·人工智能及识别技术· 文章编号:1000-3428(2019)05-0116-06 文献标志码:A 中图分类号:TP391

基于语义结构的迁移学习文本特征对齐算法

卢晨阳,康 雁,杨成荣,蒲 斌

(云南大学 软件学院,昆明 650500)

摘 要: 特征对齐在源域和目标域空间不一致时会导致负迁移现象。为此,提出一种基于 GloVe 和 WordNet 模型的迁移学习文本特征对齐算法。根据数据样本词性和类别对分类任务进行特征筛选,选择源域和目标域的领域共有词作为枢纽词,使用 GloVe 模型对齐源域和目标域中最相似的非枢纽特征。在此基础上,根据源域和目标域的非共有特征,通过 WordNet 模型对领域独立特征完成强语义对齐,同时利用含有枢纽特征的对齐三元组表示对齐特征。实验结果表明,该算法可有效降低特征维度,扩充特征空间,提高跨领域文本分类精度。

关键词: 迁移学习;特征对齐;词向量;词网;文本挖掘

中文引用格式: 卢晨阳,康雁,杨成荣,等. 基于语义结构的迁移学习文本特征对齐算法[J]. 计算机工程,2019,45(5):116-121.

英文引用格式: LU Chenyang, KANG Yan, YANG Chengrong, et al. Text feature alignment algorithm for transfer learning based on semantic structure[J]. Computer Engineering,2019,45(5):116-121.

Text Feature Alignment Algorithm for Transfer Learning Based on Semantic Structure

LU Chenyang, KANG Yan, YANG Chengrong, PU Bin

(School of Software, Yunnan University, Kunming 650500, China)

【Abstract】 Feature alignment causes a negative transfer when the source domain space and target domain space are inconsistent. Therefore, a text feature alignment algorithm for transfer learning based on the GloVe and WordNet model is proposed. According to the part of speech and category of the sample data, feature filtering is performed to classification tasks. The shared terms of the source domains and target domain are selected as pivot words, and the GloVe model is used to align the most similar non-pivot features in the source domain and target domain. On this basis, according to the unique features of the source domain and target domain, strong semantic alignment is achieved through the WordNet model for the domain independent features. At the same time, alignment features are represented by aligning triples with pivot features. Experimental results show that the algorithm can effectively reduce the feature dimension, expand the feature space, and improve the accuracy of cross-domain text classification.

【Key words】 transfer learning; feature alignment; word vector; WordNet; text mining

DOI: 10.19678/j.issn.1000-3428.0050574

0 概述

近年来,迁移学习已成为人工智能领域的研究热点。多数信息以文本数据的形式存在于网络上,因此对文本挖掘方法进行研究具有重要意义。例如,网络舆情分析使用文本挖掘方法,通过分析网民对某一新闻热点的情感极性(积极或消极)来判断某新闻事件的社会影响。关于情感分类已经有较多的研究成果[1-3]。在监督式学习模式下,情感倾向性分析依赖大量有标记的情感数据,通过训练判别模型对未标记的数据进行情感分析。但在实际情况中,有标记的数据获取成本很高,且人工标记的数据可能存在主观性问题。迁移学习可以将相似领域的源域知识迁移到只有少量标注信息的目标域中,以帮助目标域训练。因此,如何寻找可以迁移的相似领域、迁移哪些特征以及如何迁移是迁移学习研究的重点。

由于语言的多样性和同义性,不同领域描述相同情感的情感词及其概率分布会有所不同。例如,在电影评论中,人们会用“magnificent”“interesting”等词表

基金项目: 国家自然科学基金(61762092);云南省软件工程重点实验室开放基金(2017SE204)。

作者简介: 卢晨阳(1994—),男,硕士研究生,主研方向为自然语言处理;康 雁,副教授;杨成荣、蒲 斌,硕士研究生。

收稿日期: 2018-03-02 **修回日期:** 2018-04-09 **E-mail:** luchenyanglcy@163.com

达积极的语意,用“suffer”“complex”等词表达消极的语义;在厨房商品评论中,人们会用“convenience”“clean”等词语表达积极的语义,使用“noise”“trash”等词表达消极的语义。迁移学习如果直接使用电影评论的语义判别模型去预测厨房商品评论的语义倾向,效果往往较差。因此,迁移学习要想获得比较好的迁移效果,需要使源域数据特征和目标域数据的数据特征保持一致。

本文提出一种基于 GloVe 模型的文本特征对齐算法。该算法使用 GloVe 从结构和弱语义层面对不同领域的特征进行对齐,利用 WordNet 从强语义层面对不同领域的相关特征进行特征对齐。

1　相关研究

目前,国内外研究者对迁移学习中特征空间不一致问题提出多种算法。文献[4]提出结构一致学习(Structure Correspondence Learning,SCL)算法,该算法利用不同领域中共同出现的枢纽特征对非枢纽特征进行关联,以达到结构对应学习的目的。文献[5]提出谱特征对齐(Spectral Feature Alignment,SFA)算法,该算法基于聚类假设,利用谱聚类算法对非枢纽特征进行对齐。文献[6]从情感、关键词和结构 3 个方面抽取情感句,并将其划分为关键字视图和细节视图,分别训练不同的基分类,并应用集成学习实现知识迁移。文献[7]使用特征变换来解决跨领域情感分析问题,通过选取枢纽特征,计算源域和目标域中同枢纽特征关联度最高的特征,对源域和目标域的文本进行特征变换,以完成特征对齐。文献[8]构建并使用语义敏感词典选取源域和目标域的相关特征词,通过扩展特征向量完成知识迁移。文献[9]提出特征集成及样本选择算法(SS-FE),该算法将 FE 和 PCA-SS 相结合完成迁移学习。文献[10]提出混合特征核即非负矩阵分解方法,该方法针对文本的异构输入特性使用联合非负矩阵因子分解,利用枢纽特征完成不同领域的特征迁移。文献[11]使用 Word2vec,根据词性不同分别训练 Word2vec,将领域共有词作为枢纽特征,使用 Word2vec 对不同领域的非枢纽特征进行映射。文献[12]提出共有子空间重构算法(Common Subspace Construction,CSC),使用源域和目标域的情感倾向识别公共子空间,然后将领域依赖特征投射到该子空间,基于公共特征子空间来表示评论文本,并在该子空间训练目标领域情感倾向预测模型。文献[13]提出主题对应迁移(Topical Correspondence Transfer,TCT)算法,通过识别不同领域间的公共主题来映射不同领域的其他主题,以减小不同领域特征的分布差异,模型在优化目标函数时发现主题并完成情感分类。此外,研究者提出使用深度学习来获取特征,并将其应用于自然语言处理,如命名实体识别、词性标注和语言模型中的知识迁移[14-16]。

2　特征对齐算法

传统的迁移学习文本特征对齐方法考虑词语在句子结构上的特征,例如,结构一致学习算法和谱特征对齐算法利用整个语料的词共现矩阵进行矩阵运算得到在频谱或概率上相似的词语,但忽略文本类标信息与文本中特征间的联系。在情感标记为积极的样本中,“good”“bravo”等表示积极语义的词语较多;在情感标记为消极的样本中,“bad”“terrible”等表示消极语义的词语较多。如果在特征对齐时不考虑类标信息,只考虑情感词与句子结构的相关关系,则可能导致表示积极语义的词被对齐为表示消极语义的词,使特征对齐后和语义簇与句子类标特征呈现非强相关性。

本文根据类标信息和词性分别抽取特征进行对齐,在特征抽取时保留了文本特征和语义簇的强相关性。然后对领域独立词使用 WordNet 同义词/近义词进行特征对齐,构建强语义对齐二元特征组。同时使用 Glove 对领域非枢纽特征进行对齐,构建结构弱语义三元特征组。

2.1　相关定义

领域是实体世界中类别事物的抽象表示。在文本分类任务中,领域用来表达不同类别事物的文本集合。例如,在亚马逊网站不同类型的产品及产品评论中,“books”“DVD”“furniture”是 3 个不同的领域。

在跨领域情感分类问题中,源域 D_s 是某领域有标签的评论文本,可表示为:

$$D_s = \{(x_i, y_i)\}_{i=1}^{n_s} \tag{1}$$

其中,y_i 是样本评论 x_i 的情感标签,且 $y_i \in \{+1, -1\}$,$+1$ 表示积极语义,-1 表示消极语义,n_s 是源域中有标记样本的数量。

目标域 D_t 是某领域无标签的评论文本,可表示为:

$$D_t = \{(x_i)\}_{i=1}^{n_t} \tag{2}$$

其中,n_t 是目标域中无标记样本的数量。

领域独有词是指去除停用词等噪声词后只在该领域出现而不在其他领域出现的词。如“books”评论中的“obscure”,“computer”评论中的“short battery life”,其表示特定领域的词。领域共有词是指去除停用词等噪声词后在不同领域间共同出现的词。如“like”“hate”等在各个领域中都会出现的词。

枢纽特征是指在源域与目标域的迁移过程中可以起到桥接作用的特征。本文使用出现频率达到一

定数量的领域共有词作为枢纽特征。除了枢纽特征外的其他特征称为非枢纽特征。某个跟枢纽特征最为靠近或相似的非枢纽特征,称为近邻枢纽特征。

跨领域文本特征对齐通过对领域的非枢纽特征进行抽取,然后对不同领域的非枢纽特征进行特征映射来完成领域特征对齐。在对齐源域和目标文本特征后,跨领域情感分类使用源域样本构造情感分类器来对目标域样本进行情感分类。

2.2 文本预处理

文本预处理主要包括以下内容:

1)词干提取。词干提取是将英文单词的各种时态变化和后缀提取为单词最原始的样子,且具有降维作用。在词干提取之前,将文本中的单词全部转换为小写。

2)停用词过滤。停用词是没有实际意义的词以及标点符号、数字、特殊字符等。如“the”“a”“an”“that”“those”“under”等。

3)词性标注及词性过滤。词性标注是对过滤过停用词的文本进行分词,然后对每一个词的词性进行标注。名词、动词、形容词、副词在情感分类中具有重要影响,对正类样本和负类样本分别抽取名词、动词、形容词和副词。有些词在词性标注中会被标注为多种词性,例如词语“like”同时具有动词、形容词、副词和名词的词性,但在不同的句子结构中词性不同,如在句子“I like this book.”中“like”是动词,而在句子“It sounds like the violin.”中“like”是形容词。为了让词性标注更为准确,本文使用文献[17]提出的基于概率的决策树词性标注方法,该方法利用决策树方法对词性的转移概率进行估计。

2.3 文本特征抽取

本文对要迁移的源域和目标域抽取领域独有词,然后将领域独有词作为 WordNet 模型的输入。

传统的枢纽特征抽取和对齐一般只考虑到句子的结构和词性等信息,而忽略词语特征和类别信息存在一定的相关性,因此在抽取特征时应考虑类别影响。以情感分类为例,在抽取特征训练模型时,首先将源域和目标域的正类样本(积极)和负类样本(消极)分开,将不同词性的词语抽取出来,其中,在情感分类中起主要影响的是名词、动词、形容词、副词,然后对正类样本和负类样本分别抽取名词、动词、形容词和副词,原始的源域和目标域的 2 份数据变成根据样本类别和词性划分的 16 份数据。最后使用 16 份数据作为 GloVe 模型的输入来进行训练。

2.4 GloVe 模型特征

传统的词向量模型大致分为 2 类,第 1 类是以隐语义分析(Latent Semantic Analysis,LSA)为代表使用矩阵分解的方法获取词向量[18],第 2 类是以 Skip-gram 为代表使用局部上下文窗口的方法获取词向量[19]。在模型中,矩阵分解类方法使用词共现矩阵,虽然可以较好地利用全局统计信息,但无法获得理想的向量空间结构。同时局部上下文窗口类方法虽然能在词类比任务上取得较好的效果,但忽略了全文词语统计信息。GloVe 模型融合这 2 类模型的优点,利用全局统计信息和局部窗口的统计信息生成词向量模型[20]。

$$J=\sum_{i,j=1}^{V} f(X_{ij})(\boldsymbol{w}_i^{\mathrm{T}}\tilde{\boldsymbol{w}}_j+b_i+\tilde{b}_j-\lg X_{ij})^2 \tag{3}$$

$$f(x)=\begin{cases}(x/x_{\max})\alpha, & x<x_{\max}\\ 1, & \text{其他}\end{cases} \tag{4}$$

其中,V 是词汇表大小,$\boldsymbol{w}_i$ 和 $\boldsymbol{w}_j$ 是词典中词 i 和词 j 的词向量,$\boldsymbol{X}$ 是文档的词共现矩阵,X_{ij} 表示词 i 和词 j 同时出现的次数,b_i 是词向量 $\boldsymbol{w}_i$ 的偏置,$\tilde{b}_j$ 是词向量 $\tilde{\boldsymbol{w}}_j$ 的偏置,α 是权重函数 $f(x)$ 的参数,一般取 3/4。为防止过拟合,降低噪声对词向量的影响,词向量 $\boldsymbol{w}$ 和 $\tilde{\boldsymbol{w}}$ 在随机初始化时取不同的值。

本文使用 16 份根据类标和词性切分好的数据输入 GloVe 训练模型,然后利用枢纽特征找出对应领域的非枢纽特征构成特征对齐三元组。传统的迁移学习文本特征对齐算法使用 A_B 的方式对不同领域的词语进行特征对齐表示,但该表示会丢失部分特征,例如源域的特征词在不同的上下文环境中可能对齐目标域的不同词语,而使用 A_B 这种形式则无法保留不同上下文环境中需要对齐的不同特征。例如“I hate this book but my wife cherish it.”“I hate this DVD but my wife love it so much.”2 个分别来自领域 A 和领域 B 的样本,在上下文环境为“but”时,领域 A 的词“cherish”对应为领域 B的词“love”,而在“I will always cherish this book.”“My lover always remind me to value this DVD,because we had a lot of beautiful memory about it.”2 个分别来自领域 A 和领域 B 的样本,在上下文环境为“always”时,领域 A 的词“cherish”对应为领域 B 的“value”。基于上述分析,A_B 表示形式只能保留一种特征对齐,无法保留在不同上下文词出现时的不同特征对齐。

本文提出一种加入上下文环境词的迁移特征对齐三元组的表示方式。在 GloVe 训练出的模型中,如果源域和目标域的枢纽特征词为 $W_c^{s,t}$、源域对齐的非枢纽特征为 W_i^s、目标域对齐的非枢纽特征为 W_j^t,则特征对齐后的词语表示为 $W_i^s\_W_c^{s,t}\_W_j^t$。该表示方式可以表示在不同上下文环境中源域和目标域对齐的不同特征,且能够扩充特征维数。GloVe 特征对齐示意图如图 1 所示。

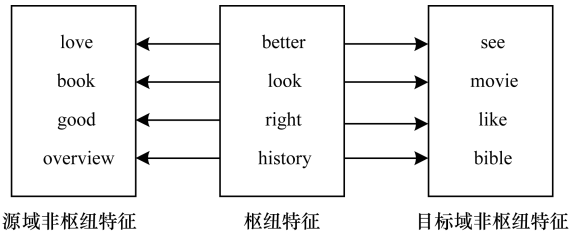

图 1　GloVe 特征对齐示意图

2.5　WordNet 模型特征

GloVe 对齐特征是在全局文档词频中统计窗口内相似的特征,在结构和部分弱语义上相似,如将“life”和“battery”“cheap”和“creak”在上下文相似和全局概率条件相似的词找出,但仍无法做到语义对齐。

WordNet 是一种基于认知语言学的英语词典,它按照单词意义将英文的各种词性组织为同义词集合,组成一个“单词的网络”,并在这些词汇概念间建立同义、反义、继承等多种词汇语义关系[21]。本文使用 WordNet 找出源域的领域独有词在目标域独有词是否存在近义词/同义词,如果源域的独有词 W_i^s 在与目标域的独有词 W_j^t 构成近义词/同义词关系,则在源域 D_s 和目标域 D_t 将 W_i^s 和 W_j^t 进行对齐并表示为 $W_i^s\_W_j^t$。由于 W_i^s 和 W_j^t 属于同一近义词簇,因此可以保证迁移学习的特征对齐的强语义相关性。WordNet 特征对齐示意图如图 2 所示。

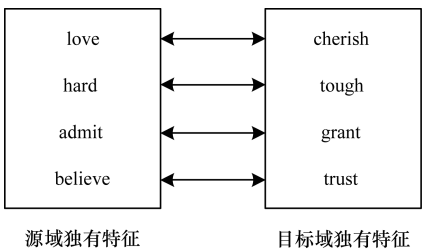

图 2　WordNet 特征对齐示意图

2.6　算法流程

文本预处理的主要步骤包括过滤词干提取、停用词和词性标注,然后将数据复制为 2 份,第 1 份根据类标和词性对文本特征进行抽取,第 2 份抽取领域独有词、领域共有词。将第 1 份按照类别和词性抽取过特征的共 16 份数据分别输入 GloVe 进行训练,此处将会产生 16 个训练出的模型,然后根据领域枢纽词找出对应模型的最相似非枢纽特征。

在迁移学习中,对 2 个领域结构不一致的特征进行迁移对齐,另一方面是对共有知识进行迁移。例如,在跨领域情感分类实验中,领域共有词不仅包含领域间的共有知识,其数据分布的不同体现不同领域内的领域知识,并且在实验中对于分类准确性影响较大。因此在特征对齐时,需并入领域共有词来构造对齐特征集。

在特征选择时,如果是领域共有词,则依次保留 WordNet 特征对齐二元组和 GloVe 特征对齐三元组。领域共有词表示领域间的公共知识域,而 WordNet 对齐特征表示在公共知识域中公认的语义相似或相同的特征。特征对齐算法如图 3 所示,具体步骤描述如下:

步骤 1　数据预处理(词干提取→过滤停用词→词性标注)。

步骤 2　数据复制 2 份,第 1 份切分为领域独立词、领域共有词,第 2 份按照类别和词性分开。

步骤 3　将分好的类别和词性数据分别输入 GloVe 模型训练,将领域共有词作为枢纽特征,在源域和目标域对应类别和词性训练好的 GloVe 模型中寻找与枢纽特征最相似的非枢纽特征,构造特征对齐三元组。

步骤 4　使用源域的领域特有词在目标域的领域特有词中寻找是否存在同义词/近义词,如果存在,则构造特征对齐二元组。

步骤 5　如果是领域共有词,则保留领域共有词,其次保留 WordNet 特征对齐二元组,再次保留 GloVe 特征对齐三元组。

步骤 6　使用特征对齐二元组和特征对齐三元组对源域和目标域对齐的词进行替换。

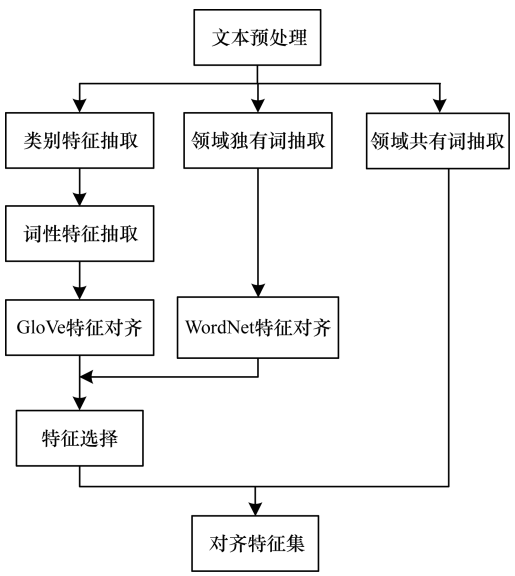

图 3　特征对齐算法流程

3　实验结果与分析

3.1　数据集

实验数据集使用亚马逊产品评论数据集,其包

含 4 个产品领域的商品评论,分别是:books(B),DVDs(D),electronics(E)和 kitchen appliances(K)。每个领域各包含 1 000 条积极评论样本和 1 000 条消极评论样本,每条样本基于评论的给分已经被情感标记为 +1(积极情感)或 -1(消极情感)。本文进行 12 组跨领域情感分类任务实验,分别是 D→B,E→B,K→B,B→D,E→D,K→D,B→E,D→E,K→E,B→K,D→K,E→K,其中,箭头左边表示迁移学习的源域,箭头右边表示迁移学习的目标域。

3.2 结果对比

为验证算法的有效性,本文在亚马逊产品评论数据集上与 SCL-MI、SFA、SS-FE、Word2vec 等文本特征对齐算法进行对比实验,结果如表 1 所示。实验使用逻辑回归(Logistic Regression,LR)作为基分类器。其中,Word2vec 算法和本文 GloVe 算法的窗口大小设置为 5,迭代次数设置为 4 000。

表 1 不同算法迁移分类准确率对比结果

领域	SCL-MI 算法	SFA 算法	SS-FE 算法	Word2vec 算法	本文算法
D→B	0.800	0.780	0.800	0.820	0.840
E→B	0.750	0.760	0.730	0.800	0.810
K→B	0.690	0.750	0.730	0.750	0.790
B→D	0.760	0.810	0.790	0.830	0.840
E→D	0.710	0.770	0.750	0.800	0.810
K→D	0.770	0.770	0.760	0.810	0.830
B→E	0.760	0.730	0.740	0.790	0.800
D→E	0.740	0.770	0.770	0.810	0.840
K→E	0.870	0.850	0.830	0.900	0.910
B→K	0.790	0.790	0.780	0.810	0.840
D→K	0.810	0.810	0.780	0.840	0.860
E→K	0.860	0.870	0.850	0.880	0.910
平均值	0.775	0.787	0.776	0.821	0.841

从表 1 可以看出,本文算法的平均准确率高于其他算法。其原因是本文算法根据样本类别和词性不同分开训练模型,利用 GloVe 模型从结构和部分语义上进行特征对齐,同时根据上下文特征,设计含有上下文的特征三元组对齐领域特征。针对强语义层面情感词,使用 WordNet 模型进行对齐领域特征。因此,本文通过 GloVe 模型和 WordNet 模型从结构和语义上对不同领域相似的特征进行空间映射,同时保留领域共有词,不仅可以保证迁移学习的共有特征空间一致,而且扩展了特征映射空间的知识域。

3.3 参数敏感性

在 GloVe 模型训练时,训练参数主要有上下文窗口大小、迭代次数、向量维度大小。为确定适合该数据集的模型参数,本文在保证模型收敛的情况下测试上下文窗口大小,同时使用平均分类准确率作为衡量指标,其中,迭代次数设为 400 次,向量维度设为 300。在目标域为 B 时,上下文窗口的分类准确率结果如图 4 所示。如果上下文窗口太小,则容易导致模型无法较好地学习全文统计信息;如果上下文窗口太大,则容易导致模型无法较好地学习语义信息。

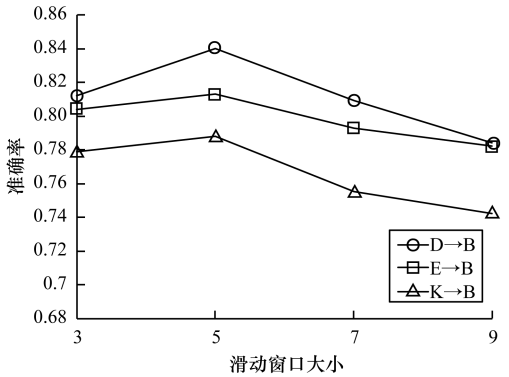

图 4 目标域为 B 时模型滑动窗口的准确率

当上下文窗口大小为 5、向量维度为 300 时,对模型最佳迭代次数进行测试。在目标域为 B 时,不同迭代次数的平均准确率如图 5 所示,可以看出,当迭代到 150 次后,准确率基本不变,即模型收敛。

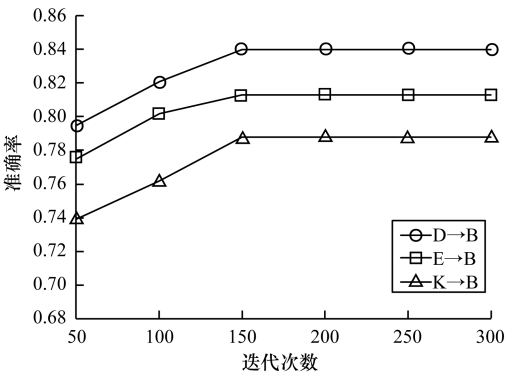

图 5 目标域为 B 时模型迭代次数的准确率

当上下文窗口大小为 5、迭代次数为 150 时,测试向量维度的最佳长度,如图 6 所示。从图 6 可以看出,向量维度的最佳长度为 300,设置过长或过短均不利于模型训练。

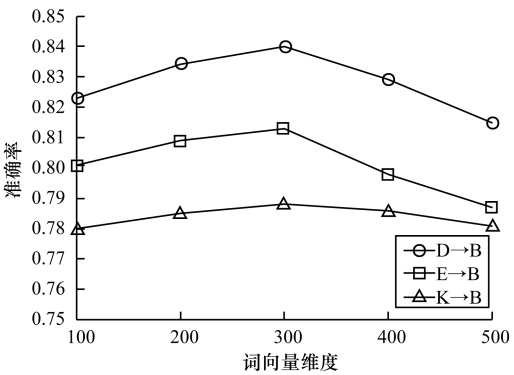

图 6 目标域为 B 时模型词向量维度的准确率

4　结束语

本文提出一种语义结构混合的迁移学习文本特征对齐算法,在抽取特征对齐时使用 GloVe 和 WordNet 对不同领域的特征从结构和语义上进行对齐。实验结果验证了该算法在准确率上优于传统算法。下一步将通过模型组合自动抽取特征进行映射,以降低模型复杂度。

参考文献

[1] LIU Bing. Sentiment analysis and opinion mining [EB/OL]. [2018-01-16]. https://www. cs. uic. edu/ ~ liub/FBS/SentimentAnalysis-and-OpinionMining. html.

[2] LIU Bing. Web data mining [M]. Berlin, Germany: Springer,2011.

[3] PANG Bo,LEE L. Thumbs up? sentiment classification using machine learning [EB/OL]. [2018-01-16]. https://arxiv. org/pdf/cs/0205070. pdf.

[4] BLITZER J, MCDONALD R, PEREIRA F. Domain adaptation with structural correspondence learning [C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics,2006:120-128.

[5] PAN Jialin, NI Xiaochuan, SUN Jiantao, et al. Cross-domain sentiment classification via spectral feature alignment [C]//Proceedings of the 19th International Conference on World Wide Web. New York, USA: ACM Press,2010:751-760.

[6] ZHANG Shaowu, LIU Huali, YANG Liang, et al. A cross-domain sentiment classification method based on extraction of key sentiment sentence [J]. Natural Language Processing and Chinese Computing, 2015, 9362:90-101.

[7] 孟佳娜,段晓东,杨亮. 基于特征变换的跨领域产品评论倾向性分析[J]. 计算机工程,2013,39(10):167-171.

[8] BOLLEGALA D,WEIR D,CARROLL J. Cross-domain sentiment classification using a sentiment sensitive thesaurus [J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(8):1719-1731.

[9] XIA Rui, ZONG Chengqing, HU Xuelei, et al. Feature ensemble plus sample selection: domain adaptation for sentiment classification [J]. IEEE Intelligent Systems, 2013,28(3):10-18.

[10] ZHOU Guangyou, HE Tingting, WU Wensheng, et al. Linking heterogeneous input features with pivots for domain adaptation [C]//Proceedings of the 24th International Conference on Artificial Intelligence. [S. l.]:AAAI Press,2015:1419-1425.

[11] 魏晓聪,林鸿飞. 面向迁移学习的文本特征对齐算法[J]. 计算机工程,2017,43(2):215-219.

[12] ZHANG Yuhong,XU Xu,HU Xuegang. A common subspace construction method in cross-domain sentiment classification [C]//Proceedings of International Conference on Electronic Science and Automation Control. [S. l.]: Atlantis Press, 2015:48-52.

[13] ZHOU Guangyou, ZHOU Yin, GUO Xiyue, et al. Cross-domain sentiment classification via topical correspondence transfer[J]. Neurocomputing,2015,159:298-305.

[14] GLOROT X, BORDES A, BENGIO Y. Domain adaptation for large-scale sentiment classification:a deep learning approach [C]//Proceedings of the 28th International Conference on Machine Learning. [S. l.]:Omnipress,2011: 513-520.

[15] BENGIO Y,GUYON G,DROR V,et al. Deep learning of representations for unsupervised and transfer learning [EB/OL]. [2018-01-16]. https://www. docin. com/p-1690924284. html.

[16] COLLOBERT R,WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning [C]//Proceedings of the 25th International Conference on Machine learning. New York, USA:ACM Press,2008:160-167.

[17] MARQUEZ L, RODRÍGUEZ H. Part-of-speech tagging using decision trees [C]//Proceedings of the 10th European Conference on Machine Learning. London, UK:Springer,1998:25-36.

[18] LANDAUER T K, DUMAIS S T. A solution to plato's problem: the latent semantic analysis theory of acquisition, induction,and representation of knowledge[J]. Psychological Review,1997,104(2):211-240.

[19] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2018-01-16]. https://arxiv. org/pdf/1301. 3781. pdf.

[20] PENNINGTON J,SOCHER R,MANNING C. GloVe:global vectors for word representation [EB/OL]. [2018-01-16]. https://nlp. stanford. edu/projects/glove/.

[21] MILLER G A. WordNet:a lexical database for English[J]. Communications of the Association for Computing Machinery, 1995,38(11):39-41.

编辑　赵　辉