Statistics and Application 统计学与应用, 2023, 12(5), 1464-1476
Published Online October 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.125150

Hans汉斯

调查问卷中开放题文本答案编码方法研究综述

刘培莹，安建业

天津商业大学理学院，天津

收稿日期：2023年9月27日；录用日期：2023年10月24日；发布日期：2023年10月31日

摘　要

调查问卷作为联系调查者与被调查者之间的“纽带”，是采集数据、发现因素间内在联系和规律的有力工具。对其中的开放题收集的答案文本数据进行编码，有助于了解文本答案蕴含的固有特征，便于后期开展统计分析，是开放题研究的热点问题。为此，从人工编码、半自动编码及自动编码三个方面对开放题文本答案的编码方法研究进行了归纳、总结与展望，为进一步开展相关研究奠定良好的基础。结果表明：关于开放题文本答案的编码方法研究，经历了从最初的完全人工编码到“机器 + 人工”的半自动编码，逐步发展到目前“基于人工智能”的自动编码探索三个阶段；虽然编码效率得到了极大的提升，但是由于缺乏普适性的编码方法，因而不同领域开放题文本答案编码的准确性存在较大差异，提高编码的普适性、准确性仍是未来研究的重点。

关键词

调查问卷，开放题，自动编码，研究综述

A Review of Text Answer Coding Methods for Open-Ended Questions in Questionnaires

Peiying Liu, Jianye An

School of Science, Tianjin University of Commerce, Tianjin

Received: Sep. 27th, 2023; accepted: Oct. 24th, 2023; published: Oct. 31st, 2023

Abstract

As the “link” between the investigator and the respondent, questionnaire is a powerful tool to collect data and find the internal relations and rules among factors. Encoding the answer text data collected by the open-ended questions is helpful to understand the inherent characteristics of the text answers, which is convenient for later statistical analysis, and is a hot issue in the study of open questions. Therefore, this paper summarizes, concludes and prospects the research on the

文章引用: 刘培莹, 安建业. 调查问卷中开放题文本答案编码方法研究综述[J]. 统计学与应用, 2023, 12(5): 1464-1476.
DOI: 10.12677/sa.2023.125150

encoding methods of open-ended question text answers from three aspects: manual encoding, semi-automatic encoding and automatic encoding, which lays a good foundation for further related research. The results show that: As for the research on the encoding method of open-ended question text answer, it has experienced three stages from the initial completely manual encoding to the semi-automatic encoding of "machine + manual", and gradually developed to the current automatic encoding exploration of "artificial intelligence". Although the coding efficiency has been greatly improved, due to the lack of universal coding methods, the accuracy of open-ended question text answer encoding in different fields is quite different. Improving the universality and accuracy of coding is still the focus of future research.

Keywords

Questionnaire, Open-Ended Question, Automatic Coding, Review

1. 引言

调查就是为了获得被调查者的真实意见，使调查者的决策依据更加充分，对未来的预测更加准确，被广泛应用于教育、农业、医学等领域[1]-[9]。在各种调查方法中，问卷调查是最常用、最重要的方法，已成为各行各业开展诸如评价学习、了解市场、征集民意等活动的主要方法。

调查问卷作为联系调查者与被调查者之间的“纽带”，是采集数据、发现因素间内在联系和规律的有力工具[10]。通常，调查问卷设计的题目包括封闭题、半开放题和开放题三种类型。封闭题设计了题目和相应的选项以及指定回答规则，由被调查者选择合适的答案；半开放题是封闭题选项的扩充，可以发表自己的观点；而开放题则按照研究目的只设置题目，不提供任何答案选项，完全由被访者自由作答。

关于调查问卷中开放题设置的必要性，不同学者拥有截然不同的观点。例如，潘绥铭[11]认为问卷中“开放题”的存在是种失误，不应该设置任何一种关于原因类的开放题；而肖富群[12]则认为问卷中开放题是很有必要的，他指出如果题目中有大量备选答案或者无法给出明确的备选答案，在问卷设计过程中不便提供，就可以设置成开放题，这样在一定程度上就能够避免由于选项对用户的限制而造成的回答偏见，收集到更加翔实的数据[13] [14]。

调查问卷中开放题文本答案虽然包含了情感倾向、商品喜好、社会热点看法等丰富的有价值信息，但是由于其具有描述信息弱、稀疏性、表达不规范等短文本数据的特点[15]，因而难以直接对其进行频数、相关性、可视化等常规统计分析。再加上开放题分析对工作人员的技术要求和处理成本较高，尤其是涉及某些特定领域的调查时，需要研究人员有较强的专业知识，所以导致调查问卷中大多数题目都采用封闭题，对于开放题的研究与应用明显不足[16] [17]。由此可知，通过对开放题文本答案进行编码，将非结构化的文本数据转换成易被机器读懂的结构化代码数据形式，对于调查研究具有重要的现实意义。

对开放题文本答案进行编码，实质上就是建立开放题文本答案数据集与码表集之间的一个映射：

已知 n 为样本容量，调查问卷中某一开放题文本答案构成的集合为 D_T，对应的码表集为 Rc，如果对于 D_T 中任意的文本答案 $x_i\left(i=1,2,\cdots,n\right)$，存在与 Rc 中唯一编码 $y_i \in R_C\left(i=1,2,\cdots,n\right)$ 之间的对应关系 f，建立的编码模型如式(1)所示：

$$y_i = f\left(x_i\right), i=1,2,\cdots,n \tag{1}$$

当编码结束后，通过对开放题文本答案编码结果与其他封闭题的一致性检验，还可以进一步判断每份调查问卷所采集到数据的有效性，提高样本数据的质量。

最初，调查研究中收集数据的方式是通过实地发放问卷进行的，这种传统发放方式所收集的开放题文本答案数量较小，通过人工方法就可以对答案进行编码处理了[18]。后来，信息技术的不断发展，发放平台逐渐向网络平台转移，问卷发放与回收效率不断提高，这为广泛运用问卷开展调查研究提供了广阔的空间[19]。

随着文本计算能力的不断加强，文本挖掘技术日益成熟，调查问卷中开放性题目的设置比例逐步增加，开放题文本答案蕴含的信息量更加丰富，此时对于开放题答案编码仍然采用人工的方式，其成本高、效率低等劣势表现得非常明显。如果要在短时间内高效地对问卷的文本数据进行处理和分析，那么开放题文本答案自动编码的重要性不言而喻[20]。近年来，语音识别技术飞速发展，语音成为人机交互的主要途径之一[21]，采用语音识别技术开展问卷调查成为数据收集的新模式，开放题文本答案半自动编码、自动编码的应用场景不断扩大。为此，许多学者同时利用网络语音调查和实地问卷调查两种形式，采取定量分析与定性描述相结合的方式开展调查研究[22] [23] [24]，并通过设置更丰富的开放性题目，收集更大量的文本答案数据，为后期统计分析奠定良好的基础。

目前，开放题编码问题受到越来越多学者的关注，如何实现文本答案准确、快速编码已成为研究的热点，然而这方面的综述文章很少，急需对开放题文本答案编码的相关研究进行系统总结。为此，下面从开放题文本答案编码经历的人工编码、半自动编码、自动编码三个阶段入手，梳理了编码方法的相关研究成果，比较分析了不同编码方式的特点与研究现状，归纳总结了分词库的建设情况，展望了未来的研究方向。

2. 人工编码

人工编码是开放题文本答案编码的基础，而要准确地进行人工编码，就要了解开放题的题型设计以及不同类型开放题的具体特征，以此选取适宜的编码策略。

关于开放题题型设置方面的研究，Popping R [25]曾在文章中指出开放题主要是为了调查“谁”“什么”“何时”“何地”以及“为什么”，因此从调查目的出发可以将开放题划分客观题、主观题和综合题，具体情况如表 1 所示：

Table 1. Basic types of common open questions
表 1. 常见开放题的基本类型

类型	发问词	题目举例	目的
客观题	谁(who)	您家有谁使用该款手机？	询问客观事实，发现事实，收集有价值信息
	什么(what)	您使用什么牌子的手机？	
	何时(when)	您在何时开始使用该款手机？	
	何地(where)	您在何地了解到该款手机的？	
主观题	为什么(why)	您为什么喜欢某牌子的手机？	征求客户的意见和态度
综合题	-	您觉得以下谁适合当该款手机的代言人？为什么？	综合考虑客观事实与客户观点

由表 1 可知，不同研究目的需要设置的开放题题目类型不同，因而收集到的文本答案具有很大差异。在遵循统一性、合理性、完备性、唯一性和可读性这一开放题文本答案编码原则的前提下，不同类型的开放题文本答案所采用的编码方式也各不相同[26] [27]。针对开放题中诸如品牌、城市等客观题型，其文

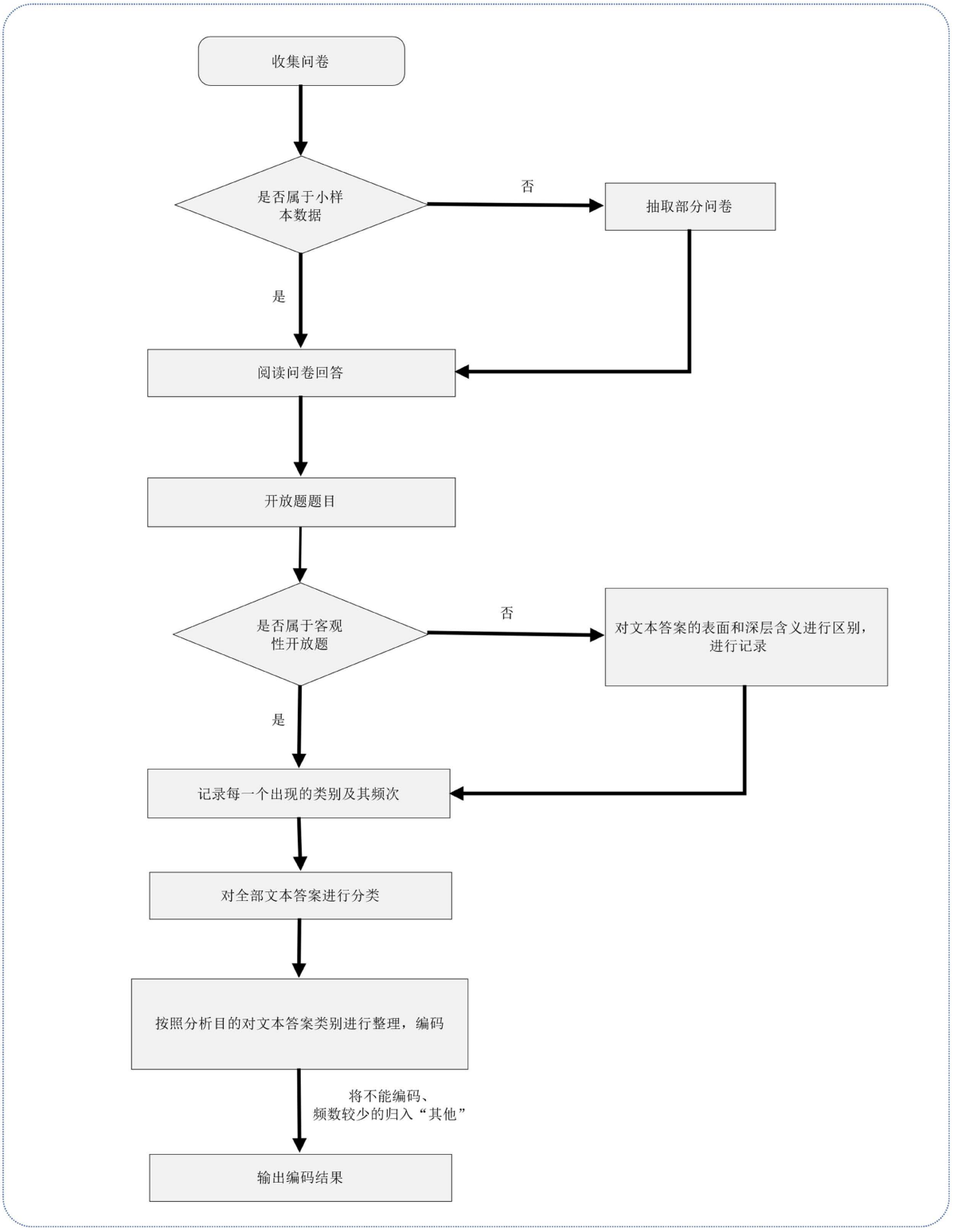

Figure 1. Manual coding flowchart
图 1. 人工编码流程图

本答案主题比较清晰，通常可以以主题的方式直接进行编码；针对开放题中诸如满意度、口味等主

观题型，通常需要在原有基本码表的基础上再应用和制定新的码表[24]；而针对回忆、心理描述等综合类型的开放题，因其文本答案主题分散、含义复杂等原因，在人工编码中通常以问卷号的方式直接对文本答案进行编码。

早在 2012 年，任莉颖[28]认为国内调查中开放题文本答案编码当时是以人工编码为主，需要编码员具有一定的专业性。文章一方面按照先大类、后细类的原则对 CFPS 数据提出了四级编码，另一方面编码员采用双向独立验证判定的方式对数据集进行了集中编码，并通过对两种编码方式的比较来验证编码的准确性。由于此方法完全由人工编码，耗费时间长，编码效率不高，因此，许多学者的研究逐步转移到如何提高开放题文本答案编码的准确性与编码效率。

Popping R [29]以荷兰选举中开放题文本答案为依托，从受访者的角度制定了相应的编码规则，并根据受访者类别对开放题文本答案进行了编码[30]。以提高编码的准确性；Zhoushanyue He [31]等人提出了基于双编码的开放题文本答案人工编码方法，提高了编码的可靠性。但是在实际应用中，由于预算成本等原因，大多数开放题文本答案的人工编码还是采用了单编码方式。

总之，目前开放题文本答案编码还是以人工编码为主，是在收集完开放题文本答案的基础上，由专业编码人员根据个人理解对答案进行手动编码，其流程如图 1 所示。

开放题文本答案编码存在以下两个问题：

一是不同领域、不同类型的开放题之间有较大的区别，编码工作耗费的时间较长，成本高；

二是由于每个编码人员对同一事物认知有偏差，缺乏统一的标准，编码结果有较强的主观性。

为了克服人工编码中的不足，国内外许多学者将机器学习、统计学习应用于开放题文本答案的编码过程中，人工编码逐渐向半自动或自动编码方法过渡。

3. 半自动编码

半自动编码以计算机辅助人工编码的方式对开放题文本答案进行编码，首先由专业的编码人员对部分采集的文本答案人工标注为预先设置的编码，然后将其作为训练集，利用机器学习方法进行模型训练，最终利用训练好的模型对其他所有文本答案进行编码，提高了编码效率。

李煜[32]在爱情观调查问卷中设置了开放题，归纳出 16 类对爱情的表述，之后通过聚类分析聚成爱情观的 9 大类别，从定性与定量统计分析两个维度探索了开放题文本答案的半自动编码。

Andrea Esuli 和 Fabrizio Sebastiani [33]开发了一种自动逐字编码系统(VCS)。该系统将人工对选项编码为所属类别的记录标记为正例，将人工对选项编码不所属类别的记录标记为反例，基于正反例进行学习，生成一个二进制编码模型，在编码的准确性、训练效率、效率等方面具有很好的效果。

Schonlau 和 Couper [34]认为半自动编码无法完全替代人工编码，并提出了一种半自动算法对开放题文本答案进行编码，即以 80%的准确度为阈值，当正确分类的概率超过 80%时，采用多项式梯度 boosting 算法进行半自动编码；当正确分类概率低于 80%时，采用人工编码。

Gweon 和 Schonlau [35]等学者针对编码成本高的问题，提出了三种职业半自动编码方法。此类方法是在建立详细职业代码和职业组合代码模型的基础上，将重复方法与统计学习算法相结合，改进了最近邻方法，构建了混合方法。在利用德国综合社会调查(ALLBUS)中开放题文本答案进行的编码实验中，有效地提高了编码的精度。

吴琼等人[36]关于职业的调查问卷中，设置了开放题，在中文职业开放题文本编码时建立了朴素贝叶斯分类器和支持向量机分类器，在大类编码上效果较好，而在细类上有所欠缺，性能上还需进一步提高。

Zhoushanyue He [37]提出“复制”“消除差异”“多数票表决”和“专家解决”共四种由两位编码人员独立编码的双编码策略，并与单编码在提高自动编码能力方面进行比较。结果表明：在预算分配方

面得出固定预算下，双编码对机器学习算法有一定的提高；在不存在预算约束且文本已经双编码的情况下，所有双编码策略的性能普遍优于单编码策略；在固定预算下，由专家解决训练文本中的分歧对准确率的提高最大，其次是消除分歧。

陈曦[18]运用 KNN 算法对社情民意调查问卷中的半开放题进行文本分类，最终各类准确率达到 85%以上。说明利用文本挖掘技术在社情民意问卷中处理半开放题有较好的效果，然而文章只对半开放题进行了应用，缺少对纯开放题的实验。

刘娅[38]首次采用四种机器学习分类器对中文职业文本进行编码，同时也是深度学习在中文职业编码上的初次尝试。对中国劳动力动态调查(CLDS) 2016 年数据集，分别测试朴素贝叶斯、逻辑回归、随机森林和卷积神经网络在中文职业编码的效果。结果显示，四种方法均在职业大类上表现良好。采用对职

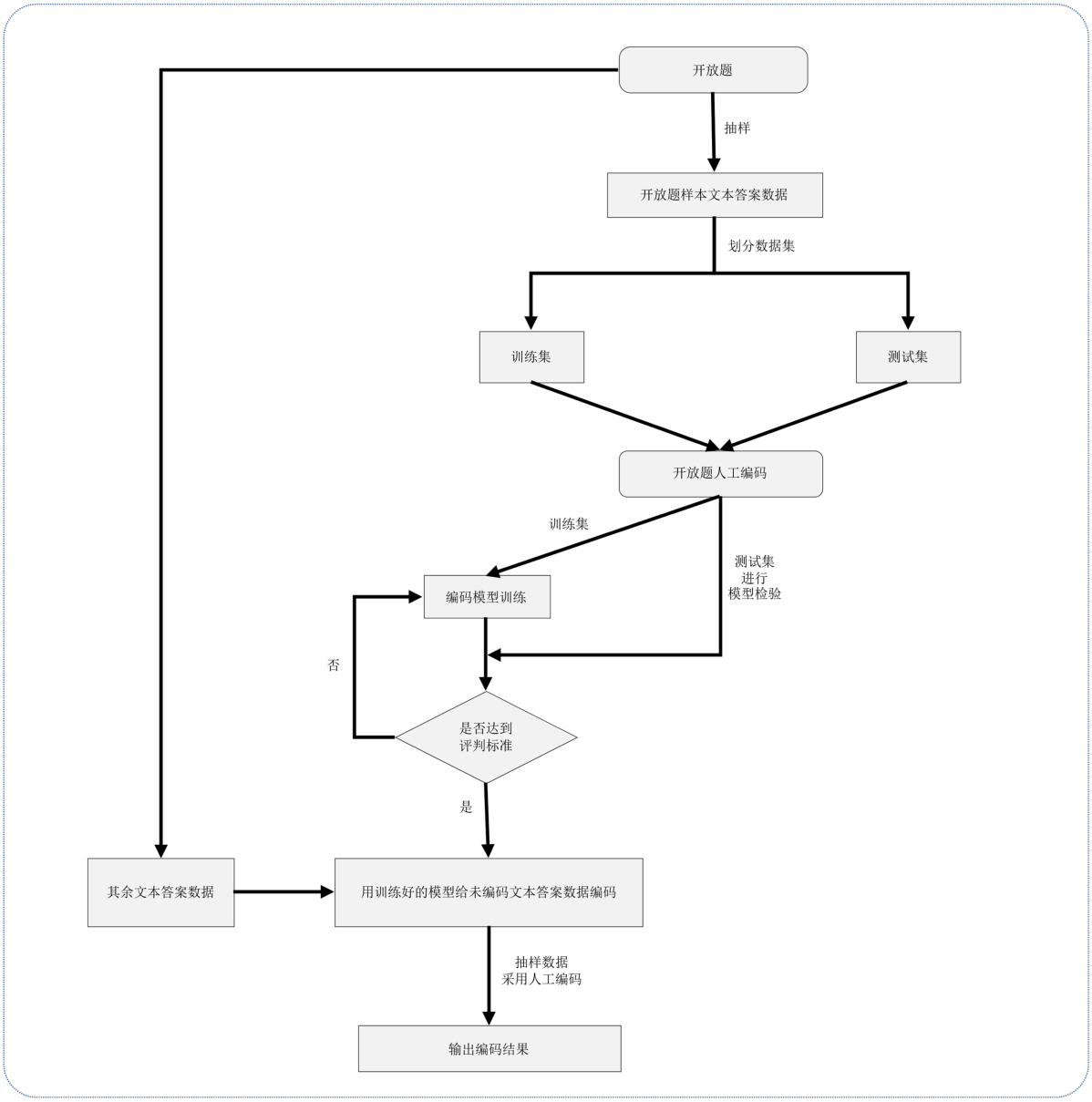

Figure 2. Semi-automatic coding flowchart
图 2. 半自动编码流程图

业类别按照相同比例扩大原始数据集，从而来提高在职业细类上的准确度。发现数据量越大，效果越优。在二十万数据量时，逻辑回归和随机森林分类器效果较令人满意，但深度学习的分类器效果并未比传统的机器学习分类器表现良好。

张静[39]针对问卷调查领域中的评价类开放性问题，结合 LDA 主题模型和基于情感词典的文本情感分析建立了“主题分类 + 五极情感”的开放题编码模型，最终得到 F 值稳定在 70%左右，然而文章仅针对于评价类问题开展，可以进一步研究对于所有开放性题目的编码模型。

Schonlau 等人[40]运用多标签算法将开放题的答案编码为多个标签，分别在三个开放性题目数据上进行实验，由于考虑答案编码之间的相关性，因而取得了良好的预测效果。同时发现，使用 0/1 损失多标签算法可以对具有多个答案的开放题实现自动分类。

总之，半自动编码采用人工 + 机器的模式，在开放题编码上取得了良好的效果，半自动编码的流程如图 2 所示。随着信息和文本相关技术的不断发展，人们更加倾向于自动化实现，自动编码的编码方式也逐渐出现。

半自动编码虽然在一定程度上克服了人工编码效率低以及个人主观原因造成的编码不准确问题，但是仍然不能很好地满足大数据背景下大规模网络调查问卷中开放题文本答案进行实时编码分析的需求，对开放题文本答案进行自动编码势在必行。

4. 自动编码

自动编码是完全不需要人工辅助进行的文本答案编码方式，他是利用已经构建的自动编码模型，建立开放题文本答案与编号代码之间的对应关系。随着深度学习在自然语言处理领域研究与应用的不断深入，开放题文本答案自动编码越来越受到关注，成为了重要的研究课题。许多学者开始尝试运用深度学习来研究并解决开放题文本答案的自动编码问题，有效地减少耗时、降低成本和提高编码质量[28] [41]。

刘泉凤[42]提出了一种改进的蚁群算法用于文本聚类，该方法适合当自动归类不断增加时的开放式自动归类方法，归类效果良好。然而对于自动阈值的研究还有待完善。

宁温馨[43]提出了一种用于中文临床诊断的自动 ICD-10 编码算法。不仅基于文字，而且基于汉字构建了文字向量，并测试了两者对精度和查全率的影响。结果表明，该算法在测试集上达到了较高的精度。

贾长娥[44]实现了用 Seq2Seq 的自动编码模型进行 BiLSTM 模型参数的预训练，之后将参数带入到双向长短期记忆网络进行句子自动编码，最终的答案选择模型结果有显著提高。与非文本特征进行结合在整体测试集上表现优于基线。

Yu 等[45]在医学方面，提出了一种多层注意力机制的双向循环神经网络来实现编码分配。通过将字符向量和词向量相结合来表示文本，将注意力机制引入 Bi-LSTM，有助于解决 RNN 在遇到长文本时性能下降的问题。实验表明，注意机制在处理中文临床记录中有着至关重要的作用。

候雪飞[46]针对常用诊断对照表所包含的疾病编码数据，采用基于词向量扩充的 CNN 模型学习从训练集中学习文本与编码之间的映射关系，给出相应编码；对训练集之外的数据(即少部分诊断无法编码的数据)，采用基于 TF-IDF 的相似度计算来筛选与疾病具有一定相似度的编码。最终使用实例对照表来解决剩余编码困难的文本。结果表明，在医疗数据，深度学习方法较传统机器学习方法表现更良好，解决了在医院诊断过程中使用最频繁和疑难编码问题，提高了在疾病自动编码方面的准确度。但该方法的不足之处在于 ICD 编码仅限于最常见的 271 个编码类别，所包含范围不全面，缺少通用性。

冯读娟[47]分别使用卷积神经网络和双向门控循环单元两个编码器构建了基于双编码器网络结构的 CGAtten-GRU 模型。在编码端，源文本并行进入双编码器，结合两种编码网络结构的输出结果构建注意

力机制，文本生成的质量得到了一定的提高，但对生成结果词语重复的现象还可以进一步提高注意力区分度进行改善。

王红斌[48]等人为了达到提高摘要质量的目的，运用英文新闻文本，在编码部分采用层级编码的方式进行，逐级向双向 GRU(门控循环单元)中加入词级注意力、句级注意力，得到结合层级注意力的文章向量表示。最终结果目的，F1 分值比 baseline 有明显提升。

贾冉冉[49]对 2014、2016 年的 CLDS 数据集，采用 N-Gram 方法构建职业词典，随后根据词典对内容进行匹配。对职业词典自动编码分类和编码员职业编码，用 Logistic 回归分析两者一致性。得到 Bigram 提取特征词所构建的词典在大类准确率 70%以上，细类上准确率 50%以上。但职业覆盖范围和样本有限，可能会存在一定量的误差。刘忠辉[50]在同样的数据集上分别采用 XGBoost、逻辑回归和 Bert 模型进行自动职业编码的实验，得出结论，XGBoost 在特征提取为 TF-IDF 时效果最优，逻辑回归在词袋模型时效果最好，Bert 在不同粒度的编码效果均优于其他两模型。

总之，开放题文本答案自动编码是在人工编码、半自动编码的基础上，借助于机器学习、深度学习等智能化算法而逐步发展起来的编码方法，具有效率高、成本低等特点，具体流程如图 3 所示。

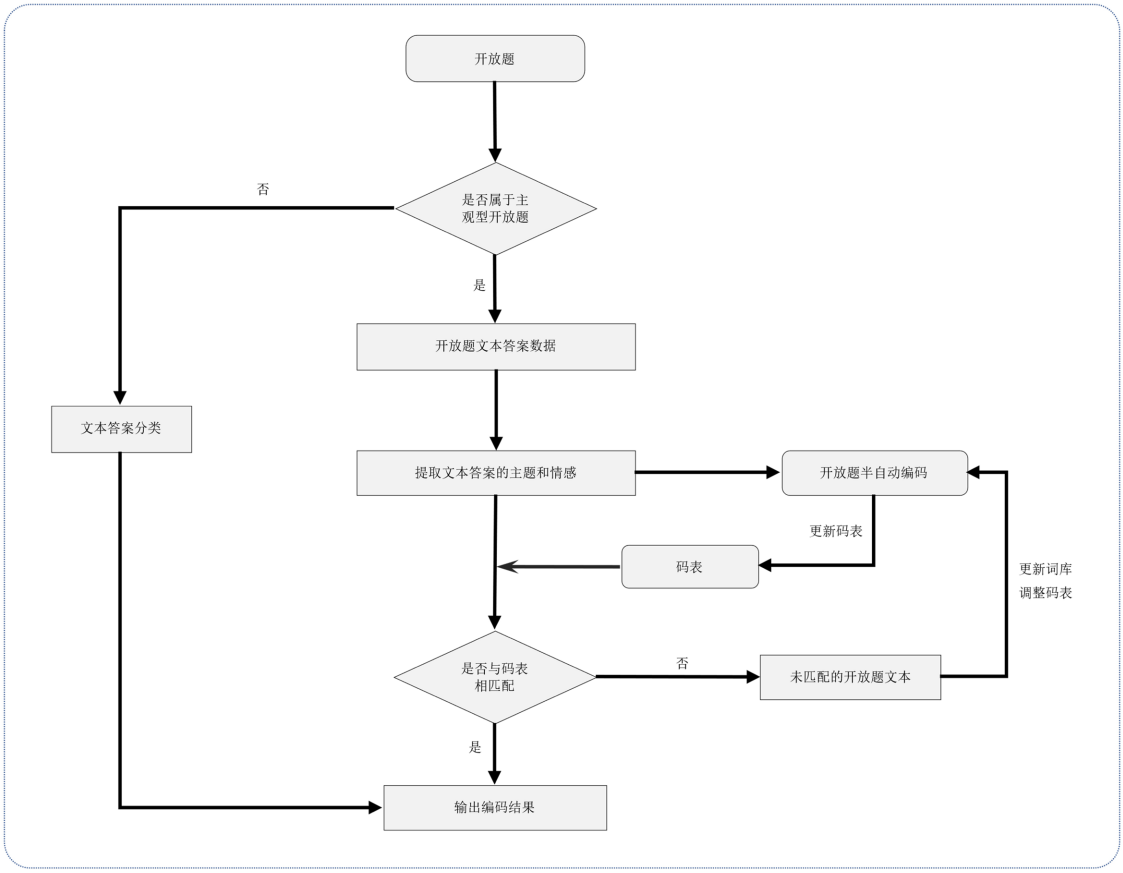

Figure 3. Automatic coding flowchart
图 3. 自动编码流程图

与人工编码、半自动编码相比，自动编码较适合对大规模调查问卷中开放题文本答案的编码预处理。不过，由于文本分析技术的局限性，自动编码仍处于探索研究阶段，其编码的准确性与泛化能力有待于进一步提高。三种编码方式的具体特征比较，如表 2 所示[38]。

Table 2. A comparison of three different coding methods
表 2. 三种不同编码方式的比较

编码方式	成本	效率	一致性	编码准确度	时效性	大数据适用性
手动编码	高	低	良好	高	差	不适用
半自动编码	较低	较高	强	较高	较好	适用
自动编码	低	高	强	较高	较好	适用

Table 3. List of common word segmentation lexicon
表 3. 常见分词词库一览表

词库名称		特点	来源
THUOCL	清华大学开放中文词库	包含词频统计信息 DF 值(Document Frequency)，方便用户个性化选择使用；词库经过多轮人工筛选，保证词库收录的准确性；开放更新，将不断更新现有词表，并推出更多类别词表。	http://thuocl.thunlp.org/
Jieba	“结巴”中文分词	支持四种分词模式。支持繁体分词。支持自定义词典。MIT 授权协议。	https://github.com/fxsjy/jieba
HanLP	汉语言处理包	功能完善、性能高效、架构清晰、语料时新、可自定义。	https://github.com/hankcs/HanLP
FoolNLTK	可能是最准的开源中文分词	可能不是最快的开源中文分词，但很可能是最准的开源中文分词；基于 Bi-LSTM 模型训练而成；包含分词，词性标注，实体识别，都有比较高的准确率；用户自定义词典；可训练自己的模型，批量处理。	https://github.com/rockyzhengwu/FoolNLTK
sego	Go 中文分词库	词典用双数组 trie 实现，分词器算法为基于词频的最短路径加动态规划；支持普通和搜索引擎两种分词模式，支持用户词典、词性标注，可运行 JSON RPC 服务；分词速度单线程 9 MB/s，goroutines 并发 42 MB/s (8 核 Macbook Pro)。	https://github.com/huichen/sego
LTP	哈尔滨工业大学语言技术平台	针对单一自然语言处理任务，生成统计机器学习模型的工具；针对单一自然语言处理任务，调用模型进行分析的编程接口；使用流水线方式将各个分析工具结合起来，形成一套统一的中文自然语言处理系统；系统可调用的，用于中文语言处理的模型文件；针对单一自然语言处理任务，基于云端的编程接口。	https://github.com/HIT-SCIR/ltp
NLPIR	中科院计算所 NLPIR-ICTCLAS 分词系统	针对大数据内容处理的需要，融合了网络精准采集、自然语言理解、文本挖掘和网络搜索技术的十三项功能，提供客户端工具、云服务、二次开发接口。	http://ictclas.nlpir.org/nlpir/

由表 2 可知，以上三种编码方式各自有各自的特点以及优势，并非可以完全互相取代，而是需要根据其特点以及不同的调查问题以及目的，去有选择性的选取最优的编码方式。人工手动编码的准确度易受编码人员的主观影响，需在进行编码操作前对编码人员进行相关的信度测试。当今的大数据时代，自动编码已成为今后的重要发展方向。

5. 常用分词与情感词库

分词词库与情感词库的构建是开放题文本答案编码的前提和基础，对准确理解文本信息、科学进行文本编码具有至关重要的作用。

首先，对国内常见的分词词库进行了整理，具体结果如见表 3 所示。

其次，由于对开放题文本答案进行情感分析，可以进一步了解用户在产品、服务、事件等方面所表达的观点、意见、情绪等主观感受[51]，并将情感极性作为编码的重要组成部分，有助于更好的分析用户的态度。情感词库是开放题文本答案编码的基础，对最终编码的准确性有很大的影响[52]，表 4 列出了常见的情感词库。

Table 4. List of common sentiment vocabulary
表 4. 常见情感词库一览表

情感词库	特点	来源
知网情感词库 HowNet	以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。该词典主要分为中文和英文两部分，包括评价(正面、负面)、情感(正面、负面)、主张、程度级别共 4 个方面的情感文本。	https://openhownet.thunlp.org/
中文情感分析库 cnsenti	默认情绪字典是知网 Hownet。 默认情绪二分法是 DLUT 情绪词典，支持 7 类情绪，如快乐/悲伤/恨……等/。 支持导入自定义 TXT 情绪词典(POS 和 NEG)。	https://github.com/hiDaDeng/cnsenti
台湾大学 NTUSD	词典包含两个子文件，分别是 negative 和 positive 子文件。其情感倾向主要分为正面和负面两大类。目前主要是应用于网络意见挖掘，领域相关情感极性分析和文件情感分类。	https://github.com/ntunlplab/NTUSD
大连理工大学中文情感词汇本体库	该资源从不同角度描述一个中文词汇或者短语，包括词语词性种类、情感类别、情感强度及极性等信息。 中文情感词汇本体的情感分类体系是在国外比较有影响的 Ekman 的 6 大类情感分类体系的基础上构建的。在 Ekman 的基础上，词汇本体加入情感类别“好”对褒义情感进行了更细致的划分。最终词汇本体中的情感共分为 7 大类 21 小类。	http://ir.dlut.edu.cn/info/1013/1142.htm
SenticNet	作为一个知识库，提供了一组语义、情感、极性关联的 100,000 个自然语言概念。特别地，语义指与输入概念在语义上最相关的概念，情感指四个情感维度的情感值和−1 到+1 之间的情感极性值。 作为一个框架，包含了一系列将常识推理、心理学、语言学和机器学习相结合的情感分析工具和技术。	http://sentic.net/
SnowNLP	情感分析方面目前训练数据主要是电商评论数据，所以在其他领域效果不是很好。	https://pypi.org/project/snownlp/

无论是分词还是情感分析，词库的选择或构建对于开放题文本答案编码的准确性都具有很大的影响。

如果要提高编码的准确性，就需要加强领域分词词库、情感词库的构建与融合，进一步完善分词与情感词库。

6. 结论与展望

本文从开放题文本答案编码研究的必要性出发，归纳与总结了编码方法的研究历程，有助于进一步深入地了解编码原理、合理地运用编码技术以及科学地探索编码方法。

对应开放题文本答案的编码，最初完全依赖人工进行的。随着文本分析技术的发展，机器学习、深度学习以及人工智能技术逐渐融入其中，向半自动、自动编码方向发展。目前，对于开放题自动编码方法模型方面的研究已经有了实质性的进展，但仍存在诸多问题需要进一步的探索与研究。

其一，自动编码将成为未来开放题文本答案编码研究的主要方向。

当今的大数据时代，随着信息技术和文本分析技术的不断发展，在调查问卷中开放性题目的比例越来越大，由此形成了结构复杂、规模庞大的多源异构开放题文本答案大数据。要对这样的文本答案大数据进行编码，人工编码因成本高、效率低几乎不可能完成，借助于机器学习、深度学习与人工智能技术以及 Transformer、BERT、GPT、文心一言等语言生成模型开展自动编码研究，将成为未来的主要方向。

其二，提高模型的泛化能力也是自动编码研究的重点内容。

由于开放题中主、客观题型之间或不同领域之间答案的文本数据具有较大的差异，而不同题型或不同领域的自动编码模型具有较强的针对性，普适性方面表现不佳，因此，增强开放题文本答案自动编码模型的自适应性使其具有良好的泛化能力，也是今后研究的重要内容。

其三，进一步完善分词词库与情感词库，将有助于提高开放题文本答案自动编码的准确性。

由于自动编码模型编码的准确性在很大程度上依赖于相关领域分词词库与情感词库的构建水平，已建的词库不够完善且涵盖的领域范围不广，无法满足层出不穷的新科技、新产业、新模式与新业态等全新领域开放题文本答案自动编码的需要，因此新建面向全新领域的词库以及对现有专业领域的词库进行扩充完善，也是不可或缺的研究工作，要统筹规划协同共建，应重视这方面的研究工作。

其四，开放题文本答案自动编码结果的应用研究将会受到越来越多业界的关注。

开放题文本答案编码的目的是为了更好地利用调查问卷中样本数据进行深层次的统计分析，从深度和广度两个维度拓展开放题编码结果的应用。例如，将开放题文本答案的编码结果运用于数据的预处理方面，可以通过与调查问卷中其他相关封闭题进行比较，探索一致性、有效性检验方法，以此提高样本数据的质量；又如，将开放题文本答案的编码结果与调查问卷中其他封闭题结合进行因子分析、关联性分析、可视化分析等多变量统计分析，可以挖掘变量之间更深层次的统计规律；另外，利用调查问卷中开放题文本答案的编码结果，也有助于统计分析报告的自动生成研究等等。

参考文献

[1] 李林梅. 试论市场调查中问卷设计的几个基本原则[J]. 统计与信息论坛, 2000, 15(2): 45-47+59.

[2] 许奎, 冷艳梅. 新时代中国特色国家审计项目组织管理影响因素研究——基于访谈和问卷的调查分析[J]. 审计研究, 2022(5): 49-55.

[3] Szűcs, V., Szabó, E. and Bánáti, D. (2015) Exploration of Healthy Nutrition Attitudes Using a Questionnaire Survey. *Orvosi Hetilap*, **156**, 636-643. https://doi.org/10.1556/OH.2015.30129

[4] Hone, K.S. and El Said, G.R. (2016) Exploring the Factors Affecting MOOC Retention: A Survey Study. *Computers & Education*, **98**, 157-168. https://doi.org/10.1016/j.compedu.2016.03.016

[5] Udo, G.J. (2013) Privacy and Security Concerns as Major Barriers for E-Commerce: A Survey Study. *Information Management & Computer Security*, **9**, 165-174. https://doi.org/10.1108/EUM0000000005808

[6] 王志刚, 刘子明, 刘超. 农产品质检体系建设对机构整合的影响——基于冀鄂鲁陕四省 210 家质检组织的调查

问卷[J]. 农业经济与管理, 2022(4): 61-70.

[7] 赵峰, 王轶. 市场化信贷、非市场化信贷对返乡创业企业绩效的影响研究——基于中国返乡创业调查问卷的证据[J]. 经济纵横, 2022(4): 67-81.

[8] 周晓清, 毛方吉, 詹春青, 焦建利. 中小学管理者对智慧课堂的认知及其态度调查——基于377份中小学管理者的调查问卷分析[J]. 现代教育技术, 2021, 31(5): 104-110.

[9] 张羽冠, 申乐, 张圣洁, 王惠珍, 张秀华, 黄宇光. 新型冠状病毒肺炎疫情期间北京协和医院手术室内医护人员头面部防护情况问卷调查[J]. 中国医学科学院学报, 2021, 43(5): 767-772.

[10] 康等银. 关于调查问卷设计应注意几个问题的研究[J]. 科技信息, 2009(23): 608+622.

[11] 潘绥铭, 黄盈盈, 王东. 问卷调查: 设置“开放题”是一种失误[J]. 社会科学研究, 2008(3): 81-85.

[12] 肖富群. 调查研究中开放式问题的编码[J]. 统计与决策, 2007(5): 73-74.

[13] 蔡鸿云, 王静, 李雪松. 文旅融合背景下云南旅游市场分析及策略研究——基于问卷和网络文本的结合分析[J]. 统计与管理, 2021, 36(1): 81-88.

[14] 王俊芳, 时俊卿. 问卷调查的类别、优缺点及实施[J]. 教育科学研究, 2004(9): 58-59.

[15] Song, G., Ye, Y., Du, X., Huang, X. and Bie, S. (2014) Short Text Classification: A Survey. *Journal of Multimedia*, **9**, 635-643. https://doi.org/10.4304/jmm.9.5.635-643

[16] 王昕. 青少年隐私调查中的“主体”反抗——基于问卷调查开放题的反思[J]. 中国青年研究, 2016(10): 10-14.

[17] 吕品, 武秦娟, 许嘉. 上市公司文本信息披露智能分析研究综述[J]. 计算机工程与应用, 2021, 57(24): 1-13.

[18] 陈曦. 文本挖掘技术在社情民意调查中的应用[J]. 中国统计, 2019(6): 27-29.

[19] 郑晶晶. 问卷调查法研究综述[J]. 理论观察, 2014(10): 102-103.

[20] 王俊杰, 韩孟杰, 陈清峰. 大学生艾滋病传播潜在风险网络测试问卷重复测试一致性分析[J]. 中国艾滋病性病, 2022, 28(10): 1150-1153.

[21] 侯俊峰. 基于编码—解码模型的序列映射若干问题研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2020.

[22] 夏海力, 朱诗晗, 李雨璇. 苏州市夜间旅游创新发展路径研究——基于网络文本和问卷调查的分析[J]. 苏州科技大学学报(社会科学版), 2021, 38(6): 24-31+107.

[23] 肖洁, 卜林, 孙婷妹. 浅析开放式问卷的调查与设计[J]. 内江科技, 2010, 30(1): 50+154.

[24] 武庆玲. 定量项目问卷的编码要求及原则[J]. 市场研究, 2006(5): 44-47.

[25] Popping, R. (2013) Analyzing Open-Ended Questions by Means of Text Analysis Procedures. *Bulletin of Sociological Methodology*, **128**, 23-39. https://doi.org/10.1177/0759106315597389

[26] 李耀. 顾客单独创造价值的结果及途径——一项探索性研究[J]. 管理评论, 2015, 27(2): 120-127.

[27] 百度文库. 问卷调查中的编码技巧[EB/OL]. https://wenku.baidu.com/view/7fdfe886de3383c4bb4cf7ec4afe04a1b071b0b7.html, 2022-12-28.

[28] 任莉颖, 邱泽奇, 李力, 严洁. 社会调查中职业问题编码的方式与质量研究[J]. 浙江大学学报(人文社会科学版), 2012, 42(3): 210-219.

[29] Popping, R. (2012) Human or Machine Coding of Open-Ended Questions. *Bulletin of Sociological Methodology*, **115**, 79-88. https://doi.org/10.1177/0759106312445710

[30] Popping, R. and Roberts, C.W. (2019) Coding Issues in Semantic Text Analysis. *Field Methods*, **21**, 244-264. https://doi.org/10.1177/1525822X09333433

[31] He, Z. and Schonlau, M. (2022) A Model-Assisted Approach for Finding Coding Errors in Manual Coding of Open-Ended Questions. *Journal of Survey Statistics and Methodology*, **10**, 365-376. https://doi.org/10.1093/jssam/smab022

[32] 李煜, 徐安琪. 普通人的爱情观研究——兼开放式问题的量化尝试[J]. 社会科学, 2007(7): 132-141.

[33] Esuli, A. and Sebastiani, F. (2010) Machines That Learn How to Code Open-Ended Survey Data. *International Journal of Market Research*, **52**, 775-800. https://doi.org/10.2501/S147078531020165X

[34] Schonlau, M. and Couper, M.P. (2017) Semi-Automated Categorization of Open-Ended Questions. *Survey Research Methods*, **10**, 143-152.

[35] Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2017) Three Methods for Occupation Coding Based on Statistical Learning. *Journal of Official Statistics*, **33**, 101-122. https://doi.org/10.1515/jos-2017-0006

[36] 吴琼, 戴利红, 张婧申. 机器学习在社会调查职业编码中的应用[J]. 调研世界, 2019(9): 56-60.

[37] He, Z. and Schonlau, M. (2020) Automatic Coding of Text Answers to Open-Ended Questions: Should You Double Code the Training Data? *Social Science Computer Review*, **38**, 754-765. https://doi.org/10.1177/0894439319846622

[38] 刘娅. 基于机器学习的自动化职业编码[D]: [硕士学位论文]. 大连: 东北财经大学, 2021.

[39] 张静. 问卷调查中评价类问题的自动编码方法及其应用[D]: [硕士学位论文]. 天津: 天津商业大学, 2021.

[40] Schonlau, M., Gweon, H. and Wenemark, M. (2021) Automatic Classification of Open-Ended Questions: Check-All-That-Apply Questions. *Social Science Computer Review*, **39**, 562-572. https://doi.org/10.1177/0894439319869210

[41] 淦亚婷, 安建业, 徐雪. 基于深度学习的短文本分类方法研究综述[J]. 计算机工程与应用, 2023, 59(4): 43-53.

[42] 刘泉凤. 一种基于文本聚类的开放式信息自动归类方法[J]. 情报杂志, 2009, 28(6): 177-180.

[43] 宁温馨, 于明. 基于语义相似度计算的临床诊断自动编码算法研究[J]. 医学信息学杂志, 2016, 37(2): 52-56.

[44] 贾长娥. 基于深度学习的答案选择[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2017.

[45] Yu, Y., Li, M., Liu, L., *et al*. (2019) Automatic ICD Code Assignment of Chinese Clinical Notes Based on Multilayer Attention BiRNN. *Journal of Biomedical Informatics*, **91**, Article ID: 103114. https://doi.org/10.1016/j.jbi.2019.103114

[46] 候雪飞. 面向医疗数据的实体分析与自动编码技术研究与应用[D]: [硕士学位论文]. 石家庄: 河北科技大学, 2019.

[47] 冯读娟, 杨璐, 严建峰. 基于双编码器结构的文本自动摘要研究[J]. 计算机工程, 2020, 46(6): 60-64.

[48] 王红斌, 金子铃, 毛存礼. 结合层级注意力的抽取式新闻文本自动摘要[J]. 计算机科学与探索, 2022, 16(4): 877-887.

[49] 贾冉冉. 基于 N-Gram 提取特征词典的职业编码研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2022.

[50] 刘忠辉. 基于机器学习的职业编码方法研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2022.

[51] 曾义夫, 蓝天, 吴祖峰, 刘峤. 基于双记忆注意力的方面级别情感分类模型[J]. 计算机学报, 2019, 42(8): 1845-1857.

[52] 邓东. 情感词典构建方法及其应用研究[D]: [博士学位论文]. 北京: 北京交通大学, 2019.