

文章编号: 1003-0077(2020)02-0046-10

基于 Bi-GRU 并包含注意力机制的文本数据真值发现

常宸¹, 曹建军², 吕国俊¹, 郑奇斌¹, 翁年凤²

(1. 陆军工程大学 指挥控制工程学院, 江苏 南京 210007;

2. 国防科技大学 第六十三研究所, 江苏 南京 210007)

摘要: 针对传统真值发现算法无法直接应用于文本数据的问题, 该文提出基于 Bi-GRU 并包含注意力机制的文本数据真值发现方法。根据文本答案的多因素性, 词语使用的多样性与文本数据的稀疏性等特点, 该文对用户答案进行细粒度划分, 并利用 Bi-GRU 表征文本答案的语义信息, 利用双层注意力机制分别学习用户答案关键词可靠度及用户答案可靠度。依据真值发现的一般假设, 无监督学习上下文向量, 并最终获得可靠答案。实验结果表明, 该算法适用于文本数据真值发现场景, 较基于检索的方法及传统真值发现算法效果更优。

关键词: 数据质量; 真值发现; 神经网络; 文本挖掘

中图分类号: TP311

文献标识码: A

Truth Discovery from Text Data by Bi-GRU with Attention Mechanism

CHANG Chen¹, CAO Jianjun², LV Guojun¹, ZHENG Qibin¹, WENG Nianfeng²

(1. Institute of Command and Control Engineering, Army Engineering University, Nanjing, Jiangsu 210007, China;

2. The Sixty-third Research Institute, National University of Defense Technology, Nanjing, Jiangsu 210007, China)

Abstract: To discovery the truth from text data, a truth discovery method by Bi-GRU with attention mechanism is proposed. According to the characteristics of the text data (the multifactorial property of text answers, the diversity of word usages, and the sparseness of the text data), the fine-grained user answers are taken as input of network, and then the Bi-GRU is utilized to extract semantic information of user answers. Moreover, the keyword reliability and answer reliability are learned with attention mechanism. Finally, the context vector for each problem can be learned without supervision according to the general hypothesis of truth discovery. The experiment results show that the proposed algorithm is better than the retrieval based methods and other traditional truth discovery methods.

Keywords: data quality; truth discovery; neural network; text mining

0 引言

随着大数据时代的来临, 互联网信息量增长迅速, 而由于互联网具备开放性和多源性, 不同互联网平台提供的数据各不相同, 大量错误、过时、不完整和虚假信息充斥着网络, 并且还存在着很多恶意数据源, 这些恶意数据源提供大量的错误信息甚至谣言, 大大干扰我们判断信息的正确性^[1]。如何从低质量数据中获得可靠信息是一个具有挑战性的课题。

随着时代的发展, 数据的形式在不断发生变化, 真值发现也在面临新的挑战。如今, 用户可以通过在线平台访问大量数据并寻找问题的答案(例如, Amazon Mechanical Turk, Stack Exchange 和 Yahoo Answers 等), 然而这种多源数据通常由非专家在线用户提供, 可能存在错误, 甚至冲突, 并且大多数以文本的形式存在, 如何充分利用自然语言的特性, 从众多文本答案中找出正确答案, 成为新的研究方向。

首先, 多个数据源对事实问题的回答可能是多因素的, 并且给定的文本答案通常很难涵盖全面。

收稿日期: 2019-09-17 定稿日期: 2019-11-22

基金项目: 国家自然科学基金(61371196); 中国博士后科学基金(20090461425, 201003797); 国家重大科技专项(2015ZX01040201-003)

如对于问题“*What are the symptoms of flu?*”, 正确答案应包含以下因素: fever, chills, cough, nasal symptom, ache, fatigue, 即使用户提供的答案涵盖两个因素(如 fever 和 ache), 现有的真值发现方法将整个答案视为一个整体单元, 可能确定该答案是完全错误的并且为该用户分配低可靠度。但是, 如果考虑到细粒度的答案因素, 那么这个用户提供的答案是部分正确的, 这意味着应该适当提高该用户的可靠度。同时, 在线用户提供的答案可能会传达与不同关键字非常相似的含义。例如, 用户 1 使用“*exhausted*”而用户 2 使用“*fatigue*”来表达诸如疲惫之类的症状。然而, 现有的真值发现方法可能会将它们视为完全不同的答案。因此, 在推断可靠信息时对文本数据中的语义信息充分利用是非常重要的。此外, 传统真值发现算法从数据源的众多观测值中挖掘数据源可靠度信息, 进行真值发现。而对于文本数据真值发现场景, 同一问题可能由众多网络用户回答, 而同一用户回答的问题却很少, 数据稀疏会对用户可靠度估计造成困难。依赖用户的众多观测值进行可靠度估计的方法对于文本数据真值发现场景是不适用的。

本文提出基于 Bi-GRU 并包含注意力机制的文本数据真值发现, 充分利用文本的语义信息, 提取文本答案的细粒度特征, 解决文本答案多因素属性及词语使用多样性为文本数据真值发现带来的挑战。同时利用神经网络无监督学习众多观测值间的复杂关系, 减少对数据源可靠度估计的依赖, 解决传统真值发现算法由于强假设数据分布导致真值发现结果质量不高的问题, 克服数据稀疏对数据源可靠度评估造成的影响。

本文的主要贡献如下:

(1) 使用神经网络解决文本数据真值发现问题, 避免了传统真值发现算法由于强假设数据分布导致真值发现结果不优的问题;

(2) Bi-GRU 及 Word2Vec 捕捉文本答案的细粒度语义信息, 解决文本数据自然语言特性为真值发现带来的挑战;

(3) 利用双层注意力机制学习用户答案及答案关键词可靠度, 对文本答案进行细粒度可靠度估计;

(4) 利用答案本身的语义信息寻找可靠答案, 不需要用户信息。

1 相关工作

1.1 结构化数据真值发现

真值发现方法研究从多个数据源提供的对于多个真实对象的大量冲突描述信息中, 为每一个真实对象找出最准确的描述。根据假设越可靠的数据源提供的事实越可信, 提供越多可信事实的数据源越可靠, 传统真值发现方法可分为基于迭代的方法、基于优化的方法和基于概率图模型的方法。基于迭代的方法^[2-5]假设数据源可靠度与观测值可信度之间的关系可用简单函数表示, 迭代真值计算与数据源可靠度估计两步, 直至损失收敛, 输出数据源可靠度与各个对象的真值。基于优化的方法根据真值发现的一般假设设置目标函数, 通过优化目标函数进行真值发现, 使用优化算法进行优化或转化为基于迭代的方法进行处理^[6-9]。基于概率图模型的方法假设观测值服从概率分布, 通过采样和参数估计的方法估计真值^[10-12]。

传统的真值发现算法假设数据源可靠度和观测值可信度之间的关系可以通过简单函数(如线性函数、二次函数等)来表示。而实际上, 源可靠度和值可信度之间的关系通常是未知的, 简单假设将会导致真值发现的结果并不理想。Marshall 等^[13]首次将神经网络应用到真值发现问题中, 利用前馈神经网络解决社会感知问题, 但这种方法需要人工标记部分对象, 无法进行无监督的学习, 且仅适用于网络观测值是否为真的场景, 不能适用于真值发现的一般情况。文献^[14-15]利用受限玻尔兹曼机隐层学习数据源可靠度分布, 利用 CD 算法^[16]训练模型参数, 但由于受限玻尔兹曼机本身的局限性, 也仅适用于属性为二值的真值发现场景。之后, Li 等^[17]利用 LSTM 神经网络进行真值发现, 将不同数据源提供的“对象—属性—值”矩阵与源可靠度矩阵乘积作为输入, 将各个观测值作为真值的概率作为输出, 通过最小化真值与各数据源观测值之间的距离来优化网络参数。该模型首次利用具有比实数更强表示能力的向量来表示数据源可靠度, 将数据源信息视为潜在的背景知识, 并存储在可靠度矩阵中用来计算观测值的可信度。

1.2 文本数据真值发现

针对文本数据真值发现, 目前大部分学者将问

题进行简化,对文本数据进行粗粒度的分析,重点对社交媒体或其他网络资源中文本数据进行是否为真的判断,将问题简化为二值属性的真值发现问题。Popat 等^[18]构建“数据源—语言风格”输入向量,通过 Logistic 回归,将真值发现问题转化为二分类问题。Marshall 等^[13]利用全连接神经网络学习数据源可靠度与观测值可信度间的关联关系,同样将用户答案抽象为 0/1,并输入网络进行真值发现。Ma 等^[19]首次考虑文本数据的真值发现问题,但没有充分考虑文本的语义信息,之后所提方法将非结构化数据转化为结构化数据进行真值发现^[20]。Li 等^[21]首次将语义信息与真值发现过程融合,但该模型适用于文本答案较短或为单词的情形,无法直接适用于大多数文本数据真值发现场景。目前仅有 Zhang 等^[22]充分利用文本数据的语义信息,提出 TextTruth 真值发现算法。该方法将从特定问题的答案中提取的关键词组合成多个可解释因子,利用基于概率图模型的方法进行真值发现,来推断答案的可信度。该方法假设了用户可靠度与观测值可信度之间的分布,而实际上数据分布是未知的,强假设分布可能对真值发现的结果产生影响。

1.3 相关问题

文本数据真值发现问题与社区问答(collaborative question answer, CQA)问题相关。针对 CQA 问题,目前有两种解决方案。部分学者从答案中抽取特征,将答案质量评估问题转化为二分类问题或排序问题^[23-24]。通常,这种方法需要高质量的训练数据以及答案特征来训练模型。在实际问题中,这样的训练数据通常是难以获得的。另一部分学者将该问题转化为专家发现问题^[25-26],通过回答者的可靠度来推断答案的可信度,然而这些方法需要诸如投票信息等额外的特征。不同的问题设置及解决方案使得 CQA 问题与本文所提真值发现有所区别。

另一个与本文相关的问题是答案选择问题(answer selection, AS),该问题旨在从候选答案中选择正确可靠的答案,是问答系统的重要研究内容。传统答案选择基于答案的词汇特征^[27-28]。之后,基于神经网络的方法通过比较问题与用户答案的相似度来寻找可靠答案^[29-30]。

近年来,注意力机制被广泛应用到答案选择等文本挖掘领域,提高了神经网络进行文本语义表征的能力。Tan 等^[31]利用 Bi-LSTM 网络以及神经网络池化层对句子进行语义表征,对问题及答案的不

同关键词赋予不同的权重,提出 QA-LSTM 模型。在此基础上,Cicero 等^[32]提出了注意力池化机制,该机制允许答案之间信息共享,同时互相影响语义的表征过程。然而,这些方法都是有监督方法,而本文提出的真值发现方法不需要类标记,是一种无监督的方法。Liu 等^[33]设计了一种新的基于多模态的注意力机制神经网络框架,该框架能够通过从异构数据中学习统一的语义表示,用于在大规模教育系统中寻找相似的练习题。Yang 等^[34]提出将多层注意力机制用于文档分类任务,引入上下文向量来发现每个词语和每个句子的重要性,给与不同重要性的字词或者句子以不同的关注度,提高了最终的文本分类效果。

2 问题定义

考虑文本数据真值发现的一般模型,给定问题 q ; 用户集合 $U = \{u_j | j = 1, 2, \dots, L\}$, 其中 u_j 表示第 j 名用户; 候选答案集合 $A = \{a_i | i = 1, 2, \dots, L\}$, 其中 L 表示候选答案个数, 本文解决从众多候选文本答案中找到问题 q 的最佳答案。

表 1 列出了一个不同用户对同一问题进行回答的实例。

表 1 不同用户对同一问题进行回答实例

| 问题 | What are the symptoms of flu? |
|---------|--|
| 用户 1 答案 | People will feel cold and cough, and sometimes they will feel exhausted. |
| 用户 2 答案 | The symptoms of flu are fever and freezing. |
| 用户 3 答案 | Maybe chills, cough, and fatigue. |
| 用户 4 答案 | I don't know. |

表 1 中,对于问题“*What are the symptoms of flu?*”,三名用户给出了各自不同的答案。传统真值发现算法将每名用户的答案以整体对待,而实际上每名用户答案包含不同的关键因素,只有部分是正确的,此时应当对用户答案进行细粒度的划分,对答案语义进行提取,从而更好地度量用户答案之间的关系。本文旨在利用神经网络强大的表达寻找用户答案间的关联,充分挖掘文本的自然语言特性,依据众多用户提供的答案,寻找问题的最优答案。

与传统方法利用多名用户提供的多个答案评估用户可靠度的方法不同,本文利用同一问题大量的用户答案挖掘其语义信息及复杂的关系,不需要提供用户的信息,每个问题独立地进行真值发现。

3 预处理

在众多用户答案中,部分答案为噪声数据,不包含任何可靠信息,并且容易从众多用户答案中分离出来,如表 1 用户 4 答案“I don't know”。本文通过构造无向图 $G=(V,E)$,对用户答案进行去噪,在真值发现前去除易区分噪声答案,有效提高真值发现的效率和准确率,具体操作步骤如下:

(1) 利用 SIF 方法^[35]对用户答案进行初步语义表征,并得到用户答案向量 $\text{emb}(a_i)$ 。

(2) 将用户答案向量 $\text{emb}(a_i)$ 设置为无向图的顶点 $v_i \in V$,同时设置阈值 $\lambda \in [0,1]$,当两用户答案相似度 $s(a_i, a_j)$ 大于阈值 λ 时,构造无向图的边 (v_i, v_j) 。由式(1)计算 $s(a_i, a_j)$ 。

$$s(a_i, a_j) = 1 - \frac{1}{\pi} \cos^{-1} \left(\frac{\text{emb}(a_i) \cdot \text{emb}(a_j)}{\sqrt{\text{emb}(a_i)} \sqrt{\text{emb}(a_j)}} \right) \quad (1)$$

(3) 根据文献[36],对象的真值情况应该尽可能与各数据源提供的观测值接近,数据源质量越高则其提供答案与正确答案越相似。噪声答案通常表现为不包含任何语义信息,与较少的其他用户答案相似,而正确答案或部分正确答案由于是对同一问题的回答,表现为与较多的用户答案相似。由此,设置阈值 $\delta > 0$,当无向图中的节点度大于 δ 时(该答案与多于 δ 名用户答案相似),则保留该节点,否则认为该节点为噪声答案并删除。

4 双层注意力机制神经网络

受文献[34]启发,图 1 为本文提出的基于 Bi-GRU 并包含注意力机制的文本数据真值发现方法架构图。

该模型以全部候选答案作为输入,输出识别真值答案向量 S^* 。上下文向量 c_s 和 c_w 分别表示用户答案关键词可靠度及用户答案可靠度,基于真值发现的假设,识别真值向量将包含可靠用户提供的可靠关键词语义信息。

4.1 GRU

RNN(recurrent neural network)是一种用来处理序列数据的特殊神经网络,能够学习文本上下文的语义信息,抽取文本的语义特征,并作为神经网络或其他模型的输入。但在随时间反向传播过程中,跨时间步和长时间学习使后续节点的梯度往往不能

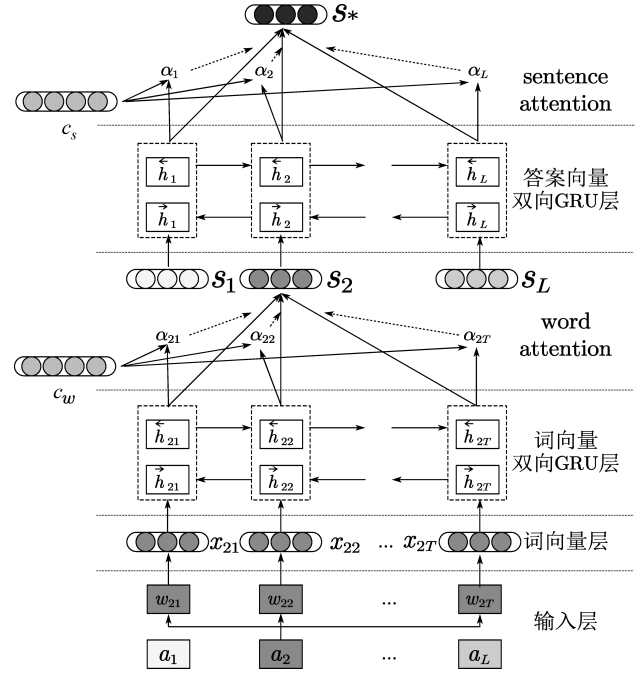


图 1 模型框架

按照初值传到最初的位置,容易出现梯度弥散问题。LSTM (long-short term memory) 作为 RNN 变体被提出并克服梯度弥散问题,而由于 LSTM 训练时间较长、参数较多且内部计算复杂,Bahdanau 等^[37]进一步提出更加简单的 GRU (gate recurrent unit) 模型,该模型将 LSTM 的单元状态和隐层状态进行合并,同时优化了网络结构。

GRU 模型具备 LSTM 优点,但同时结构更加简单,参数更少,具备更好的收敛性,能够很大程度上提高模型训练效率。GRU 由更新门和重置门两个门组成。更新门控制前一个时刻的输出隐层对当前隐层的影响程度,更新门的值越大,则前一时刻的隐层输出对当前隐层的影响越大;重置门控制前一时刻的隐层信息被忽略的程度,重置门的值越小,则忽略得越多。GRU 模型如图 2 所示。

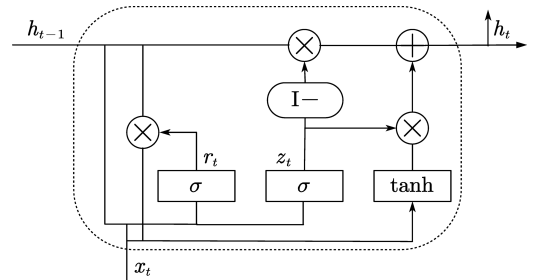


图 2 GRU 结构图

GRU 模型的更新方式如式(2)~式(5)所示。

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (2)$$

$$z_t = \sigma(W_r x_t + U_z h_{t-1} + b_z) \quad (3)$$

$$\tilde{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h) \quad (4)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (5)$$

其中, r_t 表示 t 时刻的重置门, z_t 表示 t 时刻的更新门, \tilde{h}_t 表示 t 时刻的候选激活状态, h_{t-1} 表示 $t-1$ 时刻的隐层状态。更新门 z 由当前状态需要被遗忘的历史信息和接受的新信息决定; 重置门 r 由候选状态从历史信息中得到的信息决定。本文使用 GRU 神经元作为文本答案上下文语义信息表征基本结构, 学习文本的细粒度语义信息。

4.2 双层注意力机制

设候选答案集合 $A = \{a_i | i = 1, 2, \dots, L\}$, L 表示候选答案个数。每个候选答案 $a_i = \{w_{i1}, w_{i2}, \dots, w_{it}\}$ 包含 T 个单词, $w_{it} (t \in [0, T])$ 表示答案 a_i 的第 t 单词。本文所提模型将用户答案文本进行语义抽象, 并利用网络输出识别真值。双层注意力机制能够学习对文本数据真值发现重要的关键词及可靠的用户答案。

4.2.1 单词编码

首先, 对输入层输入的候选答案进行分词, 并通过 Word2Vec 得到候选答案中每个词所对应的词向量表示。这些词向量表示保持了用户答案最原始的信息。对于用户答案 a_i 包含的单词 $w_{it}, t \in [0, T]$, 利用词向量矩阵 W_e 来获得词向量, 如式(6)所示。

$$x_{it} = W_e w_{it} \quad (6)$$

之后, 使用双向 GRU 对用户答案进行建模, 双向 GRU 可以看作两个单向的 GRU, 包含正向 GRU \vec{f} 及逆向 GRU \overleftarrow{f} , 使当前时刻的输出能与前一时刻的状态和后一时刻的状态都产生联系。 \vec{f} 从 w_{i1} 到 w_{it} 输入答案 a_i , 而 \overleftarrow{f} 从 w_{it} 到 w_{i1} 输入答案 a_i 。双向 GRU 在 t 时刻隐层状态通过前向隐层状态 \vec{h}_{it} 与后向隐层状态 \overleftarrow{h}_{it} 加权得到, 如式(7)~式(9)所示。

$$\vec{h}_{it} = \overrightarrow{\text{GRU}}(x_{it}), t \in [1, T] \quad (7)$$

$$\overleftarrow{h}_{it} = \overleftarrow{\text{GRU}}(x_{it}), t \in [1, T] \quad (8)$$

$$h_{it} = \vec{h}_{it} + \overleftarrow{h}_{it} \quad (9)$$

4.2.2 单词注意力机制

注意力机制能够在关键信息上分配足够的关注, 突出局部重要信息, 在真值发现过程中表现为单词的可信度。同理, 不是所有的单词在真值发现过程中都起到重要的作用, 不同单词应当被赋予不同

的权重。采用注意力机制能够突出重点的词语, 同时加权得到真值答案的向量表示, 使整个真值发现过程表现出更好的性能, 提高该隐藏层特征提取能力, 如式(10)~式(12)所示。

$$c_{it} = \tanh(W_w h_{it} + b_w) \quad (10)$$

$$\alpha_{it} = \frac{\exp(c_{it}^T c_w)}{\sum_t \exp(c_{it}^T c_w)} \quad (11)$$

$$s_i = \sum_t \alpha_{it} h_{it} \quad (12)$$

其中, 双向 GRU 输出 h_{it} 通过一层全连接网络得到隐层表示 c_{it} , 通过度量 c_{it} 与上下文单词向量 c_w 的相似度并进行 softmax, 得到单词可靠度 α_{it} 。至此, 用户答案 a_i 的向量表示 s_i 表示为各个隐层状态在新的隐层状态下单词可靠度系数 α_{it} 与初始输入的各个隐层状态 h_{it} 乘积的累加和。过程中, 上下文向量 c_w 担任“询问”的工作, 类似记忆神经网络中“重要词汇是什么?”的高层表征。 c_w 随机初始化, 并在真值发现过程中不断优化。

4.2.3 答案注意力机制

为了建模用户答案可靠度, 再次使用答案级别的注意力机制, 并使用上下文向量 c_s 计算各个用户答案的可靠度 α_i , 如式(13)~式(15)所示。

$$c_i = \tanh(W_s h_i + b_s) \quad (13)$$

$$\alpha_i = \frac{\exp(c_i^T c_s)}{\sum_i \exp(c_i^T c_s)} \quad (14)$$

$$s_* = \sum_i \alpha_i h_i \quad (15)$$

其中, s_* 为网络输出的识别真值向量。聚合了问题 q 所有用户答案的正确语义信息, 注意力机制赋予不同的用户答案不同的可靠度。上下文向量 c_s 随机初始化, 并在真值发现过程中不断优化。

4.2.4 真值发现

根据假设: ①问题答案的真值情况应该尽可能与各数据源提供的观测值接近; ②数据源的质量越高则其提供的问题答案越相似^[36], 将模型损失函数定义如式(16)所示。

$$L = \sum_i d(\theta | s_i, s_*) + \frac{1}{2} \|w\|^2 \quad (16)$$

其中, θ 为网络中的所有参数, $\frac{1}{2} \|w\|^2$ 为正则项, 使权重更接近原点。 s_* 为网络的最终输出, 表示本次迭代的识别答案真值向量。 s_i 为问题的第 i 名用户答案经 Bi-GRU 及单词注意力机制语义表征后得到的向量。 $d(s_i, s_*)$ 定义为识别答案真值向量与数据源观测值答案向量之间规范化后的余弦距

离,如式(17)所示。

$$d(s_i, s_*) = \frac{1}{\pi} \cos^{-1} \left(\frac{s_i \cdot s_*}{\sqrt{s_i} \sqrt{s_*}} \right) \quad (17)$$

该损失函数以识别答案真值和各数据源提供的答案观测值之间距离之和最小为优化目标。

5 用户答案评分

通过对双层注意力机制多次优化迭代直至网络参数收敛,得到最终的识别真值向量 s_* 。依据各个数据源提供的答案观测值向量 s_i 与识别真值向量 s_* 的相似度,定义各个答案的分数,如式(18)所示。

$$\text{score}_i = 1 - \frac{1}{\pi} \cos^{-1} \left(\frac{s_i \cdot s_*}{\sqrt{s_i} \sqrt{s_*}} \right) \quad (18)$$

至此,模型根据答案分数输出答案可靠度排名,上下文向量 c_w 和 c_s 存储了关键词及用户答案的可靠度信息。

6 实验与结果

本节通过在真实数据集上进行对比实验,验证基于 Bi-GRU 并包含注意力机制的文本数据真值发现算法的有效性与准确性。使用 tensorflow 框架实现网络并进行训练,CPU 为 Inter Xeon E5-2630,内存 192 GB, GPU 为 Nvidia Tesla P40 \times 2,采用 CentOS 7 64 位操作系统。

6.1 数据集

实验数据 Short Answer Scoring 来源于 Kaggle 竞赛 The Hewlett Foundation: Short Answer Sco-

ring,该数据集包含科学、英语与艺术、生物、英语四个学科 6 个问题,每个子数据集由问题及大量学生回答组成,不提供回答者信息,每个回答的平均长度为 50 个单词,每个问题候选答案的平均个数为 1 500。所有的答案均由学生撰写,并经过相关人员手动打分。

6.2 评价指标

文本数据真值发现旨在寻找众多用户答案中的可靠答案,本文为每个问题独立发现可靠答案,并以该问题的 Top-N ($N=10, 30, 50, 80, 100, 200$) 答案的平均分作为本次实验的评价指标。

6.3 对比算法

基于检索的方法 SIF Similarity 对文本答案进行 SIF 语义表征,根据答案与问题的相似度从众多答案中检索正确的答案,答案与问题相似度越高,则认为该答案越可靠。

基于检索的方法 BOW Similarity 对文本答案进行 BOW 语义表征,与 SIF Similarity 类似,但 BOW 文本表征方法粒度较粗,没有包含文本的语义信息。

真值发现算法 CRH^[8] + SIF 对文本答案进行 SIF 语义表征,利用 CRH 真值发现算法进行文本数据的真值发现,并使用本文提出的距离函数度量文本答案间的相似性。

6.4 实验结果分析

6.4.1 对比实验结果

将本文所提方法与对比算法进行比较,结果如表 2 所示。

表 2 对比实验结果统计表数据集对比算法

| 数据集 | 对比算法 | Top-10 | Top-30 | Top-50 | Top-80 | Top-100 | Top-200 |
|---------------|-----------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Short Answer1 | SIF Similarity | 0.00 | 0.36 | 0.84 | 1.16 | 1.25 | 1.58 |
| | BOW Similarity | 1.00 | 0.87 | 1.23 | 1.05 | 1.22 | 1.24 |
| | CRH + SIF | 1.20 | 1.40 | 1.58 | 1.49 | 1.46 | 1.49 |
| | Proposed Model | 1.60 | 1.72 | 1.60 | 1.53 | 1.52 | 1.50 |
| Short Answer2 | SIF Similarity | 1.70 | 1.93 | 1.88 | 1.90 | 1.90 | 1.89 |
| | BOW Similarity | 1.52 | 1.32 | 1.34 | 1.53 | 1.62 | 1.73 |
| | CRH + SIF | 2.50 | 2.50 | 2.36 | 2.28 | 2.16 | 2.17 |
| | Proposed Model | 2.70 | 2.62 | 2.38 | 2.28 | 2.26 | 2.17 |

续表

| 数据集 | 对比算法 | Top-10 | Top-30 | Top-50 | Top-80 | Top-100 | Top-200 |
|---------------|-----------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Short Answer3 | SIF Similarity | 0.99 | 1.00 | 1.05 | 1.10 | 1.05 | 1.08 |
| | BOW Similarity | 0.93 | 1.05 | 1.03 | 1.02 | 1.05 | 1.03 |
| | CRH+SIF | 0.90 | 1.01 | 1.10 | 1.03 | 1.04 | 1.07 |
| | Proposed Model | 1.15 | 1.20 | 1.17 | 1.10 | 1.08 | 1.08 |
| Short Answer4 | SIF Similarity | 0.30 | 0.13 | 0.22 | 0.29 | 0.39 | 0.51 |
| | BOW Similarity | 0.00 | 0.00 | 0.02 | 0.16 | 0.22 | 0.34 |
| | CRH+SIF | 1.40 | 1.33 | 1.30 | 1.15 | 1.09 | 0.97 |
| | Proposed Model | 1.50 | 1.42 | 1.38 | 1.27 | 1.25 | 1.10 |
| Short Answer5 | SIF Similarity | 1.00 | 1.37 | 1.42 | 1.38 | 1.41 | 1.49 |
| | BOW Similarity | 1.20 | 1.17 | 1.22 | 1.18 | 1.13 | 1.20 |
| | CRH+SIF | 1.80 | 1.75 | 1.45 | 1.63 | 1.56 | 1.37 |
| | Proposed Model | 1.83 | 1.87 | 1.63 | 1.63 | 1.60 | 1.45 |
| Short Answer6 | SIF Similarity | 1.40 | 1.20 | 1.08 | 1.10 | 1.12 | 1.12 |
| | BOW Similarity | 1.70 | 1.73 | 1.64 | 1.65 | 1.63 | 1.62 |
| | CRH+SIF | 1.60 | 1.67 | 1.62 | 1.60 | 1.59 | 1.59 |
| | Proposed Model | 1.82 | 1.78 | 1.74 | 1.70 | 1.62 | 1.60 |

由表 2 可知,本文所提方法在大部分情况下优于所有对比算法。首先,基于检索的方法根据答案与问题的相似度对答案进行排名,而实际上问题不一定包含答案需要的关键因素,基于检索的方法只是找到了与问题相关的答案,而不一定是正确的答案。另外,CRH 真值发现算法假设数据源可靠度与观测值可信度之间的关系可用简单函数表示,而这种关系实际上是未知的,强假设会对真值发现的结果产生影响。同时,该算法通过数据源提供的各个对象的大量观测值,迭代计算数据源可靠度,直至收敛,而对于文本数据,数据源众多,而每个数据源的观测值较少,算法针对同一对象进行迭代,很快收敛,对数据源可靠度的估计不准确,导致真值发现结果不理想。

与对比算法不同,本文对文本答案进行了细粒度的划分,利用注意力机制学习答案关键词可靠度和答案可靠度,不必对各个数据源观测值与真值的关系进行假设,将这种未知的关系存储在双层注意力机制网络的参数中,与传统方法通过大量观测值估计数据源可靠度的思想相比,在面对稀疏数据时有巨大的优势。同时利用双向 GRU 及 Word2Vec 充分学习文本答案的语义信息,有效解决了文本答

案的自然语言特性为真值发现带来的挑战。

6.4.2 学习率对实验结果的影响

在神经网络的优化过程中,学习率控制参数的更新速度,学习率过小,会极大降低收敛速度,可能陷入局部最优,而学习率过大,则可能导致参数在最优解两侧来回震荡,本节使用 0.01,0.001,1E-4(0.000 1),1E-5(0.000 01)进行实验,验证学习率对实验结果的影响,计算算法得到的排名前 10,30,50,80,100,200 名学生分数的真实平均分,结果如图 3 所示。

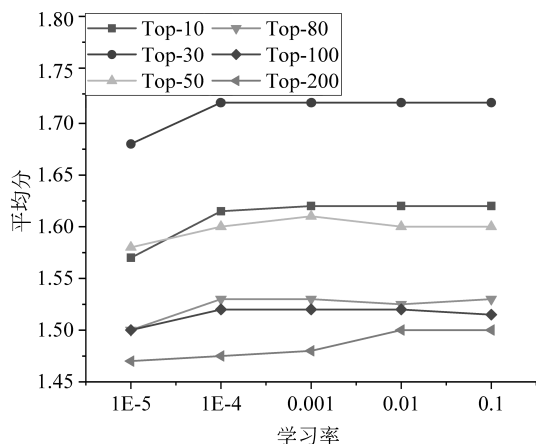


图 3 学习率对实验结果的影响

图 3 中可以发现,模型受学习率的影响较小,部

分数据集在学习率为 1E-5 时效果有所下降。

6.4.3 预处理方法案例研究

为验证预处理步骤的有效性及其意义,从 Short Answer Scoring 数据集中随机选取 100 名用户答案,并构造如图 4 所示的无向图。

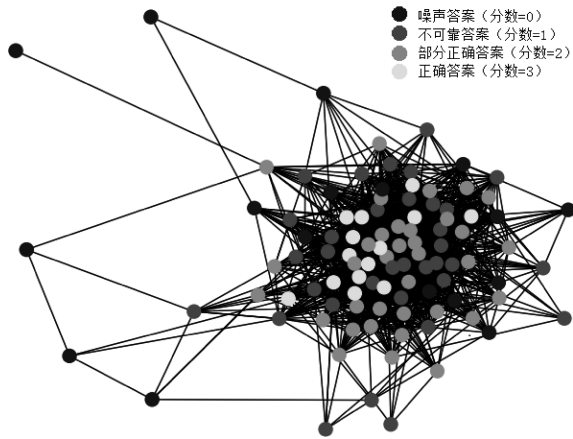


图 4 用户答案无向图

在随机选取的 100 个答案中,有 17 个用户答案评分为 0,是完全错误的答案;34 个答案评分为 1,是不可靠的答案;35 个答案被评分为 2,是部分正确的答案;14 个答案为 3,是完全正确的答案。由图 5 可知,正确答案和部分正确答案由于共享部分类似的关键语义信息,其相似度较大。而错误答案和不可靠的答案不存在相似的语义信息,在图中以孤立点或度较小的节点存在。通过预处理步骤,得到如图 5 所示的用户答案无向图。

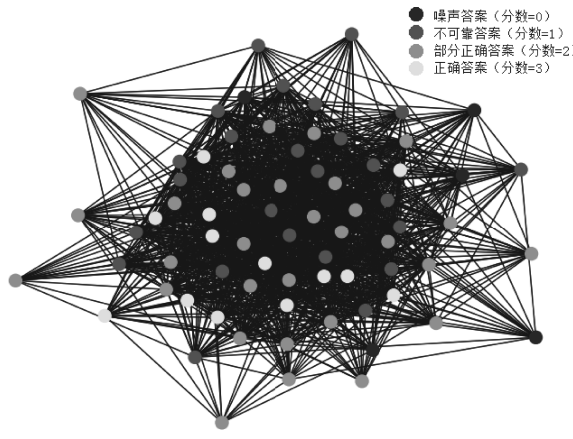


图 5 候选用户答案无向图

由图 5 可知,预处理方法删除了大部分错误答案,保留了全部的正确答案和部分正确答案。保留的答案将作为候选用户答案,输入双层注意力机制神经网络进行文本数据的真值发现。

6.4.4 真值发现方法实例分析

为了更全面阐述文本数据真值发现的过程及方法的有效性,给出数据集关于问题“*What is a variable?*”的实例分析,结果如表 3 所示。

表 3 实例分析

| 问题 | What is a variable? |
|----------|--|
| 标准答案 | A location in memory that can store a value. |
| Top 1 答案 | A variable is a location in memory where a value can be stored. |
| Top 2 答案 | It is a location in the computer's memory where it can be stored for use by a program. |
| Top 3 答案 | A location in memory where data can be stored and retrieved. |
| ... | ... |
| 部分正确答案 | A stored value used by the program. |
| 部分正确答案 | Variable is a way to store different values into the program. |
| ... | ... |
| 错误答案 | Variable can be a integer or a string in a program. |

表 3 中排名靠前的用户答案包含更多标准答案应当具备的关键因素,如“*location、memory 和 store*”,所提方法能够捕捉用户答案的细粒度语义信息,有效解决文本答案的自然语言特性为真值发现带来的挑战,而对用户答案的细粒度可靠度评估,则能找到正确答案应当包含的正确关键因素。其次,从结果也可以看出,排名靠前的答案互相之间具备较高的相似性,而随着用户答案可靠度的下降,答案的相似性也逐步下降,这也说明了本文真值发现假设的正确性。

另一方面,该结果也部分说明了基于检索的方法及真值发现方法 CRH 无法找到真正可靠答案的原因。基于检索方法利用答案与问题的相似性对答案进行排名,而实际上问题并不会包含正确答案应当具备的全部语义信息,而“*variable*”作为高频词,则无法区分不同的答案,无法真正找到可靠的用户答案。同时,可以发现正确答案具备较多的关键因素,CRH 真值发现方法不对用户答案进行细粒度的划分,导致真值发现结果不理想。由此,本文所提方法能够捕捉文本的细粒度语义信息,基于真值发现的假设,利用神经网络寻找答案间的复杂关系,同时找到可靠答案。

7 总结

由于文本数据的自然语言特性,传统真值发现方法无法直接应用于文本数据的真值发现。本文提出基于 Bi-GRU 并包含注意力机制的文本数据真值发现方法,对文本答案进行细粒度语义特征提取,并利用注意力机制学习文本的关键词可靠度及答案可靠度。不同于传统真值发现算法依赖数据源可靠度估计的思路,本文利用答案本身,使用网络无监督学习这种复杂的关系,减少对数据源可靠度估计的依赖。通过实验验证,本算法适用于文本数据真值发现场景,优于基于检索的方法与 CRH 真值发现算法。

当候选答案数量较少时,神经网络无法挖掘答案关联。同时,本文提出的损失函数建立在“大多数人提供正确答案”的假设上,当面对更为复杂的情况时,假设可能不成立。在下一步研究工作中,我们将考虑与应用场景更贴合的情况,提出更健全的损失函数,同时考虑当观测值答案较少时如何挖掘答案间的关联。

参考文献

- [1] Blenholder J, Naumann F. Data fusion[J]. *ACM Computing Surveys*, 2009, 41(1): 1-41.
- [2] Dong X L, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence[C]// *Proceedings of the VLDB Endowment*, 2009, 2(1): 550-561.
- [3] Galland A, Abiteboul S, Marian A, et al. Corroborating information from disagreeing views[C]// *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 2010: 131-140.
- [4] Pasternack J, Roth D. Knowing what to believe (when you already know something)[C]// *Proceedings of International Conference on Computational Linguistics*, 2010: 877-885.
- [5] Yin X, Han J, Yu P S. Truth discovery with multiple conflicting information providers on the web[C]// *Proceedings of the IEEE Transactions on Knowledge and Data Engineering*, 2007: 796-808.
- [6] Aydin B I, Yilmaz Y S, Li Y, et al. Crowd sourcing for multiple-choice question answering [C]// *Proceedings of the 26th IAAI Conference*, 2014.
- [7] Li Q, Li Y, Gao J, et al. A confidence-aware approach for truth discovery on long-tail data[C]// *Proceedings of the VLDB Endowment*, 2014, 8(4): 425-436.
- [8] Li Q, Li Y, Gao J, et al. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation[C]// *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2014: 1187-1198.
- [9] Li Y, Li Q, Gao J, et al. On the discovery of evolving truth[C]// *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015: 675-684.
- [10] Pasternack J, Roth D. Latent credibility analysis[C]// *Proceedings of the 22nd International conference on World Wide Web*, 2013: 1009-1020.
- [11] Zhao B, Han J. A probabilistic model for estimating real-valued truth from conflicting sources[C]// *Proceedings of the VLDB Workshop on Quality in Databases (QDB'12)*, 2012.
- [12] Zhao B, Rubinstein B I, Gemmell J, et al. A bayesian approach to discovering truth from conflicting sources for data integration[J]. *PVLDB*, 2012, 5(6): 550-561.
- [13] Marshall J, Argueta A, Wang D. A neural network approach for truth discovery in social sensing[C]// *Proceedings of the IEEE International Conference on Mobile Ad Hoc and Sensor Systems*, 2017: 343-347.
- [14] Broelemann K, Gottron T, Kasneci G. Restricted Boltzmann machines for robust and fast latent truth discovery[J]. *arXiv preprint arXiv: 1801.00283v1*, 2017.
- [15] Broelemann K, Kasneci G. Combining restricted boltzmann machines with neural networks for latent truth discovery[J]. *arXiv preprint arXiv: 1807.10680v1*, 2018.
- [16] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence[J]. *Neural Computation*, 2002, 14(8): 1771-1800.
- [17] Li L, Qin B, Ren W, et al. Truth discovery with memory network[J]. *Tsinghua Science Technology*, 2017, 22(6): 609-618.
- [18] Popat K, Mukherjee S, Weikum G. Credibility assessment of textual claims on the web[C]// *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, 2016: 2173-2178.
- [19] Ma F, Li Y, Li Q, et al. Fatercrowd: Fine grained truth discovery for crowdsourced data aggregation[C]// *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015: 745-754.
- [20] Ma F, Meng C, Xiao H, et al. Unsupervised discovery of drug side-effects from heterogeneous data sources [C]// *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017: 967-976.
- [21] Li Y, Du N, Liu C, et al. Reliable medical diagnosis from crowdsourcing: Discover trustworthy answers

- from non-experts [C]//Proceedings of the WSDM 2017, 2017: 253-261.
- [22] Zhang H, Li Y, Ma F, et al. TextTruth: An unsupervised approach to discover trustworthy information from multi-sourced text data [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery, 2018: 2729-2737.
- [23] Bouguessa M, Dumoulin B, Wang S. Identifying authoritative actors in question-answering forums: The case of yahoo! answers [C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008: 866-874.
- [24] Hong L, Davison B D. A classification-based approach to question answering in discussion boards [C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009: 171-178.
- [25] Yang L, Qiu M, Gottipati S, et al. Cqarank: Jointly model topics and expertise in community question answering [C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013: 99-108.
- [26] Zhou G, Lai S, Liu K, et al. Topic-sensitive probabilistic model for expert finding in question answer communities [C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012: 1662-1666.
- [27] Wang Z, Mi H, Ittycheriah A. Sentence similarity learning by lexical decomposition and composition [J], arXiv preprint arXiv: 1602.07019, 2016.
- [28] Yao X, Van Durme B, Callison-Burch C, et al. Answer extraction as sequence tagging with tree edit distance [C]//Proceedings of the 2013 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, 2013: 858-867.
- [29] Feng M, Xiang B, Glass M R, et al. Applying deep learning to answer selection: A study and an open task [C]//Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015: 813-820.
- [30] Wang D, Nyberg E. A long short-term memory model for answer sentence selection in question answering [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 707-712.
- [31] Tan M, Santos C D, Xiang B, et al. Lstm-based deep learning models for non-factoid answer selection [J]. arXiv preprint arXiv: 1511.04108v2 (2016), 2016.
- [32] Cicero D, Ming T, Bing X, et al. Attentive pooling networks [J]. arXiv preprint arXiv: 1602.03609 (2016), 2016.
- [33] Liu Q, Huang Z, Huang Z, et al. Finding similar exercises in online education systems [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018: 1821-1830.
- [34] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 1480-1489.
- [35] Arora S, Liang Y, Ma T. A simple but tough-to-beat baseline for sentence embeddings [C]//Proceedings of the Int'l Conf. on Learning Representations, Toulon, France, 2017.
- [36] 马如霞, 孟小峰, 王璐, 等. MTruths: Web 信息多真值发现方法 [J]. 计算机研究与发展, 2016, 52 (12): 2858-2866.
- [37] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C]//Proceedings of the ICLR, 2014: abs/1409-0743.



常宸(1994—), 硕士研究生, 主要研究领域为数据质量、冲突消解。
E-mail: c308051252@163.com



吕国俊(1995—), 硕士研究生, 主要研究领域为跨模态实体分辨。
E-mail: l1983286838@outlook.com



曹建军(1975—), 博士, 副研究员, 主要研究领域为数据质量控制、数据智能分析与应用。
E-mail: jianjuncao@yeah.com