基于网络表示的半监督问答文本情感分类方法

陈　潇[1]，　李逸薇[2]，　刘　欢[1]，　李寿山[1]

(1. 苏州大学 计算机科学与技术学院　江苏 苏州 215000；
2. 香港理工大学 人文学院 中文及双语系　香港 999077)

摘要：针对新颖的问答形式的文本展开研究，提出了基于网络表示的半监督问答文本情感分类方法，通过构建异构网络的联合学习提升半监督问答文本的情感分类性能。首先，通过分析标注和未标注样本构建一个异构网络，具体包括词-词网络、问题和答案文本-词网络、情感标签-词网络；其次，利用该异构网络学习获得词向量；最后，将学习到的词向量应用于目前性能最优的分层匹配情感分类模型(hierarchical matching network，HMN)中。实验结果表明，提出的方法在处理问答文本情感分类任务上具有一定优势。

关键词：情感分类；半监督；网络表示；问答文本

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1671-6841(2020)02-0052-07

DOI：10.13705/j.issn.1671-6841.2019079

0　引言

在电子商务环境中，新出现的面向用户的问答型评论方式很大程度上避免了虚假评论，对这些高质量、高可信度的问答文本进行情感分类具有很高的实际应用价值，如商品推荐、企业决策等。

目前，问答文本的情感分类任务主要面临两个挑战。第一个挑战是大量已标注的问答文本较难获取。另一个挑战是已有的情感分类方法大多是针对普通文本，不适用于问答文本的情感分类任务。

对自然语言处理领域的任务而言，一个有意义并且高效的词向量发挥着重要的作用。传统的词向量学习方法如 one-hot 模型、分布式词向量表示等，并不完全适用于情感分类任务。而且，相比于一般文本，问答文本包含“问题”和“答案”两部分，结构更加复杂，对词向量学习的要求也更高。

为了让表达相同情感的词学习相近的词向量表示，一个简单直接的方法就是构建词-词网络和文本-词网络，利用词与词之间的共现关系以及文本与词之间的共现关系一起学习词向量。然而，当只有少量标注问答文本时，很多词与词、文本与词的关系都无法捕捉，因此，本文加入大量未标注问答文本增强词-词网络和文本-词网络。此外，为了充分利用词的情感信息，我们加入了情感标签-词的网络，使具有相同情感标签的文本中的词更有可能表达相同的情感。

本文提出了一种基于网络表示的半监督问答文本情感分类方法。首先，构建了一个由词-词网络、文本-词网络和情感标签-词网络组合而成的异构网络，并提出了在此异构网络上利用已标注问答文本和未标注问答文本共同学习词向量的方法。接着，在得到词向量之后，将其应用到目前性能最优的问答文本情感分类模型[1](hierarchical matching network，HMN)中进行问答文本的情感分类。

1　相关工作

情感分类是自然语言处理领域的一项基础任务，主要目的是对文本中包含的情感倾向进行自动分类，这项任务已经开展了较长时间，取得了很多成果。李钝等[2]提出了中心词概念，对各词的倾向性进行计算来识别短语的倾向性和倾向强度。田胜利等[3]提出 LSI 与 KNN-naive Bayes 结合的分类模型，对中文网页作者

收稿日期：2019-03-25
基金项目：国家自然科学基金项目(61331011，61375073)。
作者简介：陈潇(1995—)，女，江苏泰兴人，硕士研究生，主要从事自然语言处理研究，E-mail：431702236@qq.com；通信作者：李寿山(1980—)，男，江苏扬州人，教授，主要从事情感分析、自然语言处理研究，E-mail：lishoushan@suda.edu.cn。

的情感态度进行分类。闻彬等[4]提出了一种基于语义理解的文本情感分类方法。庞磊等[5]提出了基于情绪知识的非监督情感分类方法,对微博文本进行情感极性自动分类。

由于自然语言固有的复杂性,在情感分类等任务中通常需要将文本中的词语转换为低维度的数字向量表示,表示的好坏会直接影响到最终结果。本文从文本表示着手,提高问答文本情感分类的性能。最早出现的 one-hot 向量形式非常简单,但它无法表示词与词之间的语义关联。1954 年,Harris 提出了“分布假说”[6],基于这一理论,产生了分布式表示方法,如 Mikolov 等[7]提出的 skip-gram 模型,很好地弥补了 one-hot 的不足。在这些基础上,蒋振超等[8]提出了一种新的基于关系的无监督词向量表示模型。刘晓蕾[9]将文本本身作为模型的上下文,实现文本信息与局部上下文信息的整合,提出了基于强化语义的词向量学习模型。吴旭康等[10]提出了主题联合词向量模型。

不同于以上研究,本文针对问答文本情感分类这一特定任务,使用半监督的方法,在搭建的异构网络上学习词向量。实验结果表明,我们的方法在性能上优于通用的词向量算法。

2　语料收集与分析

从淘宝购物平台的“问大家”板块中,我们收集了 21 万条电子产品领域的问答评论文本,并在这 21 万条样本中随机选取了 1 万条样本进行情感标注。我们主要将文本分为正面、负面、中性和冲突 4 个情感类别,其中,中性是指不包含明显的情感倾向,冲突是指既包含正面情感又包含负面情感。样本标注完成后情感类别的分布情况为:正面情感样本 3 807 条;负面情感样本 1 017 条;中性情感样本 4 648 条;冲突情感样本 528 条。

3　基于网络表示的半监督问答文本情感分类方法

3.1　整体模型

本文提出的基于网络表示的半监督问答文本情感分类方法主要包括两个部分:第一部分是表示学习;第二部分是情感分类。具体模型结构如图 1 所示。

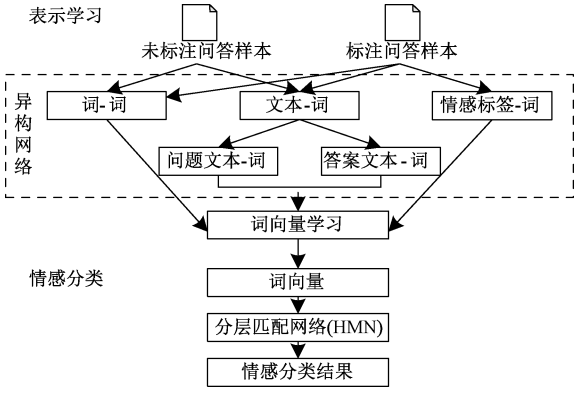

图 1　基于网络表示的半监督问答文本情感分类方法模型框架图

Figure 1　The architecture of the semi-supervised sentiment classification method towards QA text based on network representation

在表示学习的部分,我们构造了一个异构网络,这个异构网络由词-词网络、文本-词网络和情感标签-词网络这 3 个不同类型的网络组成,其中,文本-词网络又分为问题文本-词网络与答案文本-词网络。在异构网络中,我们使用半监督的方法学习问答文本的表示形式,在得到问答文本的词向量表示以后,我们将其应用于 HMN 模型中。

3.2　网络表示方法

3.2.1　词-词网络　本小节主要介绍用于捕捉问答文本中词与词之间共现关系的词-词网络 G_{vv}。

定义网络 $G_{vv}=(V,E_{vv})$,其中:V 表示词的集合;E_{vv} 表示词与词之间的边集合。词 v_i 和词 v_j 之间的边的权重 w_{ij} 定义为词 v_i 和词 v_j 在所有文本中固定大小的窗口中一起出现的次数。

对于词-词网络 G_{vv}，为了将每个词表示为一个低维向量，我们最小化目标函数

$$O_{vv} = -\sum_{(v_i,v_j)\in E_{vv}} w_{ij}\log p(v_j \mid v_i), \tag{1}$$

其中：$(v_i,v_j) \in E_{vv}$ 代表词结点 v_i 和词结点 v_j 之间的边；w_{ij} 表示边 $(v_i,v_j) \in E_{vv}$ 的权重；$p(v_j \mid v_i)$ 是指由词结点 v_i 生成词结点 v_j 的条件概率。

条件概率定义为

$$p(v_j \mid v_i) = \frac{\exp(\boldsymbol{u}_j^{\mathrm{T}} \cdot \boldsymbol{u}_i)}{\sum_{k=1}^{|V|}\exp(\boldsymbol{u}_k^{\mathrm{T}} \cdot \boldsymbol{u}_i)}, \tag{2}$$

其中：$\boldsymbol{u}_i$ 和 $\boldsymbol{u}_j$ 分别表示结点 v_i 和结点 v_j 的向量；$|V|$ 表示网络中结点的总数。

3.2.2 文本-词网络 本小节主要介绍用于捕捉问答文本中文本和词之间共现关系的文本-词网络 G_{dv}。由于问答文本包括“问题”和“答案”两部分，因此我们构建了相应的问题文本-词网络 G_{qv} 和答案文本-词网络 G_{av}。

1）问题文本-词网络 G_{qv} 如图2所示。

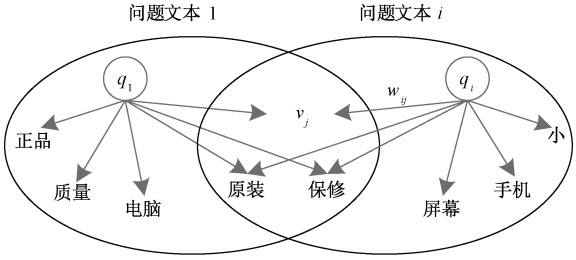

图2 问题文本-词网络

Figure 2 Question-word network

定义 $G_{qv} = (Q \cup V, E_{qv})$，其中：$Q$ 表示问题文本的集合；V 表示词的集合；E_{qv} 表示问题文本和词之间的边集合。问题文本 q_i 和词 v_j 之间的边的权重 w_{ij} 定义为词 v_j 出现在问题文本 q_i 中的次数。

对问题文本-词网络 G_{qv}，为了更好捕捉问题文本和词的共现关系，我们最小化目标函数

$$O_{qv} = -\sum_{(q_i,v_j)\in E_{qv}} w_{ij}\log p(v_j \mid q_i), \tag{3}$$

其中：$(q_i,v_j) \in E_{qv}$ 代表问题文本结点 q_i 和词结点 v_j 之间的边；w_{ij} 表示边 $(q_i,v_j) \in E_{qv}$ 的权重；$p(v_j \mid q_i)$ 是指由问题文本结点 q_i 生成词结点 v_j 的条件概率，条件概率的计算如公式(2)所示。

2）答案文本-词网络 G_{av} 如图3所示。

定义 $G_{av} = (A \cup V, E_{av})$，其中：$A$ 表示答案文本的集合；V 表示词的集合；E_{av} 表示答案文本和词之间的边集合。答案文本 a_i 和词 v_j 之间的边的权重 w_{ij} 定义为词 v_j 出现在答案文本 a_i 中的次数。

对答案文本-词网络 G_{av}，为了更好地捕捉答案文本和词的共现关系，我们最小化目标函数

$$O_{av} = -\sum_{(a_i,v_j)\in E_{av}} w_{ij}\log p(v_j \mid a_i), \tag{4}$$

其中：$(a_i,v_j) \in E_{av}$ 代表答案文本结点 a_i 和词结点 v_j 之间的边；w_{ij} 表示边 $(a_i,v_j) \in E_{av}$ 的权重；$p(v_j \mid a_i)$ 是指由答案文本结点 a_i 生成词结点 v_j 的条件概率，条件概率的计算如公式(2)所示。

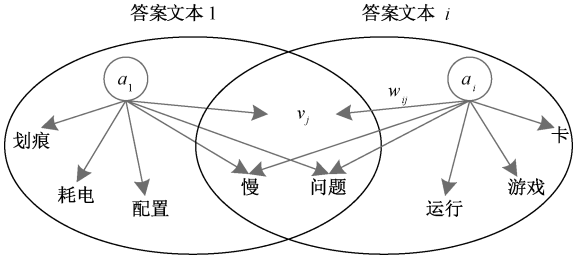

图3 答案文本-词网络

Figure 3 Answering-word network

对于由问题文本-词网络 G_{qv} 和答案文本-词网络 G_{av} 构成的文本-词网络 G_{dv},其目标函数定义为

$$O_{dv} = O_{qv} + O_{av}。\tag{5}$$

3.2.3 情感标签-词网络　本小节主要介绍用于捕捉问答文本中情感标签和词之间共现关系的情感标签-词网络 G_{sv},如图 4 所示。不同于词-词网络 G_{vv} 和文本-词网络 G_{dv},情感标签-词网络 G_{sv} 在对共现信息编码的过程中使用了情感信息,使得整体模型更适用于情感分类任务。

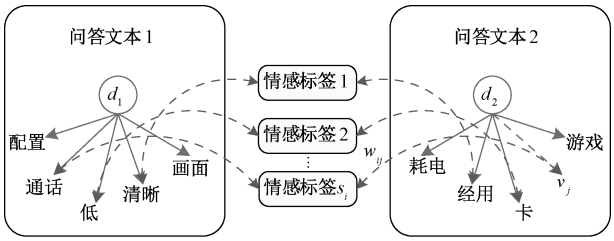

图 4　情感标签-词网络

Figure 4　Sentiment-word network

定义 $G_{sv} = (S \cup V, E_{sv})$,其中:$S$ 表示问答文本情感类别的集合;V 表示词的集合;E_{sv} 表示情感标签和词之间的边集合。情感标签 s_i 与词 v_j 之间的边的权重 w_{ij} 定义为词 v_j 出现在情感标签为 s_i 的文本中的总次数,即 $w_{ij} = \sum_{d:\text{文本}d\text{的标签为}s_i} n_{d_k}$,其中 n_{d_k} 表示词 v_j 在文本 d_k 中出现的次数。

对情感标签-词网络 G_{sv},为了更好地捕捉情感标签与词的共现关系,我们最小化目标函数

$$O_{sv} = -\sum_{(s_i,v_j)\in E_{sv}} w_{ij}\log p(v_j \mid s_i),\tag{6}$$

其中:$(s_i, v_j) \in E_{sv}$ 表示情感标签结点 s_i 和词结点 v_j 之间的边;w_{ij} 表示 $(s_i, v_j) \in E_{sv}$ 的权重;$p(v_j \mid s_i)$ 是指由情感标签结点 s_i 生成词结点 v_j 的条件概率,条件概率的计算如公式(2)所示。

3.2.4 异构网络　异构网络 G_{joint} 由词-词网络 G_{vv}、文本-词网络 G_{dv} 和情感标签-词网络 G_{sv} 共同构成。由于本文训练的异构网络使用了情感标签信息,而未标注问答样本没有情感标签,因此,G_{vv} 和 G_{dv} 由已标注问答样本和未标注问答样本共同训练,G_{sv} 则仅由已标注问答样本训练。在这个异构网络上学习词向量的直接方式就是同时训练构成它的 3 个网络,即最小化目标函数

$$O_{\text{joint}} = O_{vv} + O_{dv} + O_{sv} = O_{vv} + O_{qv} + O_{av} + O_{sv}。\tag{7}$$

对于公式(1)、(3)、(4)、(6),本文采用异步随机梯度下降的方法[11]进行优化,其中,异步随机梯度下降使用了边采样技术[12]和负采样技术[13]。在训练异构网络的每一个步骤中,我们均选取一条边 (v_i, v_j),选取的概率与边的权重 w_{ij} 成正比,与此同时,我们使用 Mikolov 等[13]提出的方法,根据噪声分布选取多条负边,然后将选取的边 (v_i, v_j) 和多条负边共同用于词向量的更新。

在词-词网络 G_{vv}、文本-词网络 G_{dv}(问题文本-词网络 G_{qv}、答案文本-词网络 G_{av})和情感标签-词网络 G_{sv} 中,我们根据不同网络的特殊性定义了 4 种含义的边的权重值。由于这 4 种网络中边的权重不能直接相互比较,所以在使用联合学习的方法优化目标函数(7)时,我们没有简单地将所有边混合在一起,而是交替地从 4 个边的集合中取样。

4　实验结果与分析

4.1　实验设置

本文使用的语料包括 1 万条已标注问答文本和 20 万条未标注问答文本,具体语料来源和标注规范如第 2 节所示。我们采用 FudanNLP(https://github.com/FudanNLP/fnlp/)工具对这 21 万条问答文本进行分词。在实验中,我们采用了传统词向量表示模型和本文提出的网络表示方式,分别对问答文本进行表示。

1) Word2Vec:我们使用 gensim(https://radimrehurek.com/gensim)实现了 skip-gram 模型,利用 1 万条已标注问答文本和 20 万条未标注问答文本在 skip-gram 模型上学习词的分布式表示,得到 Word2Vec 词向量。Word2Vec 词向量的维度为 100,窗口大小设置为 5。

2) 网络表示方式:我们将1万条已标注问答文本和20万条未标注问答文本作为输入,训练一个异构网络来获取joint词向量。具体实验中,我们抽取1万条已标注样本中70%的问答文本作为训练集,10%作为验证集,20%作为测试集。因此,在学习joint词向量的过程中,用验证集和测试集的3 000条问答文本当作未标注样本处理。joint词向量的维度为100,负边样本数为5。

本文采用正确率(*accuracy*,*Acc*)、精确率(*precision*,*P*)、召回率(*recall*,*R*)和综合评价指标*F*1值(*F*1)作为衡量分类效果的标准,其中,*P*、*R*、*F*1值采用各类别*P*、*R*、*F*1值的平均值。

4.2 实验结果与分析

为了更好地分析本文提出方法的有效性,我们选取了其他几种情感分类方法进行比较。

LSTM:使用标准的长短期记忆神经网络(long-short term memory,LSTM)模型对问答语料进行情感分类,包括一个LSTM层、一个全连接层和一个dropout层。LSTM采用Word2Vec词向量进行全监督学习的情感分类,即使用已标注问答样本。

HMN:Shen等[1]提出的分层匹配情感分类模型采用全监督学习的方式,使用已标注问答样本训练的Word2Vec词向量,是目前性能最优的问答文本情感分类模型。

co-training:协同训练算法,其基本思想是利用多个独立视图训练多个分类器,然后采用互助方式迭代地扩充已标注样本并重新训练分类器。实验中,我们采用Word2Vec词向量。

self-training:传统的自训练算法,其基本思想是使用少量标注样本构建单个分类器,然后迭代地预测未标注样本的标签,通过预测的置信度进行排序,并将置信度高的未标注样本和其预测的标签永久地添加到标注样本中,是一种增量算法[14]。实验中,我们采用Word2Vec词向量。

VAE:采用变分自编码器(variational autoencoder, VAE)[15]实现半监督情感分类,联合学习变分自编码器和分类器的损失函数提升半监督情感分类任务的性能[16]。实验中,我们采用Word2Vec词向量。

对抗学习(adversarial learning, AL)采用基于AL的半监督情感分类方法,包括两个编码器、一个分类器和一个判别器。两个编码器使用标准的LSTM神经网络,分别将标注样本和未标注样本映射到代码空间。分类器使用激活函数是softmax的全连接神经网络,预测文本的情感倾向。判别器同样使用激活函数是softmax的全连接神经网络,判断输入文本是属于标注样本还是属于未标注样本。实验中,我们采用Word2Vec词向量。

HMN(joint):该方法为第3节提出的基于网络表示的半监督问答文本情感分类方法,将得到的joint词向量应用于HMN模型中。

表1给出了上述几种方法在问答文本情感分类任务上的实验性能。同时,为了更好地验证实验结果,我们分别采用了1万、5万、10万、15万和20万的未标注问答文本进行了实验,实验的正确率如图5所示。

表1 不同方法在问答文本情感分类任务上的性能(未标注样本数量为20万)

Table 1 The performance of different methods in question-answering text sentiment classification task (The unlabeled samples were 200 000)

方法	*P*	*R*	*F*1	*Acc*
LSTM	0.570	0.556	0.562	0.715
HMN	0.652	0.632	0.640	0.779
co-training	0.413	0.411	0.356	0.420
self-training	0.450	0.427	0.401	0.494
VAE	0.619	0.580	0.595	0.735
AL	0.655	0.586	0.609	0.755
HMN(joint)	0.777	0.753	0.764	0.853

表1的实验结果中,LSTM和HMN方法是全监督学习的情感分类方法,其余5种方法为半监督学习的情感分类方法。图5中所有方法均为半监督情感分类方法。

通过比较实验,我们可以得出以下结论。

1) co-training和self-training两种方法使用分类器对未标注样本进行分类,因此产生的错误标注样本对情感分类任务产生噪音从而降低实验性能,并且随着未标注样本数量的增加,co-training和self-training两种

方法的性能都呈下降趋势。co-training 和 self-training 两种方法对问答文本进行情感分类的结果显示 P、R、$F1$ 值较低,主要原因是问答样本的情感类别分布不均匀,其中,冲突类别的样本数量为 528,而中性类别的样本数量为 4 648,样本的不平衡也增加了 co-training 和 self-training 两种方法在对未标注样本进行分类时出错的概率,从而降低整体情感分类的正确率。

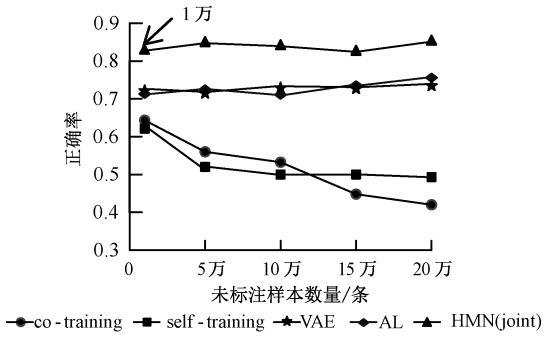

图 5　不同方法在不同未标注问答样本数量上的正确率

Figure 5　The accuracy of different methods in different sizes of unlabeled samples

2) VAE 和 AL 两种方法使用深度学习对问答文本进行半监督情感分类,且在情感分类中不对未标注样本进行标注,即不会产生错误标注样本影响实验性能。在未标注样本数目为 20 万时,VAE 和 AL 两种方法的 P、R、$F1$ 和 Acc 明显高于 co-training 和 self-training 这两种方法,说明 VAE 和 AL 两种方法利用深度学习可以缓和数据不平衡的问题,但这两种方法没有考虑到问答文本的特殊性,因此实验结果没有 HMN(joint)高。

3) 本文提出的 HMN(joint)方法在所有实验中正确率是最高的。在未标注问答样本数目为 20 万的情况下,在基于全监督学习的情感分类方法中,HMN(joint)方法比 LSTM 方法提高 13.8%,比 HMN 方法提高 7.4%。在基于半监督学习的情感分类方法中,HMN(joint)方法比 VAE 和 AL 两种方法平均提高 10.8%,比 co-training 和 self-training 两种方法平均提高 39.6%。实验结果表明,我们提出的 HMN(joint)方法在构建异构网络时能够有效处理未标注样本的信息,且在未标注样本数量不同的情况下能够保持稳定的正确率。因此,本文提出的方法在问答文本上具有较强的情感分类能力。

5　结论

本文提出了一种基于网络表示的半监督问答文本情感分类方法。具体而言,首先构建一个由标注样本的词-词网络、文本-词网络、情感标签-词网络和未标注样本的词-词网络、文本-词网络组成的异构网络,并根据问答文本形式的特殊性,将文本-词网络分为问题文本-词网络和答案文本-词网络,然后采用异步随机梯度下降的方法联合学习这个异构网络,最后将得到的词向量应用于分层匹配网络。实验结果表明,本文提出的方法能够有效提高问答文本情感分类的实验性能。

在下一步研究工作中,我们尝试将通过标注样本和未标注样本学习得到的词向量与其他半监督情感分类方法进行融合,解决已标注问答样本匮乏的问题。

参考文献:

[1] SHEN C L, SUN C L, WANG J J, et al. Sentiment classification towards question-answering with hierarchical matching network [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, 2018: 3654-3663

[2] 李钝, 曹付元, 曹元大, 等. 基于短语模式的文本情感分类研究[J]. 计算机科学, 2008, 35(4): 132-134.
LI D, CAO F Y, CAO Y D, et al. Text sentiment classification based on phrase patterns[J]. Computer science, 2008, 35(4): 132-134.

[3] 田胜利,熊德兰. 中文网页作者情感态度倾向性分类研究[J]. 信阳师范学院学报(自然科学版),2009,22(2):307-309.
TIAN S L, XIONG D L. Research on emotion attitude orientation classification of Chinese web page author[J]. Journal of Xinyang normal university (natural science edition), 2009,22(2):307-309.

[4] 闻彬, 何婷婷, 罗乐, 等. 基于语义理解的文本情感分类方法研究[J]. 计算机科学, 2010, 37(6): 261-264.
WEN B, HE T T, LUO L, et al. Text sentiment classification research based on semantic comprehension[J]. Computer science, 2010, 37(6): 261-264.

[5] 庞磊, 李寿山, 周国栋. 基于情绪知识的中文微博情感分类方法[J]. 计算机工程, 2012, 38(13): 156-158.
PANG L, LI S S, ZHOU G D. Sentiment classification method of Chinese micro-blog based on emotional knowledge[J]. Com-

puter engineering, 2012, 38(13): 156-158.
[6] HARRIS Z S. Distributional structure [J]. Word, 1954, 10(2/3): 146-162.
[7] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, 2013: 3111-3119.
[8] 蒋振超, 李丽双, 黄德根. 基于词语关系的词向量模型[J]. 中文信息学报, 2017, 31(3): 25-31.
JIANG Z C, LI L S, HUANG D G. Word representation based on word relations[J]. Journal of Chinese information processing, 2017, 31(3): 25-31.
[9] 刘晓蕾. 面向情感分析的词向量学习及其应用[D]. 北京: 中国科学院大学, 2016.
LIU X L. Word vector learning with its application for sentiment analysis[D]. Beijing: University of Chinese Academy of Science, 2016.
[10] 吴旭康, 杨旭光, 陈园园, 等. 主题联合词向量模型[J]. 计算机工程, 2018, 44(2): 233-237.
WU X K, YANG X G, CHEN Y Y, et al. Topic combined word vector model[J]. Computer engineering, 2018, 44(2): 233-237.
[11] NIU F, RECHT B, RE C, et al. HOGWILD!: a lock-free approach to parallelizing stochastic gradient descent [J]. Advances in neural information processing systems, 2011, 24: 693-701.
[12] TANG J, QU M, WANG M, et al. LINE: large-scale information network embedding [C]//Proceedings of the 24th International Conference on World Wide Web. Florence, 2015: 1067-1077.
[13] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2014-02-10)[2019-01-05]. http://arxiv.org/pdf/1301.3781.pdf.
[14] HADY M F A, SCHWENKER F. Semi-supervised learning[C]//Proceedings of the 20th International Conference on Neural Information Processing. Daegu, 2013:215-239.
[15] XU W, SUN H, DENG C, et al. Variational autoencoder for semi-supervised text classification [C]//Proceedings of the 31th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence. San Francisco, 2017: 3358-3364.
[16] 刘欢, 徐健, 李寿山. 基于变分自编码器的情感回归半监督领域适应方法[J]. 郑州大学学报(理学版), 2019, 51(2): 47-51.
LIU H, XU J, LI S S. A semi-supervised domain adaptation method of sentiment regression on variational autoencoder[J]. Journal of Zhengzhou university(natural science edition), 2019, 51(2): 47-51.

A Semi-supervised Sentiment Classification Method Towards Question-answering Text Based on Network Representation

CHEN Xiao[1], LEE Sophia[2], LIU Huan[1], LI Shoushan[1]
(1. *School of Computer Science and Technology, Soochow University, Suzhou* 215000, *China*; 2. *Department of Chinese & Bilingual Studies, Faculty of Humanities, Hong Kong Polytechnic University, Hong Kong* 999077, *China*)

Abstract: A semi-supervised sentiment classification method towards question-answering text based on network representation was proposed. And the performance of semi-supervised sentiment classification of question-answering text was improved by constructing joint learning of heterogeneous network. A heterogeneous network was firstly constructed by analyzing labeled and unlabeled samples, which was composed of word-word network, question and answering document-word network, and sentiment-word network. Secondly, the heterogeneous network was used to learn word embedding. Finally, the word embedding was applied to the currently best-performing hierarchical matching network. Empirical results showed that the proposed method had certain advantages in processing the sentiment classification task on question-answering text.

Key words: sentiment classification; semi-supervised; network representation; question-answering text

(责任编辑:王浩毅)