基于深度迁移网络的 Twitter 谣言检测研究*

刘　勘　杜好宸
(中南财经政法大学信息与安全工程学院　武汉　430073)

摘要: 【**目的**】解决网络谣言分领域检测时某些领域标注数据不足的问题，帮助在无标注数据的领域构建谣言检测模型。【**方法**】提出一种深度迁移网络，以 Multi-BiLSTM 网络为基础，加入 MMD 统计量计算的领域分布差异，训练过程中同时学习源领域的标签损失与领域间的分布差异，完成标签信息在领域间的有效迁移。【**结果**】相较于未分领域的谣言检测方法和分领域但不使用迁移学习的谣言检测方法，本文方法在 F1 指标上分别提升 10.3%与 8.5%。【**局限**】在数据分布差异大的领域迁移效果受到限制，未涉及多个领域的谣言检测。【**结论**】本文方法可以有效地将迁移学习技术应用在分领域谣言检测场景下，为网络谣言识别提供新思路。

关键词: 谣言检测　深度迁移网络　多层双向长短时记忆网络　领域适配　推特
分类号: TP393
DOI: 10.11925/infotech.2096-3467.2018.1250

1　引　言

Web2.0 时代，以社交网站、微博、博客、论坛为代表的社交媒体成为互联网世界中最大的信息产生与交换渠道[1]，同时也为谣言的生成与传播构筑了温床。社会学家将谣言定义为：没有相应事实基础，却被捏造出来并通过一定手段推动传播的言论[2]。相较于社会谣言，网络中的谣言常常以吸引流量或制造恐慌为目的，因而往往带有较大的煽动性与恶意性，成为破坏网络空间秩序与环境的毒瘤，对社会生活造成极大的负面影响[3]。因此对网络谣言进行有效鉴别，具有较为紧迫的社会意义、经济意义与现实意义。

目前国内外针对网络谣言的检测主要依赖于一些公司或公益性组织建立的辟谣平台，其中国际包括 Snopes.com、FactCheck.org、UrbanLegends.about.com、Emergent 等网站，国内主要有中国互联网联合辟谣平台(http://www.piyao.org.cn/,中央网信办)、微博辟谣(https://weibo.com/weibopiyao，新浪微博)等渠道。这些渠道通过调查溯源等手工方式对网络谣言进行检测，在取得较高准确度的同时也需要很大的时间开销，尤其是当面对海量的网络消息时，人工鉴谣就显得力不从心。因此，网络谣言的自动检测也是谣言分析的热点问题之一。Twitter 是全球范围内具有深远影响力与代表性的社交媒体，以 Twitter 为对象开展网络谣言的自动检测研究对于网络谣言的治理具有广泛借鉴性。

2　文献综述

网络谣言的自动检测最早开始于 Twitter，众多学者从识别谣言的要素入手，构造相关特征实现对网络谣言的检测。Castillo 等[4]归纳了来自 4 个方面的要素以识别谣言，分别是文本特征、用户特征、传播特征和话题特征，在此基础上总结出 15 项关键

通讯作者：刘勘, OCRID: 0000-0001-9339-7315, E-mail: liukan@zuel.edu.cn。
*本文系国家社会科学基金项目“基于文本挖掘的网络谣言预判研究”(项目编号: 14BXW033)的研究成果之一。

特征并利用决策树算法 J48 实现对谣言的检测。Ma 等[5]考虑谣言演变的时间特征，使用动态时间序列模型对谣言进行检测。Zhao 等[6]通过构建线索词等特征实现了谣言的早期预测。祖坤琳等[7]则关注微博的评论消息，将微博评论的情感倾向性加入模型，实现了谣言的有效检测。这些方法虽然取得一定效果，但是大多依赖于手工构造特征，也属于谣言识别要素的浅层提取，因而无法进一步提升准确率。

近年，很多学者尝试使用深度学习模型自动构造深层特征实现谣言检测。Ma 等[8]尝试使用 RNN 及其衍生模型对谣言事件进行检测，抽取相关的 Tweet 组成谣言事件，然后通过词嵌入提取谣言事件的特征，最后借助 RNN 网络实现对谣言事件的检测，实验证明了 RNN 在谣言检测领域的有效性。Chen 等[9]在此基础上，加入注意力机制，实现对 Tweet 特征的深层抽取，取得了较好效果。深度学习模型的构建依赖于大量标注数据，而网络谣言的数据获取一直是谣言检测领域的一大难题，因此数据标注问题成为基于深度学习模型进行谣言检测的最大瓶颈。一些学者尝试避开数据的标注，借助无监督学习的思想进行网络谣言检测。Zhang 等[10]提出一种无监督的谣言检测方式，借助于多层自编码器获取谣言的文本编码规则，实现对谣言的有效检测。Chen 等[11]在自编码器的前端加入多层 RNN 网络，进一步提升了模型的效果。无监督学习的方法虽然避开了数据标注的问题，但是模型的不稳定性会带来较大局限。

无论使用传统模型还是深度学习模型，通过对文本信息、用户信息、传播信息的加工可有效识别网络谣言[12]，借助深度网络自动构建特征进一步提升了谣言检测的准确性与稳定性。然而网络谣言涉及众多领域，比如经济类谣言、社会类谣言、政治类谣言等。不同领域具有不同的特征，但现有的谣言检测研究缺乏对领域的关注，常用统一的检测模型应对各个领域的谣言。这样的方法忽视了领域差异对谣言检测的影响，导致识别准确率低。针对网络谣言的分领域问题，也可以利用已有方法进行识别，但有些领域数据量较少会导致识别结果不佳。本文将这一问题视为领域数据标注不足的问题，利用迁移学习的思想来探寻其中的解决方案。

3 问题分析

3.1 网络谣言的领域差异

Twitter 仿照传统媒体的分类标准，根据来源与内容将平台内的消息划分为不同领域，如社会类、军事类、娱乐类等①，从而使具有不同偏好的用户方便地获取自己喜欢的信息。从新闻学角度出发，不同领域的信息具有不同的描述对象、描述技巧和描述词汇，这一点也反映在社交媒体的文本中。

网络谣言的产生往往从真实事件出发，对事实进行篡改、夸张与嫁接，因而网络谣言也会涉及不同的领域，具有不同的语言特征。如娱乐类的谣言往往充斥着“私生子”、“隐婚”、“潜规则”等字眼，而生活类的谣言则与健康、日常饮食、养生等话题相关。从计算语义学的角度看，这些领域的差异可以被理解为不同领域的信息具有不同数据分布，因此不同领域的谣言检测方法也不尽相同。

传统的网络谣言检测方法使用机器学习算法建模时，常常将具有不同数据分布的各个领域信息作为一个整体去构建分类器，这样建立的分类器往往缺乏领域针对性，导致某些领域的网络谣言不能被很好地鉴别。因而需要针对各个领域分别构建分类器，从而捕捉特定领域的关键特征，实现精细化检测，也可以有效避免来自领域高频词的误导。比如食品类谣言中经常出现的“致癌”、“癌症”、“细菌”等词，在医学领域则作为正常词出现。

事实上网络谣言在各领域的分布是极不平均的。以国外辟谣网站 Snopes②为例，2018 年第三季度披露的网络谣言共 676 条，其中政治类占比最大，约 28%；新闻类占 22%，排第二；娱乐类和商业类分别占 13% 和 11%，排第三和第四；其他多达 18 个细分类别均不到 10%。这种不平衡性使谣言识别可能在某些领域准确性较高，而另外的领域则效果不佳，这也为分领域的谣言检测提出了挑战。

因此，本文提出一种跨领域的谣言检测模型，将利用具有丰富标注数据的领域谣言构建的分类模型迁

①https://twitter.com/search?q=#ThisHappened.

②https://www.snopes.com/archive/.

移到含有少量数据或缺少标注数据的领域中，解决领域间数据或标注不平衡的问题。

3.2　跨领域谣言检测模型

(1) 模型框架

本文的建模场景针对谣言在某些领域(目标领域T)存在数据量少且缺少标注的问题，借助标注数据丰富的领域(源领域 S)数据通过迁移学习，提高目标领域的检测效果。借鉴文献[4]，结合谣言的文本特征、用户特征、传播特征构建深度迁移谣言检测模型，如图 1 所示。因为用户特征与传播特征不受领域影响，因此先对文本特征进行跨领域迁移，在网络训练时再引入用户特征与传播特征。

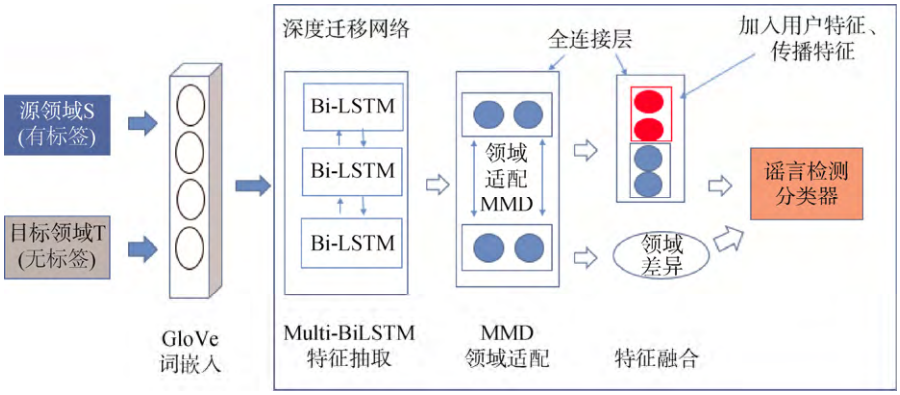

图 1　跨领域谣言检测流程

整个流程可分为词嵌入处理、特征抽取、领域适配、特征融合、预测分类等步骤。具体而言，模型通过预先训练 GloVe 词向量实现对输入文本的向量化；将两个领域的文本向量经过同一个多层双向 LSTM 网络[13]，完成对文本特征深度提取(特征抽取)；然后将经过特征抽取过的两个领域数据输入全连接层，这时会计算两个领域数据分布的差异；最后将输出向量与对应的用户特征向量及传播特征向量进行拼接，并送入 Softmax 层进行预测训练。

考虑极端情况下的领域迁移：即源领域存在标注数据集 SE，目标领域存在无标注数据集 TE，y_i 表示类别变量，定义 $SE=\{(SE_1,y_1)\ldots(SE_i,y_i)\ldots(SE_n,y_n)\}$，$TE=\{TE_1\ldots TE_i\ldots TE_m\}$。

(2) 特征抽取

特征抽取通过深度网络实现对谣言文本特征的表示，使用 Multi-BiLSTM 网络作为特征抽取的网络结构。这是因为，一方面文本数据具有天然的序列化特征，LSTM 能保留文本的位置信息；另一方面，一般社交媒体信息属于典型的短文本，LSTM 加入了门控机制，能有效解决短文本中上下文信息少、语义不明确等问题。

根据 Ma 等[8]的研究，RNN 网络能够在网络谣言检测任务中有效对文本特征进行抽取，在此基础上通过对比实验证明了使用复杂 RNN 衍生结构与增加网络层数可显著提升特征抽取效果，继而提升谣言检测效果。借鉴 Ma 等[8]的研究成果，本文选择双层双向 LSTM 网络(Multi-BiLSTM)作为特征抽取层，其中双向机制保证了每一个词都在充分考虑上下文的条件下获得语义[14]，这样双层网络保证了文本特征被深度抽取，其结构如图 2 所示。

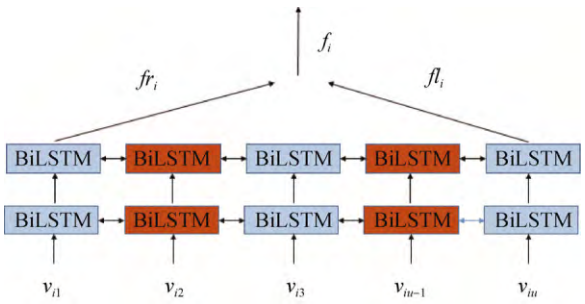

图 2　Multi-BiLSTM 网络

定义源领域 S 与目标领域 T 经过特征抽取输出的结果分别为 $Sf=\{Sf_1\ldots Sf_i\ldots Sf_m\}$ 与 $Tf=\{Tf_1\ldots Tf_i\ldots Tf_m\}$。

(3) 领域适配

领域适配旨在减小领域间的数据分布差异[15]，使两个领域数据分布趋同，如图3所示，蓝色表示源领域 S，红色表示目标领域 T，圆点表示谣言，三角表示非谣言数据，在跨领域构建谣言检测模型时，源领域 S

与目标领域 T 具有不同的数据分布(左图)，因而不能直接使用源领域 S 帮助目标领域 T 构建谣言检测。利用基于特征的迁移学习技术进行领域适配，即通过特征变换使得源领域与目标领域数据分布趋于相似(右图)，继而目标领域可以借助源领域的标签信息进行训练建模，这一过程也称为文本的特征对齐(Feature Alignment)。

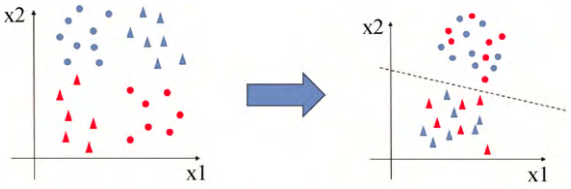

图 3　领域适配示意

最大均值差异(Maximum Mean Discrepancy, MMD)方法是由 Borgwardt 等[16]提出的判断两类样本是否同属于一个总体分布的指标，最初用于双样本检测中。因为 MMD 统计量的大小表示两个数据分布之间的距离，很多学者利用这一点将 MMD 应用在迁移学习领域[17-19]。本文基于此将 MMD 方法应用到网络谣言检测的领域适配中，在 Multi-BiLSTM 层后添加领域适配层，计算源领域 S 与目标领域 T 间的距离，并将其计入整个深度迁移网络的损失函数。借助于特征抽取层 Multi-BiLSTM 网络良好的特征变换能力[20]，在训练过程中可以不断减小源领域 S 与目标领域 T 的分布差异，从而实现领域的适配。

设领域适配层权值为 w_a，偏置量为 b_a，$\sigma(\cdot)$ 为 sigmod 函数，则领域适配层的输出分别如公式(1)和公式(2)所示。

$$Sa_i = \sigma(w_a \cdot Sf_i + b_a) \tag{1}$$

$$Ta_i = \sigma(w_a \cdot Tf_i + b_a) \tag{2}$$

同时计算基于 MMD 的领域分布距离并计入最终的误差函数。

设源领域数据 $Sa = \{Sa_1 \ldots Sa_i \ldots Sa_n\}$ 服从 p 分布，目标领域数据 $Ta = \{Ta_1 \ldots Ta_i \ldots Ta_n\}$ 服从 q 分布。在一个再生核希尔伯特空间(Reproducing Kernel Hilbert Space, RKHS) H 中存在映射函数 $\phi(\cdot): X \to H$ 表示从原始空间到希尔伯特空间的一个映射[16]，则源领域与目标领域的 MMD 距离如公式(3)所示。

$$MMD(Sa,Ta) = \left\| E_p[\phi(Sa)] - E_q[\phi(Ta)] \right\|_H \tag{3}$$

其中，H 是完备的内积空间，进一步描述为 $L2$ 范式和内积形式如公式(4)、公式(5)所示，其中 μ_p, μ_q 表示 p,q 分布的总体分布[16]。

$$MMD^2(Sa,Ta) = \| < \mu_p - \mu_q, \phi(\cdot) >_H \|^2 \tag{4}$$

$$MMD^2(Sa,Ta) = E[\phi(Sa)\phi(S\hat{a})] + E[\phi(Ta)\phi(T\hat{a})] - 2E[\phi(Ta)\phi(Sa)] \tag{5}$$

存在再生核 $k(x,y) = \phi(x)\phi(y)$，所以存在MMD的估计量如公式(6)所示[16]。

$$MMD^2(Sa,Ta) = [\frac{1}{n^2}\sum_{i,j=1}^{n} k(Sa_i, Sa_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(Ta_i, Ta_j) - \frac{2}{mn}\sum_{i,j=1}^{n,m} k(Sa_i, Ta_j)] \tag{6}$$

仿照 SVM 的核方法，在实际计算中核函数 $k(x,y)$ 可选择高斯核、线性核等。本文选用高斯核，如公式(7)所示。

$$k(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2}) \tag{7}$$

至此基于 MMD 计算的领域距离已经计算完成。如果将其作为误差项直接计入损失函数，则使用 SGD 求优时的时间复杂度为 $o(n^2)$，这样会大大降低迁移效率。采用 Gretton 等[21]提出的线性无偏估计量作为最终领域距离，可以使计算的时间复杂度由 $o(n^2)$ 降为 $o(n)$。

定义函数 $q(d_i)$ 如公式(8)所示，进而可以计算 MMD 的值如公式(9)所示[21]。

$$q(d_i) \triangleq k(Sa_{2i-1}, Sa_{2i}) + k(Ta_{2i-1}, Ta_{2i}) - k(Sa_{2i-1}, Ta_{2i}) - k(Ta_{2i-1}, Sa_{2i}) \tag{8}$$

$$MMD^2(Sa,Ta) \triangleq \frac{2}{n}\sum_{i=1}^{n/2} q(d_i) \tag{9}$$

至此，公式(9)将作为误差函数将计入总体迁移网络的总体误差。

(4) 特征融合

特征融合是将不同角度的特征结合，为分类提供更多可依据的信息。网络谣言的检测除了使用文本特征，也可以利用其他特征。如从用户角度出发，造谣者作为非正常用户常常具有粉丝少、关注多、注册地不明确、缺少认证等特点。从传播角度出发，谣言在传

播中蛊惑性较强，具有较多的转发数、较少的评论数等。这两类特征并不随文本领域的不同而产生差别，因此本文设计特征融合层，在结合 Castillo 等[4]、祖坤琳等[7]、Zhang 等[10]研究成果的基础上，选择 8 项用户特征、5 项传播特征与输出的文本特征向量拼接，形成完整的特征向量，最终送入 Softmax 层生成预测类别。用户特征与传播特征如表 1 和表 2 所示。

表 1　用户特征

特征名	特征解释
reg_period	注册时间，从注册日到数据获取日，以月计
followers_count	作者粉丝数
friends_count	作者朋友数
listed_count	作者关注数
statuses_count	作者发布信息的数量
Description	作者个人描述的长度
Sex	作者性别
Location	用户位置

表 2　传播特征

特征名	特征解释
retweet_count	转发数
retweet favorite_count	转发点赞数
retweet_comment_count	转发评论数
retweet_ followers_count	转发用户平均粉丝数
retweet_reg_period	转发用户平均注册时间

设特征融合层权值为 w_m，偏置量为 b_m，Su_i 为第 i 个用户的用户特征，Sp_i 为第 i 个用户的传播特征。源领域 S 数据特征融合层输出为 Sm_i，类别预测变量为 $\hat{y}_i$，则有网络结构如公式(10)和公式(11)所示。

$$Sm_i = \sigma(w_m \cdot [Sa_i, Su_i, Sp_i] + b_m) \tag{10}$$

$$\hat{y}_i = Softmax(Sm_i) \tag{11}$$

(5) 目标函数

本文提出的深度迁移网络类似于多任务学习的形式，一方面通过学习源领域与目标领域的数据分布差异实现领域适配，另一方面学习了源领域的标签信息，反映在目标函数上就是整个网络的损失函数由两部分组成，分别是代表领域差距的 MMD 统计量与反映源领域标签信息的 $\sum_{i=1}^{n} Loss(y_i, \hat{y}_i)$，因此整个迁移网络的损失函数可表示如公式(12)所示，其中 K 为迁移常数。

$$Global\_Loss(S,T) = \sum_{i=1}^{n} Loss(y_i, \hat{y}_i) + K \cdot MMD^2(Sa, Ta) \tag{12}$$

4　实验与分析

4.1　数据采集与预处理

(1) 数据获取

本实验在 Twitter 数据集上进行。数据一部分来自文献[8]公开的 Twitter 数据集，其中包含 498 条谣言事件(对内容相似 Twitter 的总结文本)与 494 条非谣言事件，以及这些事件后对应的 Twitter 文本。从中选择 5 个领域(Politics、News、Business、Food、History)的谣言文本，从每个领域人工抽取 10 个关键词组成查询项，将这些查询项通过 Twitter 官方的 API 进行检索，得到包含 10 314 条 Twitter 的数据集，其明细如表 3 所示。

表 3　Twitter 数据集明细

文本领域	谣言文本数(条)	非谣言文本数(条)
Politics	1 780	1 934
News	1 744	1 659
Food	562	676
History	488	440
Business	576	455

原始的 Tweet 数据包含较多信息，通过文本提取获取到 Tweet 的正文部分，作为实验的原始输入。同时提取其中的用户信息与传播信息，经过归一化处理后在模型后半部分进行特征融合。实验采用交叉验证的方式，实验结果图表展示了相关指标的均值。

(2) 文本预处理

使用 GloVe(Global Vectors for Word Representation)实现对 Tweet 文本的词嵌入建模。GloVe 是 Pennington 等[22]提出的一种无监督词嵌入技术。相较于 Word2Vec 模型，GloVe 通过加入全局的共现矩阵(Co-occurrence Matrix)信息，有效解决了过于依赖局部信息、多义词处理乏力等问题，适合 Twitter 这类短文本[23]。

选择 Stanford 自然语言处理小组公开的 Twitter 文本集①作为词库(本文选择其中的 2 000 万条构建词库)，

①https://nlp.stanford.edu/projects/glove/.

在此基础上训练 GloVe 词向量。常见的 Twitter 文本一般由文字、符号、表情、@、hashtag(标签)、链接等元素构成，传统的文本预处理方式会直接删除其他元素只保留文字元素。但 Castillo 等[4]的研究表明符号、表情、@、hashtag(标签)、链接等元素可作为有效识别谣言的要素，并在模型中加入这些元素的统计量实现了较好的预测效果。本文在此基础上，使用特定标签对 Twitter 文本中的特殊元素进行替换，并训练得到对应标签的词向量，有效保留了这些元素的位置信息与数量信息。具体处理如图 4 所示。

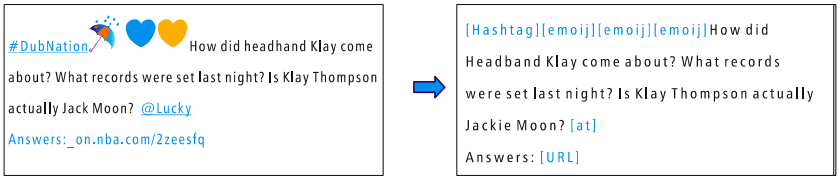

图 4　保留特殊元素的文本预处理

4.2　迁移常数的选择

选择从 Politics 到 Business 与 News 到 Food 的迁移场景，不同迁移场景最佳迁移常数不同，但是迁移常数一般在 1 附近取得最佳迁移效果，不同迁移常数下 F1 值的变化如图 5 所示。说明在误差函数中，领域损失项与标签损失项同样重要，这从另一个角度也印证了领域适配的重要性。P、B、N、F 分别表示 Politics, Business, News, Food 等领域。

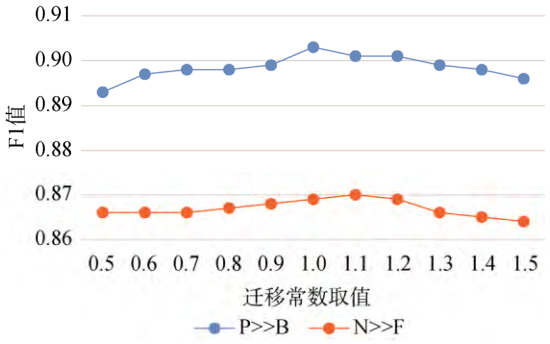

图 5　不同迁移常数下 F1 值的变化

4.3　迁移有效性度量

(1) 迁移有效性

迁移有效性度量迁移前后目标领域谣言检测模型的准确率变化。实验选定 Politics 与 News 领域为源领域，Food、History 与 Business 为目标领域，设计 6 组迁移实验(目标领域的标签信息仅用作测试集，在训练时不带入模型)。选择的模型分别是:

①深度迁移网络(1D1A)。即本文的方法，D 代表领域，A 代表迁移，1D1A 代表既划分了谣言的领域又实现了迁移学习。

②不划分领域(0D0A)。将所有领域当作一个整体，不考虑领域差异的影响，因而也不存在领域适配和迁移。

③分领域但未做领域适配(1D0A)。即分领域构建谣言检测模型，但对于无标注的目标领域 T，直接应用源领域数据得到的检测模型。源领域训练时网络结构不变，数据输入只有源领域数据，误差函数也删除了 MMD 损失项。

实验结果如表 4 所示。

表 4　不同迁移任务实验结果

迁移领域	建模方式	Precision	Recall	F1
P>>F	1D1A	**0.814**	**0.822**	**0.817**
	0D0A	0.734	0.725	0.729
	1D0A	0.760	0.715	0.722
P>>H	1D1A	**0.874**	**0.869**	**0.872**
	0D0A	0.762	0.743	0.755
	1D0A	0.794	0.786	0.792
P>>B	1D1A	**0.901**	**0.895**	**0.888**
	0D0A	0.792	0.789	0.795
	1D0A	0.811	0.802	0.807
N>>F	1D1A	**0.869**	**0.872**	**0.871**
	0D0A	0.762	0.759	0.760
	1D0A	0.784	0.780	0.782
N>>H	1D1A	**0.865**	**0.872**	**0.863**
	0D0A	0.754	0.762	0.758
	1D0A	0.789	0.786	0.787
N>>B	1D1A	**0.904**	**0.884**	**0.891**
	0D0A	0.788	0.781	0.786
	1D0A	0.803	0.805	0.802

①深度迁移网络有效性。划分领域又做迁移学习(1D1A)的结果相较于另外两种不做迁移的建模思路准确率平均提升 10.6%(0D0A)与 8.1%(1D0A)，召回率平均提升 10.9%(0D0A)与 9.0%(1D0A)，F1 值平均提升 10.3%(0D0A)与 8.5%(1D0A)。这说明深度迁移网络通过加入领域适配层有效利用源领域数据的标签信息，减小了领域分布的差异，有效提升检测效果。

②领域间数据分布差异明显，分领域进行谣言检测更

加精细化。1D0A 忽略了领域间的数据分布差异，直接将源领域的分类器应用于目标领域。其各项指标平均比深度迁移网络低 8.5%左右，这说明领域间确实存在数据分布差异，分领域的建模方式具有更高准确性。

③数据在领域上的不平衡会降低数据贫瘠领域的准确度。0D0A 遵循传统的建模思想将所有领域看作一个整体进行建模。相较于分领域的深度迁移网络，其在谣言检测的准确率上平均低 10.6%。这种准确度降低主要来自于数据集的不平衡，数据越丰富效果越显著。

④分布越相近领域，迁移效果越好。从传播学角度，Business 作为目标领域与源领域 Politics 和 News 有更相近的数据分布，因而经过迁移后，谣言检测的准确率平均提高约 10.0%。而从 Politics 到 Food 的迁移任务中，谣言检测准确率仅提升 6.1%左右，这也符合人的直观认识。

(2) 比对实验

将迁移学习的结果与已有方法从有标签数据和无标签数据两个方面展开对比。有标注为监督学习方法，使用目标领域的标签信息构建谣言检测模型，分类器为支持向量机 SVM 和线性回归 LR。另一方面，更常见的情况是目标领域往往缺少标注数据，这就需要无监督的方式进行谣言检测。采用文献[10-11]提出的无监督模型作为对比实验，其中文献[10]的方法定义为 Multi-AN，文献[11]的方法定义为 RNN+AN。实验结果如表 5 所示。

表 5　与其他学习方法的比对结果

目标领域	建模方式	Precision	Recall	F1
Food	深度迁移学习(P>>F)	0.814	0.822	0.818
	深度迁移学习(N>>F)	0.869	0.872	0.870
	SVM(有监督学习)	**0.881**	**0.875**	**0.878**
	LR(有监督学习)	0.873	0.869	0.871
	Multi-AN(无监督学习)	0.802	0.801	0.801
	RNN+AN(无监督学习)	0.823	0.813	0.818
History	深度迁移学习(P>>H)	0.874	0.869	0.871
	深度迁移学习(N>>H)	0.865	0.872	0.868
	SVM(有监督学习)	**0.903**	**0.890**	**0.896**
	LR(有监督学习)	0.873	0.865	0.869
	Multi-AN(无监督学习)	0.869	0.871	0.870
	RNN+AN(无监督学习)	0.882	0.862	0.872
Business	深度迁移网络(P>>B)	0.901	0.895	0.898
	深度迁移网络(N>>B)	0.904	0.884	0.894
	SVM(有监督学习)	**0.913**	**0.906**	**0.909**
	LR(有监督学习)	0.907	0.901	0.904
	Multi-AN(无监督学习)	0.824	0.819	0.821
	RNN+AN(无监督学习)	0.831	0.826	0.828

在有监督学习方面，SVM 在三个领域的准确率平均比深度迁移网络高 2.7%左右，比 LR 则平均高 1.3%，这说明准确率差别并不显著，迁移学习已经非常接近有监督学习的检测效果。而且在不同的目标领域表现较为稳定，说明深度迁移网络的领域适配机制使得源领域标签信息可以被目标领域使用。由于实际应用中目标领域往往不具备充足的标注数据，人工标注数据意味着巨大的时间成本与人工代价，深度迁移网络就可以在无标注的数据集上得到谣言检测模型，且准确率与在有标注数据集学习到的模型十分接近。

在无监督学习方面，从表 5 的实验结果来看，文献[11]提出的 RNN+AN 模型是无监督方法中效果最好的，以 RNN+AN 的模型为对比分别分析本文方法在三个领域内的迁移效果。在 Food 领域，以 Politics 为源领域的迁移方法 F1 值与 RNN+AN 方法的 F1 值持平，但以 News 为源领域的迁移方法在 F1 指标上却比对比方法高出 5.2%，这说明在 Food 领域内本文方法总体优于无监督方法；在 History 领域，两组迁移任务在 F1 值上的表现略低于 RNN+AN 的无监督方法(分别低 0.1%与 0.4%)，但是在 Recall 指标上均高于无监督方法，这说明更多的潜在谣言被找到，这一点在谣言的检测中显得更加重要；在 Business 领域，本文方法

在 F1 值上比 RNN+AN 方法分别高出 7.0%与 6.6%。综合来看，本文方法优于无监督的谣言检测方法。无监督方法是解决谣言数据无标注问题的一种思路，其放弃了对标签信息的使用，只选择微博数据的自身特征进行谣言检测，因而检测效果有待进一步提升。本文方法在面对大量无标注数据时，借助深度迁移网络，有效利用源领域的标签信息，因而在实验结果上优于无监督方法。

5 结 语

本文针对不同领域网络谣言的识别问题，尝试对网络谣言实现分领域的有效检测。提出一种基于深度迁移网络的跨领域谣言检测模型，在源领域拥有标注数据、目标领域是无标注数据的情况下，通过双层双向 LSTM 网络实现对文本特征的深度提取，在领域适配层计算源领域与目标领域 MMD 作为领域误差并送入全局误差，与用户特征及传播特征相融合，使用 Softmax 进行分类。实验在 Twitter 数据集上展开，结果证明深度迁移网络能够有效迁移源领域标注数据，帮助目标领域构建谣言检测模型，显著提升目标领域谣言的检测准确率。

本文模型在面对不同领域的迁移任务时表现存在一定差别，在进行迁移任务之前不能明确选择最优的源领域，原因在于并未深入探讨领域分布差异对领域迁移任务的影响，这将是下一步的工作重点，以提升跨领域迁移的可扩展性与可控性。另一方面，本文模型假定微博属于某一单一领域，实际生活中很多网络谣言难以划分到某一个领域，常常会同时属于社会、生活、经济等多个领域，这一点广泛地存在于中文微博谣言。因此本文模型面对中文微博谣言时尚有一定的局限性，也没有在中文微博上进行相关实验，未来需要利用多标签学习技术进行深入分析，探索适应中文数据的跨领域迁移方法。另外，进一步的研究还包括如何提升深度迁移网络训练效率、改进本方法在监督学习上的表现等。

参考文献:

[1] 曹博林. 社交媒体: 概念、发展历程、特征与未来——兼谈当下对社交媒体认识的模糊之处[J]. 湖南广播电视大学学报, 2011(3): 65-69. (Cao Bolin. Social Media: Definition, History of Development, Features and Future—The Ambiguous Cognition of Social Media[J]. Journal of Hunan Radio & Television University, 2011(3): 65-69.)

[2] 雷霞. 谣言: 概念演变与发展[J]. 新闻与传播研究, 2016(9): 113-118. (Lei Xia. Rumor: Concept Evolution and Development[J]. Journalism & Communication, 2016(9): 113-118.)

[3] Fanti G, Kairouz P, Oh S, et al. Hiding the Rumor Source[J]. IEEE Transactions on Information Theory, 2017, 63(10): 6679-6713.

[4] Castillo C, Mendoza M, Poblete B. Information Credibility on Twitter[C]// Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India. 2011: 675-684.

[5] Ma J, Gao W, Wei Z, et al. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites[C]// Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Australia. ACM, 2015: 1751-1754.

[6] Zhao Z, Resnick P, Mei Q. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts[C]// Proceedings of the 24th International Conference on World Wide Web, Florence, Italy. ACM, 2015: 1395-1405.

[7] 祖坤琳, 赵铭伟, 郭凯, 等. 新浪微博谣言检测研究[J]. 中文信息学报, 2017, 31(3): 198-204. (Zu Kunlin, Zhao Mingwei, Guo Kai, et al. Research on the Detection of Rumor on Sina Weibo[J]. Journal of Chinese Information Processing, 2017, 31(3): 198-204.)

[8] Ma J, Gao W, Mitra P, et al. Detecting Rumors from Microblogs with Recurrent Neural Networks[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, USA. 2016: 3818-3824.

[9] Chen T, Li X, Yin H, et al. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection[C]// Proceedings of the 2018 Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2018: 40-52.

[10] Zhang Y, Chen W, Yeo C K, et al. Detecting Rumors on Online Social Networks Using Multi-Layer Autoencoder[C]// Proceedings of the 2017 IEEE Technology & Engineering Management Conference. IEEE, 2017: 437-441.

[11] Chen W, Zhang Y, Yeo C K, et al. Unsupervised Rumor Detection Based on Users' Behaviors Using Neural Networks[J]. Pattern Recognition Letters, 2018, 105: 226-233.

[12] 刘雅辉, 靳小龙, 沈华伟, 等. 社交媒体中的谣言识别研究综述[J]. 计算机学报, 2018, 41(7): 1536-1558. (Liu Yahui, Jin Xiaolong, Shen Huawei, et al. A Survey on Rumor Identification over Social Media[J]. Chinese Journal of Computers, 2018, 41(7): 1536-1558.)

[13] Zhou J, Xu W. End-to-End Learning of Semantic Role Labeling Using Recurrent Neural Networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 1127-1137.

[14] Chen T, Xu R, He Y, et al. Improving Sentiment Analysis via Sentence Type Classification Using BiLSTM-CRF and CNN[J]. Expert Systems with Applications, 2017, 72: 221-230.
[15] Blitzer J, McDonald R, Pereira F. Domain Adaptation with Structural Correspondence Learning[C]// Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia. ACM, 2006: 120-128.
[16] Borgwardt K M, Gretton A, Rasch M J, et al. Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy[J]. Bioinformatics, 2006, 22(14): e49-e57.
[17] Ghifary M, Kleijn W B, Zhang M. Domain Adaptive Neural Networks for Object Recognition[C]// Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence. 2014: 898-904.
[18] Tzeng E, Hoffman J, Zhang N, et al. Deep Domain Confusion: Maximizing for Domain Invariance[OL]. arXiv Preprint, arXiv: 1412.3474.
[19] Long M, Cao Y, Wang J, et al. Learning Transferable Features with Deep Adaptation Networks[C]// Proceedings of the 32nd International Conference on Machine Learning, Lille, France. 2015: 97-105.
[20] Mou L, Meng Z, Yan R, et al. How Transferable are Neural Networks in NLP Applications?[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA. 2016: 479-489.
[21] Gretton A, Sriperumbudur B, Sejdinovic D, et al. Optimal Kernel Choice for Large-Scale Two-Sample Tests[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, USA. 2012: 1205-1213.
[22] Pennington J, Socher R, Manning C D. GloVe: Global Vectors for Word Representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar. 2014: 1532-1543.
[23] Wu K, Yang S, Zhu K Q. False Rumors Detection on Sina Weibo by Propagation Structures[C]// Proceedings of the 31st International Conference on Data Engineering, Seoul, South Korea. IEEE, 2015: 651-662.

作者贡献声明:

刘勘: 提出研究思路, 制定研究方案, 论文撰写、修改及最终版本修订; 杜好宸: 采集和分析数据, 进行实验, 撰写论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: dududuhc@163.com。
[1] 杜好宸. data.zip. Twitter 谣言的训练集和测试集数据.
[2] 杜好宸. Tweet_word_embedding_200d.csv. 使用 GloVe 算法预训练的 Tweet 词向量.
[3] 杜好宸. result.rar. 文中实验数据及图表数据.

收稿日期: 2018-11-09
收修改稿日期: 2019-04-18

Detecting Twitter Rumors with Deep Transfer Network

Liu Kan　Du Haochen
(School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073,China)

Abstract: [Objective] This paper proposes a new model to address the issue of insufficient data facing network rumors detection. **[Methods]** We proposed a deep transfer network based on the Multi-BiLSTM network as well as domain distributions of MMD statistics calculation. Then, we trained the model to learn the data loss of source domain and the distribution difference among domains. Finally, we realized the effective migration of label information across domains. **[Results]** Compared with two traditional rumor detection methods, the proposed model's F1 index was increased by 10.3% and 8.5% respectively. **[Limitations]** The effect of transfer was not obvious in skewed data distribution and multiple domains. **[Conclusions]** The proposed method could improve the rumor detection results. The deep transfer network could achieve positive outcomes among domains, and provide new directions for Internet rumor recognition.
Keywords: Rumor Detection　Deep Transfer Network　Multi-BiLSTM　Domain Adaption　Twitter