计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(1):75−86 [doi: 10.15888/j.cnki.csa.008901]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

人身保险知识图谱的构建与应用①

陈浩远[1], 何震瀛[2], 刘晓清[2], 杨　阳[3], 汤路民[3]

[1](复旦大学 软件学院, 上海 200438)
[2](复旦大学 计算机科学技术学院, 上海 200438)
[3](星环信息科技 (上海) 股份有限公司, 上海 200233)
通信作者: 刘晓清, E-mail: xqliucs@fudan.edu.cn

摘　要: 辅助投保人了解保险产品的条款是保险应用关注的热点问题之一, 借助知识图谱技术辅助人身保险业务开展是一种可行的方法. 本文首先从多源数据中提取并构建人身保险知识图谱 LIKG. 具体而言, 构建 BERT-IDCNN-BiLSTM-CRF 模型提取非结构化文本数据的实体, 通过多种短文本相似度算法以及集成排序算法完成实体对齐; 设计并使用 Bootstrapping 和分类预测两阶段抽取方法对保险产品进行属性填充. 然后, 根据构建的 LIKG, 设计开发原型系统, 该系统使用实体抽取和属性抽取算法提供知识获取功能、设计 CF-IIF 指标提供属性推荐功能以及实现可视化界面帮助用户快速掌握人身保险产品的信息, 展示 LIKG 的应用价值.

关键词: 人身保险; 知识图谱; 实体抽取; 属性抽取; 智能推荐

引用格式: 陈浩远,何震瀛,刘晓清,杨阳,汤路民.人身保险知识图谱的构建与应用.计算机系统应用,2023,32(1):75–86. http://www.c-s-a.org.cn/1003-3254/8901.html

Construction and Application of Life Insurance Knowledge Graph

CHEN Hao-Yuan[1], HE Zhen-Ying[2], LIU Xiao-Qing[2], YANG Yang[3], TANG Lu-Min[3]

[1](Software School, Fudan University, Shanghai 200438, China)
[2](School of Computer Science, Fudan University, Shanghai 200438, China)
[3](Transwarp Technology (Shanghai) Co. Ltd., Shanghai 200233, China)

Abstract: Assisting users in understanding the clauses of insurance products is one of the hot issues in insurance applications. It is feasible to assist the life insurance business with knowledge graph technology. The life insurance knowledge graph (LIKG) is extracted and constructed by multi-source data. Specifically, the BERT-IDCNN-BiLSTM-CRF model is applied to extract entities from unstructured data, and the entity is aligned by a variety of short text similarity algorithms and ranking ensemble algorithm. A two-stage extraction algorithm is designed to fill the attributes of insurance products by Bootstrapping and classification prediction. Then a prototype system is designed based on LIKG. The system uses the entity extraction and the attribute extraction to provide knowledge acquisition, designs an index called CF-IIF to provide attribute recommendation function, and realizes a visual interface to help users quickly master the information of life insurance, which demonstrates the application value of LIKG.

Key words: life insurance; knowledge graph; entity extraction; attribute extraction; intelligent recommendation

人身保险是一种以人的寿命或身体为保险标的的险种, 在被保险人的生命或身体发生保险事故或保险期满时, 依照保险合同的规定, 由保险人向被保险人或受益人给付保险金的保险形式, 主要包含人寿保险、

① 基金项目: 国家自然科学基金 (61732004, 62072113)
收稿时间: 2022-05-31; 修改时间: 2022-06-27; 采用时间: 2022-07-06; csa 在线出版时间: 2022-08-26
CNKI 网络首发时间: 2022-11-15

伤害保险、健康保险 3 种[1]. 面对愈发激烈的市场竞争, 辅助用户了解合同条款, 并有效匹配客户需求可提高保险公司在该业务竞争中的实力. 由于保险条款中包含了大量令人难以理解的专有名词, 这会极大地降低消费者的用户体验. 在面对这些晦涩难懂的条款时, 消费者有了解待购买保险产品条款的需求, 如何从保险条款中提取出有用的领域知识帮助消费者快速了解保险产品是人们面临的一个重要问题.

近年来, 知识图谱技术受到了学术界和工业界的广泛关注[2]. 知识图谱常用 RDF[3] 等形式来表示和管理数据, 并且具有强大的语义表达能力, 被应用到医疗[4]、教育[5] 等不同领域中.

基于人身保险领域应用知识图谱进行业务辅助的工作已有出现, 但市场中存在的部分保险知识图谱产品, 如 InsuranceAI[6]、中国疾病保险知识图谱[7] 等, 普遍存在不能面向用户提供知识抽取功能、保险产品属性较少、保险类型较少等问题. 现存的实体抽取任务通常利用序列标注模型完成, 但其需要大量的数据对模型进行训练, 而当前缺少已标注的人身保险数据. 对于属性抽取任务, 由于人身保险条款数据通常使用一段文本内容描述某个属性, 常用的序列模型主要用于抽取短文本内容形式的实体, 难以应用于人身保险的属性抽取任务.

针对上述问题, 本文通过构建知识图谱的本体和设计的知识抽取方法, 从结构化和非结构化数据中构建人身保险知识图谱 (life insurance knowledge graph, LIKG), 并基于 LIKG 开发了原型系统. 本文完成的主要工作如下.

1) 构建人身保险条款知识图谱 LIKG, 设计人身保险知识图谱 LIKG 原型系统. 根据知识抽取算法提供知识获取功能; 根据知识图谱多种人身保险类型设计条款属性相关性指标 CF-IIF 并提供知识推荐功能; 根据知识图谱中详细的人身保险产品条款属性提供属性查询功能.

2) 用结构化数据中存在的实体信息在非结构化数据中反向标注形成可训练的语料库. 构建 BERT_IDCNN-BiLSTM-CRF 模型, 并对比多种基于神经网络的实体识别模型, 选择最适合人身保险实体抽取的方案.

3) 设计了两阶段属性抽取方法, 通过使用 Bootstrapping 算法获取实体候选属性, 然后使用分类模型将其映射到对应保险实体属性槽中, 实现端到端的属性抽取, 解决了序列模型不适用将长文本属性作为预测目标的问题.

1 背景知识及相关工作

1.1 知识图谱

知识图谱是一种语义网络知识库, 可以用来表示和管理数据, 其数据存储形式使得知识图谱具有强大的语义表达能力. 知识图谱主要包含 3 类元素: 实体、关系和属性. 其中实体可以代表人或者具体事务; 关系用来连接两个实体, 表示它们之间的一些联系; 属性用来充实一个实体, 使实体具有更多的信息. 近年来有许多大型知识图谱被构建和发布, 如 Freebase[8]、YAGO[9]、DBpedia[10] 等, 它们包含大量的通用知识, 可以提供智能问答[11]、知识推荐[12] 等功能. 然而这些通用知识图谱缺少某些特定领域的知识, 当涉及领域知识的获取时, 通用知识图谱不能够很好地提供服务, 因此构建特定领域的知识图谱是很有必要的. 王建勋等人[13] 利用知识图谱技术探寻我国干旱遥感监测研究领域的研究现状, 为干旱遥感监测方法的完善与发展提供支持; 江双五等人[14] 构建高质量的气象档案知识图谱, 为我国气象档案的知识组织提供理论框架; Zhao 等人[15] 在汽车工业中构建了知识图谱, 用以帮助了解汽车相关知识. 本文在人身保险条款领域中探索如何构建知识图谱及其应用, 实验结果和系统展示分别证明了本文使用的知识抽取算法的有效性和人身保险知识图谱的应用价值.

1.2 实体识别

命名实体识别旨在从给定非结构化文本数据中抽取出自定义类别的实体数据, 是知识图谱自动化构建过程中的关键技术. 目前研究将命名实体识别任务视为序列标注任务, 利用神经网络对待抽取文本进行预测. Huang 等人[16] 提出使用双向长短时记忆网络配合条件随机场模型 (BiLSTM-CRF) 对序列数据进行预测, 其中 BiLSTM 可以有效地学习待抽取实体的上下文信息, CRF 可以通过当前状态选择下一个状态, 有效地解决 BiLSTM 在实体识别中产生的问题, 是命名实体识别任务中代表性工作之一. Stubell 等人[17] 提出 IDCNN (iterated dilated convolutional neural network) 模型, 在实体识别任务中的效果与 BiLSTM-CRF 方法相当, 由于 IDCNN 模型结构相对简单, 其训练速度与预测速度和 BiLSTM 相比相对更快. 近年来, 预训练模型的成功使得模型能很好地理解文本的语义信息, 如 Google 提

出的 BERT 模型[18]. Dai 等人[19] 构建了 BERT-BiLSTM-CRF 模型从中文电子病历中抽取出实体, 利用 BERT 模型强化词向量的学习并取得了不错的效果.

1.3 属性抽取

属性抽取旨在从非结构文本中抽取出给定实体的属性值, 使实体富有更多的信息. 早期的属性抽取任务主要通过人工制定的规则来完成. 张凯伦[20] 使用基于关键词和简单人工规则方法从非结构化和半结构化的文本数据中抽取出人物属性. 但由于不同的领域通常会有不同的关键词和语言特点, 基于规则的方法扩展性不强, 难以从某个领域数据迁移到另一个领域中, 而且随着规则的增多, 制定的规则会产生内部冲突. 基于深度学习的属性抽取方法类似于命名实体识别技术, 该类方法可以使用模型学习不同领域的数据特点, 根据训练数据泛化到不同领域中. 其通常将属性抽取视为序列标注任务, 通过序列模型识别文本中实体的属性值, 解决基于规则的方法扩展性不强和规则冲突等问题.

然而由于人身保险条款通常使用整个段落来描述实体属性, 其不适合使用用于短文本识别的序列标注任务, 因此本文提出两阶段属性抽取算法. 首先, 针对实体设计其包含的属性槽, 再使用弱监督算法 Bootstrapping[21] 抽取出所有候选属性, 然后使用集成排序 (rank ensemble, 也被称为 ranking-based ensemble 或 ensemble ranking)[16] 和分类模型将候选属性填充到对应的属性槽中, 完成人身保险属性抽取任务.

1.4 本体构建

本体是对特定领域之中某套概念及其相互之间关系的形式化表达, 是描述客观事物的抽象模型, 其目标是为了捕获相关领域的知识, 是对知识图谱模式层的管理[22]. 通过构建本体模型, 可以规范管理知识图谱中的实体、关系以及实体属性.

2 人身保险知识图谱构建

由于保险业属于第三产业中的金融服务业, 其专业知识存在一定封闭性, 人身保险条款文件中包含大量保险条款内容, 使得用户很难快速定位到想要了解的保险条款. 本文针对这一痛点, 根据人身保险条款的数据特点以及业务逻辑, 自顶向下构建人身保险条款知识图谱, 其具体流程主要包括知识图谱本体构建、知识获取、知识融合和知识存储, 构建流程图如图 1 所示.

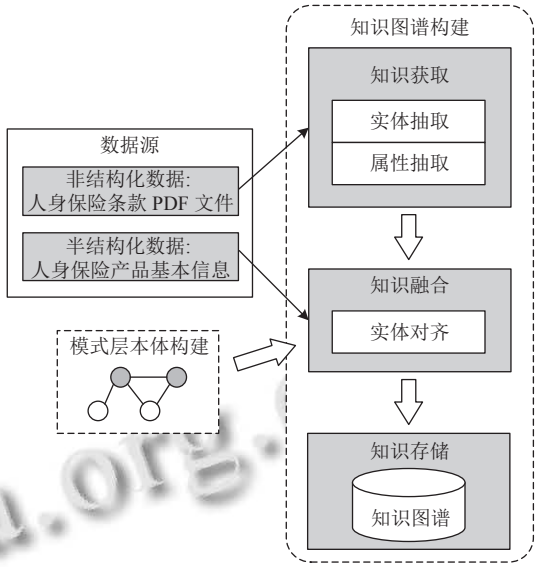

图 1 LIKG 构建整体流程

人身保险条款内容通常存储在 PDF 文件中, 少量信息如保险公司名称、保险产品可以从半结构化数据 HTML 中获取. 对于 PDF 文件中的信息, 本文根据本体构建内容, 利用实体抽取算法和属性抽取算法对其中非结构化文本数据进行知识提取, 然后对提取出的知识与现有知识融合, 最终存入到数据库中完成从异构数据中构建 LIKG.

本节将介绍 LIKG 构建的关键技术, 组织架构如下: 第 2.1 节介绍 LIKG 的本体构建过程; 第 2.2 节介绍 LIKG 实体抽取的内容, 主要包括实体抽取的数据来源、实体抽取算法和实体对齐算法; 第 2.3 节介绍人身保险实体的属性抽取算法.

2.1 本体构建

利用 LIKG, 将保险条款中各个属性条款中的内容用图谱的形式展示给用户, 以辅助用户快速定位保险条款的重要内容, 是本文的核心任务. 在这一过程中, 保险实体的属性是本文本体模型设计的核心. 本文利用保险业领域专家设计的人身保险的 schema 来构建本体模型, 然后以本体模型为核心构建人身保险条款知识图谱. 人身保险业务的部分实体和属性间的层次结构如图 2 所示.

2.2 人身保险实体抽取

2.2.1 数据来源

本文共收集 14 124 份脱敏的人身保险条款文件, 每个人身保险 PDF 文件均对应一个保险产品, 包含保险产品所属公司、保险产品名称和该保险产品的所有

属性. 对于人身保险条款领域中的实体, 本文将其分为两类: 保险公司 (organization) 和保险产品 (production). 对于非结构化文本数据中实体的自动抽取任务, 需要基于已标注数据对序列模型进行训练, 因此本文首先从人身保险文件对应的网页链接获取该保险产品对应的保险公司和保险产品名称, 然后使用获取到的保险公司和保险产品名称在保险文件中进行反向标注, 收集训练数据. 数据集具体信息如表 1 所示.

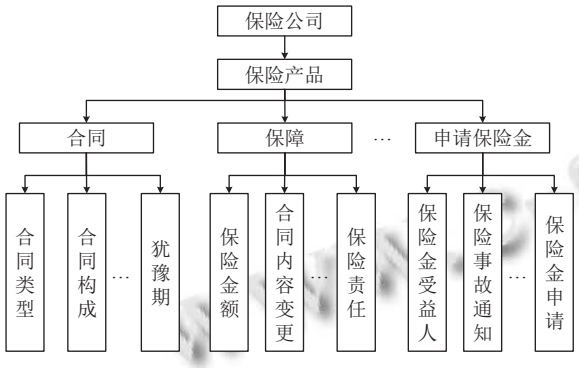

图 2 人身保险条款部分核心概念

表 1 实体标注情况

实体类型	标签含义	实体数量
ORG	保险公司	21 286
PROD	保险产品	15 032

对于数据集的标注, 本文采用序列标注中的两种常用方式: BIO 和 BIOES. BIO 标注法共有 3 类标签, 分别是 B、I、O, 其中 B 代表实体的开头, I 代表实体的中间部分或结尾, O 代表非实体. BIOES 标注法共有 5 类标签, 分别是 B、I、O、E、S, 其中 B 代表实体的开头, I 代表实体的中间部分, E 代表实体的结尾, O 代表非实体, S 代表单个字符, 其本身就是一个实体.

2.2.2 BERT-IDCNN-BiLSTM-CRF

本文构建的 BERT-IDCNN-BiLSTM-CRF 模型主要由 4 部分组成: BERT 特征表示层、IDCNN-BiLSTM 特征学习层、CRF 知识推理层. 首先将待预测文本字符序列输入到 BERT 特征表示层中, 该模块对每个字符进行编码并得到对应的词向量; 然后将词向量特征输入到 IDCNN 模型中初步提取特征; 之后将 IDCNN 提取到的特征输入到循环神经网络 BiLSTM 中提取深度特征; 最后, 利用 CRF 预测出概率最大的标签序列. 模型结构如图 3 所示.

BERT 是由 Google 提出的预训练模型, 其采用多层双向 Transformer 结构[23], 每个单元由自注意力机制 (Self-Attention) 组成. 该模型使用 Masked 语言模型和下一句预测 (NSP) 两种无监督预训练任务, 并在大规模的语料库中进行训练, 可以充分学习每个字词的含义. 该模型可以解决在训练 IDCNN 和 BiLSTM 模型过程中, 训练集规模较小而产生模型不能较好的学习文本语义的问题.

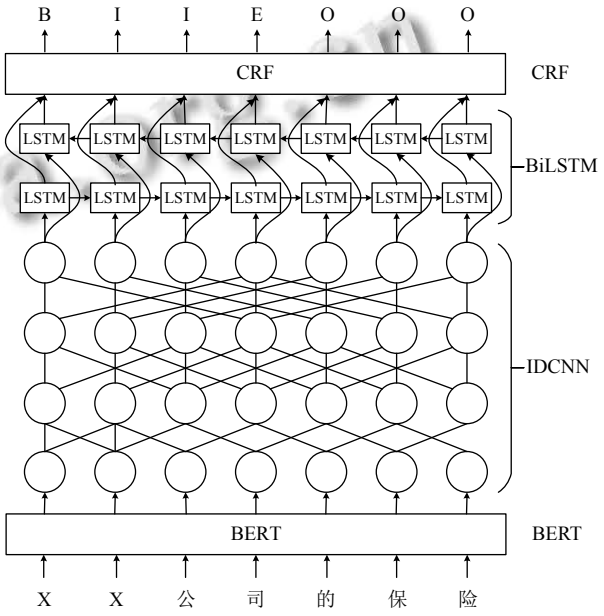

图 3 BERT-IDCNN-BiLSTM-CRF 模型结构

IDCNN 和 BiLSTM 模型利用卷积神经网络和循环神经网络学习文本的整个语义信息, 根据当前字词的上下文特征以及词向量信息预测每个字词在序列中对应的标签.

CRF 可以考虑标签之间的相邻关系, 在训练过程中学习到标签之间的约束信息, 如实体的头标签为“B”“I”标签不能跟在“O”标签之后等约束, 解决了 IDCNN 和 BiLSTM 模型忽略这些约束的问题. 其算法具体步骤如下.

对于给定序列$x=(x_1,x_2,\cdots,x_n)$和对应的标签序列$y=(y_1,y_2,\cdots,y_n)$, 其评分公式如式 (1):

$$s(x,y)=\sum_{i=1}^{n}(W_{y_{i-1},y_i}+p_{i,y_i}) \tag{1}$$

其中, W是转移矩阵, W_{y_{i-1},y_i}表示标签转移分数. 之后采用 Softmax 函数得到预测序列的最大概率标签, 如式 (2) 所示:

$$p(y|x)=\frac{e^{s(x,y)}}{\sum_{\hat{y}\in Y}s(x,\hat{y})} \tag{2}$$

2.2.3 实体对齐

人身保险文件以 PDF 的格式存储, 因此在抽取实体或者属性前需要将其转换成可处理的文本数据. 对于数据格式的转换, 本文使用百度开源的飞桨平台 PaddlePaddle[24] 中的光学字符识别 (OCR[25]) 工具识别 PDF 文件中的文本数据并保存到 TXT 文件中. 由于 OCR 技术在文字识别过程中会产生误差, 如将“中国人寿”识别为“中国大寿”等, 因此算法还需对识别到的实体内容进行实体对齐.

本文采用多种短文本相似度计算方法计算序列模型抽取到的实体与现有实体之间的相似度, 完成实体对齐. 本文使用以下几种方式计算实体间的相似度.

1) Jaccard 系数: Jaccard 系数可以用于比较有限样本集之间的相似性与差异性, Jaccard 系数值越大, 样本相似性越高, 反之, Jaccard 系数值越小, 样本差异性越大, 其计算公式如式 (3) 所示. 本文对短文本进行字符级别的相似度计算, 将实体名称分成字符级别的集合并计算两个集合间的 Jaccard 系数.

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \tag{3}$$

2) 莱文斯坦距离: 莱文斯坦距离是编辑距离的一种, 它指两个字符串之间, 由一个转变成另一个所需的最少编辑操作次数. 其中编辑操作包括将一个字符替换成另一个字符、插入一个字符或者删除一个字符, 该距离也可以衡量两个字符串的差异程度.

3) FASPell[26]: FASPell 是 2019 年由爱奇艺发表的错别字检测纠错算法, 它提出了一种新的范式去噪自动编码 (DAE) 和解码器. 其中 DAE 利用预训练模型 BERT 减少监督学习中所需文本拼写检查数据量, 解码器为把握度-字符相似度解码器 (CSD), 在字形上采用 Unicode 标准的 IDS 表征, 可以准确描述汉字的各个笔画和它们的布局形式. 该模型可完成对任何一种中文文本的拼写检查, 包括 OCR 识别结果.

通过使用上述 3 种相似度计算方式, 可以分别得出 3 种不同的实体相似性排序列表. 为了综合考虑每种相似度计算方式, 本文采用一种集成排序算法 (ranking ensemble), 最终可以获得相似实体的最终排名, 如式 (4)、式 (5) 所示:

$$r_t^i=\sum_{a_j\in A}I(score(a_i|sim_t)\leqslant score(a_j|sim_t)) \tag{4}$$

$$final\_r_i=\sum_{t-1}^{T}\frac{1}{r_t^i} \tag{5}$$

其中, $I(\cdot)$为指示函数, r_t^i为使用第t个相似度计算方式sim_t计算结果中候选实体a_i的排名, 每个候选实体的最终排名可以通过使用每个相似度排序列表中排名的倒数求和得到, 如式 (5) 所示.

2.3 人身保险属性抽取

对于人身保险产品的属性抽取, 由于人身保险条款数据自身的特点, 其通常使用一段文字描述某个属性值, 如投保范围对应的属性值为“凡年满 18 周岁 (见释义 8.1), 具有完全民事行动能力且对被保险人具有保险利益的人, 均可以作为投保人向本公司投保”, 不适用常规的序列标注任务. 因此, 本文根据本体构建制定的保险实体属性, 将人身保险属性抽取任务转变为人身保险属性槽填充任务. 本文将该任务划分为两个阶段: 属性标签抽取和属性填充. 首先, 本文根据文本特征使用 Bootstrapping 算法抽取出保险产品的候选属性标签, 进而可以找到每个属性标签对应的属性段落, 然后使用模型对候选属性段落分类, 将其填充到对应属性槽中.

2.3.1 属性标签抽取

通过对人身保险条款数据的观察, 本文发现保险条款中的标题数据通常对应一个条款属性, 如“第一条 合同的构成”, 属性标签的上下文符合一定的特征, 如通常会出现在标题序号后的特点等, 因此, 本文将这些标题视为属性标签. 对于属性标签的抽取, 本文使用 Bootstrapping 算法获取属性标签候选集, 如算法 1 所示.

首先使用给定种子规则供 Bootstrapping 算法冷启动, 通过种子规则抽取出文本的部分属性标签; 然后使用抽取到的属性标签回到文本中挖掘新的规则, 并通过已有的候选属性标签对规则打分, 保留大于一定阈值的规则; 最后使用新的规则继续抽取候选属性标签. 特别的是, 如果算法不能通过新规则抽取出新的候选属性标签, 或者不能通过候选属性挖掘出新的规则, 则算法结束并返回抽取到的候选属性标签. 由于在使用 Bootstrapping 算法抽取候选属性标签和扩展规则时, 每次迭代需要遍历新规则集合以及候选属性标签集合,

因此算法的时间复杂度是O(n^2).

算法 1. 基于 Bootstrapping 的属性标签抽取算法

输入: 人工制定规则种子集合 R, 待抽取文本内容 *content*, 规则阈值ε
输出: 候选属性标签集合 A

```
for r in R
    A'+=find(content, s)
    if len(A')==0
        return A
    A.append(A')
    R'=generate_rule(content, A')
    for r' in R':
        score=get_score(r', A)
        if score>ε
            R_new.append(r')
    if len(R_new)==0
        return A
```

2.3.2 属性填充

通过算法 1, 本文可以从保险条款文本中抽取出候选属性标签. 由于不同人身保险产品条款文件的属性标签由每个公司的员工制定, 其名称并不统一, 因此不能直接通过抽取到的候选属性标签将属性映射到对应属性槽中. 由于属性槽的个数是一定的, 因此本文将属性填充任务转换为分类任务. 首先通过抽取到的候选属性标签将文本划分成 n 个段落, 这 n 个段落则对应每个候选属性标签的属性值, 然后抽取特征并使用模型对这些属性值分类, 将其填充到对应属性槽中. 特别的是, 本文将某些段落不属于任何一个属性槽的噪音属性值预测为“None”类型. 本文所使用的模型结构如图 4 所示, 主要包含两部分.

1) 属性标签信息: 人身保险中的属性标签与本文定义的属性名称类似, 可以通过第 2.2.3 节中实体对齐的方式进行映射. 但是该方法对于一词多义的情形不能很好地处理, 如“职业或工种变更”和“工作变化”等. 因此, 除了使用实体对齐的算法, 本文还利用模型学习标签语义, 首先将属性标签转换成词向量, 然后使用 CNN[27] 获取该部分信息.

2) 段落信息: 文本的属性段落主要描述对应属性的属性内容, 其中包含大量的文本信息, 该内容的描述对属性段落填充到对应属性槽发挥着一定程度的作用. 对于文本信息的获取, 现今有许多预训练模型通过大量的语料库训练, 可以很好地理解文本并提取语义信息, 因此本文使用 BERT 来抽取出段落内容的特征.

最后, 本文将两部分特征拼接起来输入到分类器中预测出最终属性类型, 损失函数如下:

$$L = -\sum_{i=1}^{K} y_i \log(p_i) \tag{6}$$

其中, y_i为第i个数据的真实标签, p_i为模型预测为该标签的概率.

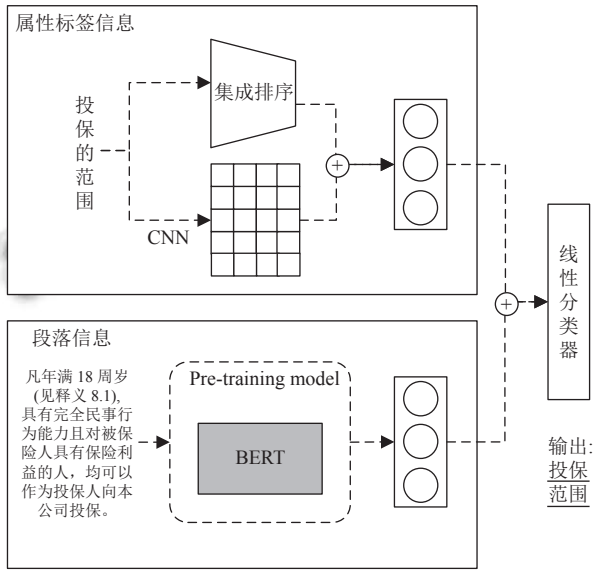

图 4 属性段落分类模型结构

2.4 推荐指标

当用户检索保险产品时, 系统需要根据保险产品的特点将其条款属性依次展现. 对于不同类型或不同公司的人身保险条款, 其包含的条款属性不尽相同, 如对于人寿保险, 其条款内容普遍包含寿命等信息, 疾病保险则会注重疾病条款属性. 因此为了描述不同条款属性对不同类型人身保险的重要程度, 本文参考 TF-IDF (term frequency-inverse document frequency) 指标, 设计了人身保险条款属性相关性指标 *CF-IIF* (clause frequency-inverse insurance frequency).

*CF*指某一个条款属性在一个确定的保险类型中出现次数的占比, 它可以描述该条款属性对这个保险类型的重要程度, 其计算公式如式 (7):

$$CF_{ij} = \frac{n_{ij}}{\sum_{k} n_{kj}} \tag{7}$$

其中, CF_{ij}表示第i个条款属性对第j种类型保险的占比, n_{ij}表示第i个条款属性与第j种类型保险的共现次数.

*IIF*可以描述某一个条款属性区分不同类型保险

的能力, 如果大部分保险产品都含有这个条款属性, 它将会有一个低的保险相关性, 其计算公式如式 (8):

$$IIF_i = \log \frac{\sum |I_k|}{|\{j : t_i \in I_j\}|} \tag{8}$$

其中, $\sum |I_k|$表示保险文件数, 分子为第i个条款属性在第j种类型保险文件中出现次数. 最终, 使用上述两个指标的乘积即可计算出$CF\text{-}IIF_{ij}$:

$$CF\text{-}IIF_{ij} = CF_{ij} \times IIF_i \tag{9}$$

3 实验结果

3.1 实验设置

3.1.1 实验数据

本文实验所使用的数据集来自作者收集的 14 124 份人身保险条款合同 (脱敏后). 本文的实验包含两部分: 人身保险实体抽取和人身保险属性抽取, 因此针对这两部分实验, 本文分别构建了对应的数据集对模型进行训练, 数据集介绍如下.

对于人身保险实体抽取任务, 根据第 2 节数据来源介绍, 本文通过反向标注得到训练数据用于对命名实体识别序列模型训练. 最后, 本文使用 60% 数据作为训练集, 30% 数据作为测试集, 10% 数据作为验证集, 数据统计信息如表 2 所示.

表 2 实体识别数据统计

实体	训练集	测试集	验证集	总计
保险公司	12 771	6 385	2 130	21 286
保险产品	9 019	4 509	1 504	15 032
总计	21 790	10 894	3 634	36 318

对于人身保险属性抽取任务, 由于缺少已标注的数据训练属性段落分类模型, 本文随机选择 1 000 个人身保险条款文件作为数据集, 然后对其进行标注. 本文首先使用短文本相似度计算方法和集成排序算法找出文件中每个属性标签最相似的属性名称列表, 然后保存其中属性标签与属性定义中相似度大于 0.8 的数据. 对于差异较大的数据, 本文对其进行人工修正, 最终收集到的数据信息如表 3 所示.

表 3 属性段落分类数据

文件数	段落数	训练集	测试集	验证集
1 000	25 346	15 207	7 603	2 437

3.1.2 评价指标

对于实体识别任务以及属性抽取端到端的结果评估, 本文采用精确率 (precision, P)、召回率 (recall, R) 和 $F1$ 分数作为算法的评估指标, 计算公式如式 (10)–式 (12):

$$P = \frac{TP}{TP+FP} \times 100\% \tag{10}$$

$$R = \frac{TP}{TP+FN} \times 100\% \tag{11}$$

$$F1 = 2 \times \frac{P \times R}{P+R} \times 100\% \tag{12}$$

对于属性段落分类, 本文采用准确率 (accuracy, acc) 来进行评估, 计算公式如式 (13):

$$acc = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \tag{13}$$

3.1.3 实验环境和参数设置

本文的实验均在 Python 3.7.11 环境中运行, 服务器的操作系统是 Ubuntu 18.04.3LTS, 环境配置如表 4 所示.

表 4 实验环境

实验环境	参数
CPU	Intel(R) Xeon(R) Gold 5215 CPU @ 2.50 GHz
GPU	NVIDIA TITAN RTX
内存	64 GB
Torch	1.11.0
Transformers	4.9.0
PyTorch_crf	0.7.2

对于实体识别的参数设置, 本文使用 Transformers 库中 BERT-Base 预训练模型作为词向量表示层, 其输出为 768 维的词向量, IDCNN 特征层数都为 64, LSTM 的大小为 128, 训练过程使用 Adam 优化器, 学习率设置为 0.000 03; 对于属性段落分类模型, 本文使用基于中文维基百科训练的 300 维词向量[28] 来对文本进行初始化输入到 CNN 中, 训练过程使用 Adam 优化器, 学习率设置为 0.000 01; Bootstrapping 算法中ε设置为 0.5.

3.2 实验结果及分析

3.2.1 实体识别结果

本文从多维度比较本文构建的序列模型 BERT-IDCNN-BiLSTM-CRF 和现有常用的 NER 模型, 实验结果如表 5 和表 6 所示. 其中表 5 展示了使用 BIO 标注方案下不同模型训练的结果, 表 6 展示了使用 BIOES 标注方案下不同模型训练的结果. 从这两个表中, 可以看出 BERT-IDCNN-BiLSTM-CRF 模型的最终效果最

好, 在 BIO 标注方案中 *F*1 分数为 93.91%, 在 BIOES 标注方案中 *F*1 分数为 96.07%.

表 5 不同模型基于 BIO 标注方案的实体识别结果 (%)

模型	实体类别	*P*	*R*	*F*1
BiLSTM-CRF	保险公司	86.17	96.56	91.07
	保险产品	67.21	80.01	73.05
	总计	78.99	90.29	84.26
IDCNN-CRF	保险公司	88.76	98.23	93.26
	保险产品	80.84	82.16	81.49
	总计	85.76	92.23	88.87
IDCNN-CRF2	保险公司	88.92	98.45	93.44
	保险产品	83.71	85.62	84.65
	总计	90.74	93.59	92.14
BERT-BiLSTM-CRF	保险公司	**97.19**	**98.44**	**97.81**
	保险产品	86.16	87.45	86.80
	总计	92.94	94.20	93.57
BERT-IDCNN-BiLSTM-CRF	保险公司	95.28	95.18	95.23
	保险产品	**90.32**	**93.44**	**91.85**
	总计	**93.33**	**94.51**	**93.91**

表 6 不同模型基于 BIOES 标注方案的实体识别情况 (%)

模型	实体类别	*P*	*R*	*F*1
BiLSTM-CRF	保险公司	85.30	95.46	90.09
	保险产品	78.25	89.52	83.50
	总计	82.63	93.21	87.60
IDCNN-CRF	保险公司	90.74	97.53	94.01
	保险产品	83.74	92.17	87.75
	总计	88.09	95.50	91.64
IDCNN-CRF2	保险公司	93.38	95.78	94.56
	保险产品	85.13	93.76	89.23
	总计	90.26	95.02	92.27
BERT-BiLSTM-CRF	保险公司	**99.18**	96.68	**97.92**
	保险产品	87.81	94.66	91.11
	总计	**94.53**	95.90	95.21
BERT-IDCNN-BiLSTM-CRF	保险公司	93.67	**98.57**	96.06
	保险产品	**94.53**	**97.68**	**96.08**
	总计	94.00	**98.23**	**96.07**

针对表 5 和表 6 的内部数据进行对比可以发现, 本文构建的模型 BERT-IDCNN-BiLSTM-CRF 在大部分指标上高于现有模型 BERT-BiLSTM-CRF, 在实体“保险公司”的类别中性能略低于 BERT-BiLSTM-CRF 模型, 但是在实体“保险产品”的类别中本文模型性能则会高出其很多. 本文通过对“保险公司”和“保险产品”实体数据的观察发现, 许多保险产品的名称会包含保险公司的名字, 模型在识别的过程中会容易将保险产品中的一些字符识别为“保险公司”实体, 因此该实体类别的识别性能普遍低于“保险公司”实体类别. 本文在 BiLSTM 模型层前插入 IDCNN 模型, 该模型可以先提取文本的全局特征, 然后供 BiLSTM 模型学习, 因此本文模型可以更好地学习到“保险产品”实体信息, 表现效果最好.

对比表 5 和表 6 实验结果, 可以发现基于 BIOES 标注方案训练的模型效果普遍高于基于 BIO 标注方案训练的模型. BIOES 额外提供实体 End 的信息, 并给出单个词汇的 S-tag, 提供更多的信息, 因此该标注方案在本文面临场景中表现效果更优.

3.2.2 属性抽取结果

属性抽取算法主要分为两个阶段: 首先, 使用 Bootstrapping 算法抽取出候选属性标签以及对应的属性段落; 然后再使用分类模型对属性段落分类填充到对应属性槽中. 算法主要区别为预测段落类型的分类算法, 本文选择以下几种方法进行对比.

1) 集成排序: 使用本文介绍的集成排序算法, 可以获得每个候选属性标签对应的属性名称相似度排序列表, 取 top-1 作为预测结果.

2) LSTM: LSTM 是常用文本分类模型, 该模型以学习文本内容提取文本信息, 完成文本分类任务.

3) BERT: BERT 是今年来最成功预训练模型之一, 该模型的出现刷新了自然语言处理中 11 个基本任务的分数, 包括文本分类任务.

4) RoBERTa[29]: RoBERTa 由 Facebook 和华盛顿大学共同发表, 该模型通过对 BERT 训练过程的改进, 如采用动态 Mask 以及不适用 NSP 训练方法, 达到很好的效果.

5) DistilBERT[30]: DistilBERT 采用知识蒸馏的方法, 使得该模型的参数量比 BERT 少 40%, 速度比 BERT 快 60% 并保留了 BERT97% 的语言理解能力.

表 7 展示了属性抽取结果, 其中 *acc* 表示每个模型在属性分类中的性能. 由于每个人身保险文件都对应多个属性, 因此每个人身保险文件都会计算出 *P*、*R* 和 *F*1 分数, 本文计算这些分数的平均值作为每个算法的最终结果. 从表中可以看出本文使用的 CNN-BERT 分别学习属性标签和属性段落的信息预测的结果最好, avg_F1 分数为 92.36%. 集成排序在属性抽取中 avg_R 的分数最高, 该方法根据相似度计算方式得出与候选属性标签最相似的属性名称, 一定会给出一个属性标签对应的属性名称, 因此其会有较高的 avg_R 分数, 但是 *acc* 值会比较低. 对于其他预测方法, 本文加入特殊

标签“None”来指明某些属性标签和属性段落不对应属性槽中的任何一个, 因此可能会对一些属性误判, 将其预测为“None”类型造成 avg_R 略低.

表 7 属性段落分类及属性抽取结果 (%)

算法	*acc*	avg_P	avg_R	avg_F1
集成排序	79.63	83.72	**98.30**	89.38
LSTM	81.92	86.93	98.21	90.86
BERT	83.91	87.70	98.11	91.97
RoBERTa	83.56	87.24	98.01	91.24
DistilBERT	84.62	88.16	97.74	92.03
本文方法	**85.99**	**88.81**	97.62	**92.36**

表 8 展示了在不同阈值下挖掘新规则的 Bootstrapping 抽取算法结果. 从表中可以看出, 随着阈值 ε的增加, avg_P 会逐渐增高, 当其等于 1 时, 即不会扩展新的规则, avg_P 最高, 但是召回率会很低. 在本文方法中, 需要尽可能抽取出候选属性, 然后在第 2 阶段中使用分类模型筛选掉不符合的属性, 因此对于召回率应越高越好. 本文发现阈值从 0.75–0.5 中 avg_R 增长较慢, 因此最终选择阈值为 0.5.

表 8 不同阈值下 Bootstrapping 抽取结果

阈值ε	avg_P (%)	avg_R (%)	avg_F1 (%)
1	**99.85**	60.24	80.34
0.75	75.04	98.23	**82.01**
0.5	73.39	**99.20**	81.78

4 人身保险知识图谱原型系统

4.1 系统设计

本文根据构建的 LIKG 设计了人身保险知识图谱应用原型系统. 本文前端采用 JavaScript 和 Echart 来完成图表展示; 后端采用 Django 框架, 使用 RESTful 风格 API 接收前端请求并生成数据返回前端; 本文使用 Neo4j 图数据库存储抽取出的实体和属性, 对于本体构建过程中实体、关系和属性的定义, 本文将其存储到 MySQL 关系数据库中, 供上层应用使用. 本文系统的整体架构如图 5 所示.

本文人身保险知识图谱原型系统所提供的功能如下.

1) 知识获取: 输入人身保险文件, 抽取其实体和属性并存入数据库中, 将抽取结果以 JSON 格式以及图表形式展现给用户.

2) 知识查询: 知识查询主要分为两部分: 属性推荐和属性查询. 对于属性推荐, 用户查询某个保险产品的所有属性, 根据 *CF-IIF* 指标对属性排序并展现给用户; 对于属性查询, 用户明确查询某个保险产品的具体属性, 系统构建对应 Cypher 语句查询 Neo4j 数据库, 将查询结果展现给用户.

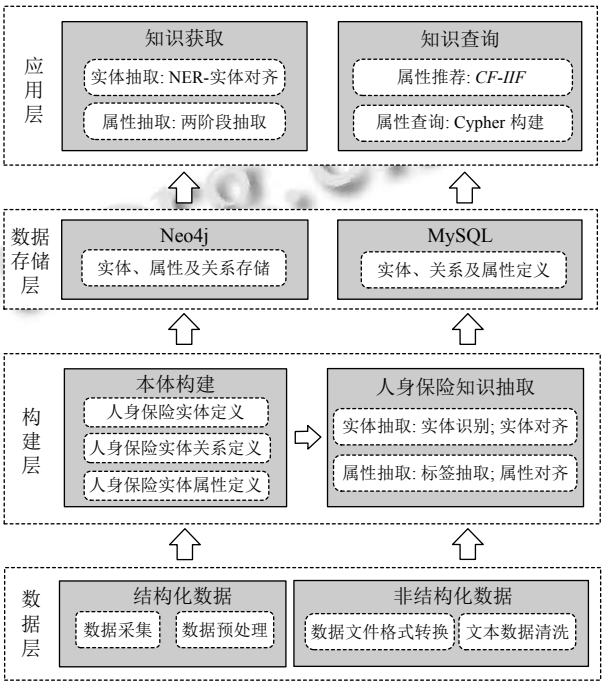

图 5 系统整体架构图

4.2 系统功能展示

4.2.1 知识获取

本系统支持用户上传新的人身保险条款文件, 系统则利用本文提到的实体抽取算法和属性抽取算法, 从上传文件中抽取出其包含保险公司实体、保险产品实体以及保险产品属性, 并可视化地展现给用户. 如图 6 所示, 用户从本地选择上传的人身保险产品文件并提交, 系统抽取出其实体和属性并将结果返回给用户. 其中“抽取结果”则为系统从上传的人身保险文件中抽取出的实体信息和属性信息, 并以 JSON 数据格式展示该部分内容. “关系图”则为系统将抽取的实体结果和属性结果存入 Neo4j, 并利用 Echart 可视化技术展示给用户, “关系图”的具体展示内容可见图 7.

4.2.2 知识查询

本系统向用户提供知识查询功能, 可以帮助用户快速获取 LIKG 中存在的知识, 得到用户想要了解的保险产品的信息. 其主要分为两部分: 人身保险属性推

荐和属性查询.

对于人身保险属性推荐, 本系统向用户提供查询某个保险产品实体的所有属性内容, 并使用 *CF-IIF* 指标对其排序, 按照指标的倒序结果展示给用户. 如图 8 所示, 用户在输入框中输入想要查询的保险产品名称, 系统则会在 Neo4j 中查询其所有属性返回给用户. 图 8 则查询了一个和疾病保险有关的保险产品, 因此该产品的“疾病”属性的 *CF-IIF* 的指标会比较高, 因此排在首位, 也符合购买疾病保险产品的用户对该类保险条款属性的关注重点.

图 6 知识抽取功能展示

除了属性推荐列表展示外, 系统还会将保险产品的属性图展示给用户, 如图 7 所示. 属性图展示主要实现了知识图谱的图形可视化, 由于每个节点中包含的内容是长文本, 属性图的节点无法将节点的内容展示完整, 因此在展示过程中只展示每个节点的前 3 个文字, 用户可以点击每个节点查看具体内容.

对于人身保险属性查询, 根据本体构建内容, 每个保险产品的属性是有一定的范围的, 因此本系统给用户展示了一个下拉框, 供用户选择想要查询保险产品的具体属性, 系统则根据保险产品名称和属性名称构建 Cypher 语句, 查询对应属性值并在前端展现给用户, 如图 9 所示.

图 7 属性图展示

图 8 属性推荐功能展示

5 结论与展望

本文根据给定人身保险条款数据, 制定了一套完整的知识图谱构建流程, 包括本体构建、知识抽取、知识存储以及知识图谱应用, 并设计了原型系统以展

示该知识图谱的应用价值. 本文在实体识别任务中构建了 BERT-IDCNN-BiLSTM-CRF 模型, 该模型在本应用场景中表现的效果最好; 在属性抽取任务中, 根据人身保险数据的特点, 制定了两阶段抽取流程, 将抽取任务转变为属性槽填充任务, 最终使用分类模型完成属性填充; 在系统功能展示模块中, 除了可视化展示以及新增数据的知识抽取任务, 本文还制定了 *CF-IIF* 指标, 用于向用户推荐不同类别保险产品的属性内容, 并使用户快速了解保险产品的重要内容. 由于知识抽取涉及非结构数据, 在训练模型的过程中, 训练数据的质量是至关重要的, 因此在后续任务中会关注如何获取高质量数据训练模型, 或者考虑让模型从少样本或低质量数据中训练的更好. 在系统应用层面, 则考虑加入知识图谱问答功能, 自动化解析输入文本语义, 完成智能问答.

图 9 属性查询功能展示

参考文献

1 邹瑜, 顾明. 法学大辞典. 北京: 中国政法大学出版社, 1991.

2 Zhu XR, Li ZX, Wang XD, *et al*. Multi-modal knowledge graph construction and application: A survey. arXiv:2202.05786, 2022.

3 Ali W, Saleem M, Yao B, *et al*. A survey of RDF stores & SPARQL engines for querying knowledge graphs. The VLDB Journal, 2022, 31(3): 1–26. [doi: 10.1007/s00778-021-00711-3]

4 Li LF, Wang P, Yan J, *et al*. Real-world data medical knowledge graph: Construction and applications. Artificial Intelligence in Medicine, 2020, 103: 101817. [doi: 10.1016/j.artmed.2020.101817]

5 罗明. 教育测评知识图谱的构建及其表示学习. 计算机系统应用, 2019, 28(7): 26–34. [doi: 10.15888/j.cnki.csa.006977]

6 InsuranceAI. 保险大脑. http://39.96.206.144/. (2016-05-16).

7 中国疾病保险知识图谱. 保险知识图谱. https://ai.zhibao-tech.com/kg. (2022-03-10).

8 Bollacker K, Evans C, Paritosh P, *et al*. Freebase: A collaboratively created graph database for structuring human knowledge. Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver: ACM, 2008. 1247–1250.

9 Suchanek FM, Kasneci G, Weikum G. YAGO: A core of semantic knowledge. Proceedings of the 16th international conference on World Wide Web. Banff: ACM, 2007. 697–706.

10 Auer S, Bizer C, Kobilarov G, *et al*. DBpedia: A nucleus for a Web of open data. Proceedings of the 6th International Semantic Web Conference on the Semantic Web. Busan: Springer, 2007. 722–735.

11 王智悦, 于清, 王楠, 等. 基于知识图谱的智能问答研究综述. 计算机工程与应用, 2020, 56(23): 1–11. [doi: 10.3778/j.issn.1002-8331.2004-0370]

12 Bouraga S, Jureta I, Faulkner S, *et al*. Knowledge-based recommendation systems: A survey. International Journal of Intelligent Information Technologies, 2014, 10(2): 1–19. [doi: 10.4018/ijiit.2014040101]

13 王建勋, 华丽, 邓世超, 等. 基于 CiteSpace 国内干旱遥感监测的知识图谱分析. 干旱区地理, 2019, 42(1): 154–161.

14 江双五, 刘惠兰, 温华洋, 等. 气象记录档案知识图谱构建. 计算机系统应用, 2022, 31(1): 73–82. [doi: 10.15888/j.cnki.csa.008315]

15 Zhao MX, Wang H, Guo J, *et al*. Construction of an industrial knowledge graph for unstructured Chinese text learning. Applied Sciences, 2019, 9(13): 2720. [doi: 10.3390/app9132720]

16 Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv:1508.01991, 2015.

17 Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 2670–2680.

18 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: NAACL, 2018. 4171–4186.

19 Dai ZJ, Wang XT, Ni P, *et al*. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. Proceedings of the 2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI). Suzhou: IEEE, 2019. 1–5.

20 张凯伦. 中文人物属性抽取技术的研究与实现 [硕士学位论文]. 北京: 北京邮电大学, 2016.

21 Mooney CZ, Duval RD. Bootstrapping: A Nonparametric Approach to Statistical Inference. Thousand Oaks: SAGE Publications Inc., 1993.

22 李勇, 张志刚. 领域本体构建方法研究. 计算机工程与科学, 2008, 30(5): 129–131. [doi: 10.3969/j.issn.1007-130X.2008.05.039]

23 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

24 Ma YJ, Yu DH, Wu T, *et al*. PaddlePaddle: An open-source deep learning platform from industrial practice. Frontiers of Data and Computing, 2019, 1(1): 105–115.

25 Memon J, Sami M, Khan RA, *et al*. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE Access, 2020, 8: 142642 – 142668. [doi: 10.1109/ACCESS.2020.3012542]

26 Hong YZ, Yu XG, He N, *et al*. FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm. Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). Hong Kong: EMNLP, 2019. 160–169.

27 Chen YH. Convolutional neural network for sentence classification [Master's thesis]. Waterloo: University of Waterloo, 2015.

28 Li S, Zhao Z, Hu RF, *et al*. Analogical reasoning on Chinese morphological and semantic relations. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 138–143.

29 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

30 Sanh V, Debut L, Chaumond J, *et al*. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv:1910.01108, 2019.

(校对责编: 牛欣悦)