取任务中。其架构主要由四大模块组成。

2.2.1 文本结构分析模块　该模块将输入技能文本进行结构分析，将其划分成为不同的语义段落，并输出相应的段落信息。

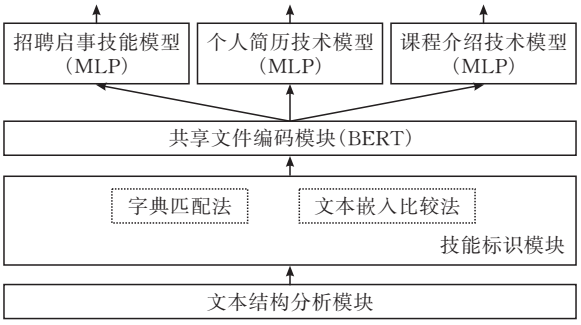

图 4　MTSE 的模型架构

1 篇技能文本通常由多个语义段落组成,每个段落具有特定的模板和功能。例如,招聘启事通常包含以下段落:1)公司概况:介绍公司的使命和愿景;2)技能要求:列举工作所需的必备技能和首选技能;3)福利待遇,描述公司提供的休假和五险一金等福利信息。类似地,个人简历也包含多个段落,如个人总结、工作经验和职业证书等。

文本结构分析模块的输出信息有助于 MTSE 理解技能文档的结构,并生成有效的特征和信号来辅助建立技能抽取模型。例如,个人简历通常包括一个特定的"技能"段落,其中列举了简历持有者具备的所有技能。如果文本结构分析模块能够识别该段落,则其中的所有词语都可以被视为是技能表达的候选项。

文本结构分析模块的处理逻辑与输入技能文本类型相关,因为不同类型的文本具有不同的段落模板,因此需要采用不同的处理方法。文本结构分析可以制定规则或者建立机器学习模型来处理。具体的算法不是本文介绍的重点,感兴趣的读者可以参考文献[23]中的相关内容。

2.2.2　技能标识模块　该模块处理输入的技能文本,标记并输出技能表达的候选项。

MTSE 参考了 Sayfullina 等[8]的字典匹配法来标记技能表达候选项。首先,MTSE 让领域专家建立一个技能分类表。该表包含了 9617 个技能条目,每个条目由一个技能的标准名称和对应的多个别称组成。例如,标准名称为"Java"的技能条目,可能有别称"Java SE"和"Java EE"等。得到了技能分类表后,技能候选项标识模块会根据技能标准名称和其别称在输入技能文本中进行完全字符匹配来标识技能表示候选项。技能表达匹配的过程使用了文本结构分析模块的段落分析结果。例如,在招聘启事中,MTSE 不会在"公司福利"段落中匹配和发现工作所需的技能,因为招聘公司通常不会在"公司福利"段落中提及工作所需技能。

字典匹配方法在技能表达的精确率方面表现较好,但召回率相对较低。有些文本中潜在的技能表达无法通过字符匹配直接发现。因此,除了字符匹配法,MTSE 还提供了文本嵌入匹配法来抽取技能表达的候选项。该方法的步骤如下:1)首先对输入句子进行词类分析,获取句子中所有字符的词类标识;2)定义词类标识序列模板,并根据模板将具有特定词类标识的字符序列作为技能表达可能项(表 1 展示了词类标识序列模板的示例);3)使用词嵌入技术(MTSE 使用了 BERT),分别生成技能表达可能项的嵌入,和技能分类表中所有标准技能的嵌入,然后计算它们之间的文本相似度(如余弦相似度);最后,根据预先设定的阈值,确定最终的技能候选项。例如,在文本中出现的技能表达"管理一个团队",将其嵌入与分类表中所有技能嵌入进行对比,发现其与条目"团体管理"的嵌入值相似度较高,大于阈值 0.6,则"团体管理"被认为是一个技能表达候选项。相比字典匹配法,文本嵌入匹配法的优点是技能抽取召回率较高,但是准确度较低。在实验阶段,将比较这两种不同的方法,并进行评估。

表 1　词类标识序列模板

模板名称	模板序列	模板示例
名词序列	"名词"或者"名词"+"名词"…	机器/名词 学习/名词
动名词序列	"动词"+(量词 /形容词)+"名词"	驾驶/动词 汽车/名词
动词短语	"动词"	设计/动词

2.2.3　共享文本编码模块　该模块用于生成在不同文本上的技能抽取任务所共享的特征。正如前文所述,技能的上下文是技能抽取模型的重要共享特征。共享文本编码模块将技能的上下文作为输入,并进行编码(上下文的字符窗口大小为 10)。MTSE 使用了预训练模型 BERT 作为上下文编码器。该模块最终使用 BERT 输出的[CLS]编码作为技能的上下文表征。

2.2.4　不同文本的特定技能分类器　该模块用于针对不同类型的文本建立特定的技能分类模型。如图 4 所示,MTSE 针对招聘启事、个人简历和课程介绍分别建立了三个独立的技能分类器。每个分类器都采用神经网络模型(MLP) 进行建模,其输入是技能标识模块输出的技能表达候选项。这些模型共享技能的上下文嵌入特征,但也具有针对各自文本类型的特定特征(例如,在招聘启事文本中,检查技能表达候选项是否出现在"公司基本情况"段落)。给定输入的技能候选项和文本,各个模型输出对应的技能概率值。

MTSE 采用了多任务学习的硬共享参数架构。

所有任务的训练数据可以共享。通过增加训练数据，MTSE可以同时学习多个任务，得到更通用的表达，而且对其中训练数据较少或者难学到特征的任务，可以通过其他任务来辅助学习。

3 实验及结果分析

3.1 数据收集

从在线招聘网站和在线学习网站等资源中随机采集了30 128个招聘启事（简称JP），1630份个人简历（简称PR）和2450多个课程介绍（简称CI）。所有的技能文本都通过人工进行了技能标注。各个不同技能文本的技能统计信息如表2所示。

表2 技能文本的统计信息

文本类型	工作启事（JP）	个人简历（PR）	课程介绍（CI）
文本数量	30 128	1630	2450
包含独特技能名称数量	3250	2839	2790
文本平均包含技能数量	28	19	16
文本平均包含字符数	1256	1170	860
文本最大包含字符数	3890	2980	1698

在训练和评价不同模型时，我们将训练数据按3∶1∶1的比例分别作为训练数据集，验证数据集和测试数据集。

3.2 模型比较

为了评价MTSE模型的技能抽取结果，使用了几个基本模型与MTSE模型来进行对比实验。使用的基本模型包括：1）CNN模型[24]；2）LSTM模型[25]；3）BiLSTM模型[26]；4）BiGRU模型[27]。以上模型与MTSE模型均在单一文本上进行实验且均使用文本结构分析模块和技能标识模块，即各个模型从输入文本中抽取的技能标识候选项是一致的。

使用了如下指标来评价模型：1）精确率（precision），表示模型抽取出技能的准确性，值的区间在[0,1]；2）召回率，表示模型抽取出准确技能的覆盖率，值的区间在[0,1]；3）平均精度均值（MAP），表示模型抽取出技能的精确性，值的区间在[0,1]；4）平均倒数排名（MRR），表示模型抽取出第一个正确技能的结果在列表中位置，值的区间在[0,1]。

使用基本模型和在三个不同技能文本集合（分别用JP，PR和CI表示）上的实验结果，如表3所示。

表3 技能抽取模型的结果

模型	Precision			Recall			MAP			MRR		
	JP	PR	CI	JP	PR	CI	JP	PR	CI	JP	PR	CI
CNN	0.682	0.633	0.595	0.703	0.673	0.647	0.621	0.563	0.542	0.819	0.785	0.761
LSTM	0.691	0.655	0.601	0.711	0.688	0.657	0.631	0.571	0.552	0.837	0.791	0.772
BiLSTM	0.714	0.668	0.621	0.723	0.693	0.668	0.652	0.581	0.556	0.851	0.796	0.781
BiGRU	0.726	0.678	0.632	0.734	0.697	0.674	0.663	0.602	0.577	0.877	0.813	0.792
MTSE-Single	0.748	0.692	0.655	0.759	0.714	0.693	0.713	0.631	0.602	0.912	0.836	0.812
MTSE-Multi	0.751	0.732	0.730	0.769	0.733	0.726	0.749	0.652	0.627	0.918	0.862	0.844

从表中可以看出，MTSE在单个数据集上的表现均强于其他模型。主要原因是MTSE使用了预训练模型BERT。此外，MTSE的更大优势在于能通过多任务学习利用不同文本的训练数据和特征。MTSE-Single和MTSE-Multi显示了MTSE分别在没有使用多任务学习与使用多任务学习下得到的不同结果。明显可以看出MTSE-Multi技能抽取模型的效果更好。

为了显示MTSE框架中不同模块的作用，还做了模块消融实验，来了解各个模块在MTSE中的作用。MTSE- w/o-seg表示不使用文本结构分析模块；MTSE-w/o-emb表示不使用共享文本编码模块，只是共享训练数据；MTSE-all表示所有模块都被使用。

表4 MTSE消融实验结果

模型	Precision			Recall			MAP			MRR		
	JP	PR	CI	JP	PR	CI	JP	PR	CI	JP	PR	CI
MTSE-w/o-seg	0.743	0.711	0.698	0.750	0.701	0.710	0.724	0.641	0.610	0.913	0.830	0.813
MTSE-w/o-emb	0.748	0.721	0.717	0.764	0.711	0.712	0.731	0.648	0.613	0.915	0.842	0.822
MTSE-all	0.751	0.732	0.730	0.769	0.733	0.726	0.749	0.652	0.627	0.918	0.862	0.844

此外，还对MTSE技能标识模块的两种不同技能标识方法（字典匹配法和文本嵌入比较法）进行了对比实验。其精确率和召回率对比结果分别如图5和图6所示。

从图中可以看出，使用字典匹配法的模型的技能抽取精确率较高，而使用文本嵌入比较法技能抽取的召回率较高。

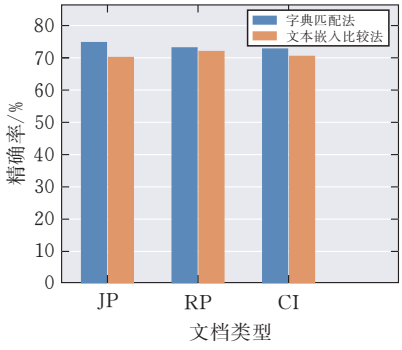

图 5　技能标识方法精确率对比

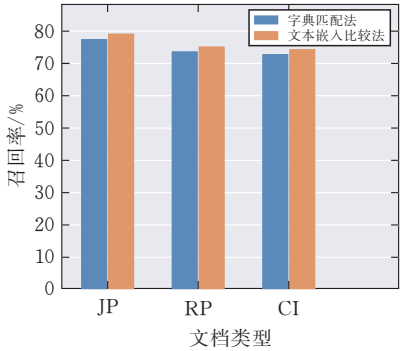

图 6　技能标识方法召回率对比

4　结论

本文提出了一种基于多任务学习的技能抽取框架 MTSE。其利用多任务学习在不同文本上共享标记数据和模型特征,而且又考虑了不同任务的特殊性,分别建立了单独的技能抽取分类模块。实验结果表明,MTSE 能显著提高在不同文本上建立技能抽取模型的效率,而且各个模型的技能抽取效果均有所提升。为了进一步提高 MTSE 模型表现,将增加训练数据,提高共享文本编码表征能力以及改进特定技能分类器模型。

[参 考 文 献]

[1] JING LI, AIXIN SUN, JIANGLEI HAN, et al. A survey on deep learning for named entity recognition [J]. IEEE Transactions on Knowledge and Data Engineering. 2020, 34(01): 50-70.

[2] 刘浏,王东波.命名实体识别研究综述[J].情报学报, 2018, 37(03):329-340.

[3] ZHIHENG HUANG, WEI XU, KAI YU. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. (2015-8-19). https://doi.org/10.48550/arXiv.1508.01991.

[4] XUEZHE MA, EDUARD HOVY. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [C]. // In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, 1: 1064-1074.

[5] GUILLAUME LAMPLE, MIGUEL BALLESTEROS, SANDEEP SUBRAMANIAN, et al. Neural architectures for named entity recognition[C]. // In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2016: 260-270.

[6] JASON P C CHIU, ERIC NICHOLS. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016(04), 357-370.

[7] IMANE KHAOUJA, ISMAIL KASSOU, MOUNIR-GHOGHO. A survey on skill identification from online job ads[J]. IEEE Access. 2021(09):118134-118153.

[8] LUIZA SAYFULLINA, ERIC MALMI, JUHOKANNALA. Learning representations for soft skill matching[C]. // In Proceedings of the International Conference on Analysis of Images, Social Networks and Texts, 2018: 141-152.

[9] DAMIAN A TAMBURRI, WILLEM-JAN VAN DEN HEUVEL, MARTIN GARRIGA. Dataops for societal intelligence: a data pipeline for labor market skills extraction and matching[C]. // In Proceedings of the IEEE 21st International Conference on Information Reuse and Integration for Data Science, 2020: 391-394.

[10] JACOB DEVLIN, MING-WEI CHANG, KENTON LEE, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. // In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, (01):4171-4186.

[11] MARIIACHERNOVA. Occupational skills extraction with FinBERT[D]. Helsinki: Metropolia University of Applied Sciences, 2020.

[12] AKSHAY BHOLA, KISHALOY HALDER, ANIMESH PRASAD, et al. Retrieving skills from job descriptions: a language model based extreme multi-label classificationframework[C]. In Proceedings of the 28th International Conference on Computational Linguistics, 2020: 5832-5842.

[13] MIKE ZHANG, KRISTIAN JENSEN, SIF DAM SONNIKS, et al. SkillSpan: hard and soft skill extraction from English job postings[EB/OL]. (2022-04-27) https://doi.org/10.48550/arXiv.2204.12811.

[14] 庄福振,罗平.迁移学习研究进展[J]. 软件学报,2015, 26(01):26-39.

[15] RICH CARUANA. Multitask learning[J]. Machine Learning, 1997, 28: 41-75.

[16] SEBASTIAN RUDER. Neural transfer learning for natural language processing [D]. Galway: University of Galway, 2019.

[17] ALEX KENDALL, YARIN GAL, ROBERTO CIPOLLA. Multi-task leanring using uncertainty to

weigh losses for scene geometry and semantics[EB/OL]. (2018-4-24) https://doi.org/10.48550/arXiv:1705.07115.

[18] 张钰,刘建伟.多任务学习[J].计算机学报,2020,43(07):1340-1378.

[19] LILI MOU, ZHAO MENG, RUI YAN, et al. How transferable are neural networks in NLP applications [C]. // In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 479-489.

[20] RONAN COLLOBERT, JASON WESTON. A unified architecture for natural language processing: deep neural networks with multitask learning[C]. // In Proceedings of the 25th International Conference on Machine Learning, 2008: 160-167.

[21] LONG DUONG, TREVOR COHN, STEVEN BIRD, et al. Low resource dependency parsing: cross-lingual parameter sharing in a neural network parser[C]. // In Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2015(02): 845-850.

[22] MARTIN LE VRANG, AGIS PAPANTONIOU, ERIKA PAUWELS, et al. Esco: Boosting job matching in Europe with semantic interoperability[J].Computer, 2014, 47(10):57-64.

[23] SONG MAO, AZRIEL ROSENFELD, TAPAS KANUNGO. Document structure analysis algorithms: a literature survey[J]. Document Recognition and Retrieval, 2003: 197-207.

[24] YOON KIM. Convolutional neural networks for sentence classification[C]. // In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1746-1751.

[25] TIM ROCKTASCHEL, EDWARD GREFENSTETTE, KARL MORITZ HERMANN, et al. Reasoning about entailment with neural attention [EB/OL]. (2015-9-22). https://doi.org/10.48550/arXiv.1509.0666.

[26] CHENGJIE SUN, YANG LIU, CHANGE JIA, et al. Recognizing text entailment via bidirectional lstm model with inner-attention[C]. // In Proceedings of the International Conference on Intelligent Computing, 2017: 448-457.

[27] KISHALOY HALDER, LAHARI PODDAR, MINYEN KAN. Cold start thread recommendation as extreme multilabel classification[C]. // In Proceedings of the International Web Conference, 2018: 1911-1918.

Skill Extraction Based on Multi-Task Learning

QIAN Yi[1], CHEN Feiyang[2]

(1 *Hubei Communications Technical College, Wuhan* 430073, *China*;

2 *The Chinese University of Hong Kong (Shenzhen), Shenzhen* 518172, *China*)

Abstract: Skill extraction from text is an increasingly important natural language processing task. For example, enterprises can extract skills from job postings to understand popular skills required in job market and their trends. Skill extraction is generally performed through building a machine learning model. However, the model building process needs much effort in data labeling and feature design, and the learned machine learning model from a single data source may not perform well in other texts. Thus, we have to label data and design features again based on other texts, which is time-consuming and label-intensive. The paper proposed a skill extraction framework based on multi-task learning- MTSE. The experiment results show that MTSE can leverage data and features across different texts and improve model building efficiency and performance on different texts significantly.

Keywords: skill extraction; Transfer Learning; Multi-task Learning; NLP

[责任编校:赵 瞳]